THE UNIVERSITY OF HULL

The Development of Computational Methods for Large-Scale Comparisons and Analyses of Genome Evolution

being a Thesis submitted for the Degree of

Doctor of Philosophy

in the University of Hull

by

Stephen Paul Moss BSc (Hons) PGCert

December 2015

For Dorothy, Kelsey, Ashleigh, and James.

ACKNOWLEDGEMENTS

I feel I must first extend my sincerest gratitude to Dr. David Lunt and Dr. Domino Joyce, my primary and secondary supervisors, for their patience, resolve, motivation and humour. My PhD has been a particularly challenging journey, as one might expect (if it were easy, then it wouldn't be worthy of its prestige), and I am sure that, at times, I have induced more than my fair share of stress. Dave and Domino have been there to guide me through the rough patches and likewise, to celebrate the successes. They have elicited, like blood out of a stone, the makings of a scientist, from deep within, where seldom an action potential doth go, and allowed me the intellectual freedom to investigate many of the exciting areas of genome informatics and evolution. In addition to their academic support, they have also been wonderful friends, and I hope we can maintain our relationship for many years to come.

I should also like to thank Dr. Magnus Johnson. Magnus and I were introduced to one another by Domino, who had discussed with him my delightful habit of being distracted by interesting problems of a computational nature. I spent a fortnight looking after his picturesque cottage in Hackness (complete with Dog, Cat and Chickens), perfectly located at the fringes of the North York Moors National Park, whilst simultaneously attempting to translate an ancient QBASIC program into the Python programming language. The task, a computational model of the reflective superposition compound eye, was successful (see Appendix 1.1) and I hope will lead to many collaborative publications with Magnus and his associates in the future. In addition to our research collaborations, I feel we have also become very good friends, not in the least due to our shared eclectic interests and distractible personalities.

Another source of great support throughout my PhD has been Dr. Chris Venditti. Chris came to Hull from Reading, where he completed his PhD and a subsequent Leverhulme Early Career Fellowship under the supervision of the ingenious Prof. Mark Pagel. Chris brought a fresh perspective to the university and to my PhD. He is an inspirational, talented and brilliant researcher, who has been somewhat of an icon to me during his time in the department. During our various discussions over coffee (many spent rectifying his computer problems), he provided invaluable advise in both a professional and personal capacity, without which, I feel I would have walked away a long time ago. Our friendship is one that I hope will be a long a fruitful one and I

should very much like to collaborate with him in the future. His move back to Reading, will, I hope, not impede that.

These are people that have predominated my thoughts throughout my PhD, but of course there are a number of other, no less important people in the department, without whom I would also have likely fallen at the wayside. I should like to thank Dr. Stuart Humphries, who has provided several consultations on the use of statistical methods for my research. In addition he has also been a source of personal support and I have enjoyed spending many a summer day partaking in barbecues at his residence, and wandering round farms with our inquisitive children. In that regard, I would also like to extend my gratitude to Dr. Lesley Morrell, Stuart's better half. I have enjoyed collaborating with Lesley on an ecological modelling project, where I enabled the integration of a C program with MATLAB. The project involved modelling predator-prey interactions in the "selfish herd" using game theory computation. I have some interesting ideas for the computational direction of that project and should hope that we can also collaborate more in the future. Other academic staff that have been of great assistance are Dr. Bernd Haenfling, Dr. Africa Gomez, Dr. Lori Lawson-Handley and Dr. John Adams. This doesn't mean to say that other academics have not had any role in maintaining my sanity or clarity of thought, everyone has played their part in nourishing my focus and resolve.

My fellow postgraduate researchers have been a huge source of inspiration and support through the last three years of my life. Tom Mathers, and Dan Jeffries have provided critical input to my research and friendly lunchtime chats to distract me from my punishing schedule in the bioinformatics lab. In addition Greg McCullough, Helen Kimball, Renske Gudde, Victoria Smith, Andrea Simon, Cathleen Thomas, Michael Orchard, Paul Nichols, Laura Davidson, Fiona Hatch, Jackson Sage and many others have been a source of support, that perhaps they didn't even realise, purely by being there to talk to and maintain some remnants of social contact, without which I would likely have regressed to a vegetative state, only stimulated by the plethora of zeros and ones that have so imperceptibly infested my psyche. That being said, the computers have often been my "friend", when I have wanted to experience nothing more than the whir of electrons, through the printed circuit boards that bring them "life". It is, therefore, perhaps also relevant to extend my thanks to the valuable and ever-conscious world of social networking. I have made a great number of friends across the world and been able to implement valuable networking opportunities, due to the bridging of distance and time, by services such as Google+, Twitter and Facebook. I would particularly like to extend my gratitude to the members of the #phdchat and #eswphd communities.

Outside of the university, in addition to the contacts I have made across the world via social networking sites, I also feel indebted towards the excellent and highly proficient staff at the European Bioinformatics Institute in Hinxton near Cambridge. I was fortunate enough to be awarded a European Molecular Biology Organisation Short-Term Fellowship Grant in August 2011, which allowed me to spend two weeks working with the Ensembl Software Development Team as part of the "Geek for a Week" programme. I was able to enjoy an exceptionally productive fortnight working alongside some of the brightest minds in bioinformatics, in order to produce a prototype Ensembl REST API that would allow programmatically independent access to Ensembl's data. I learnt an amazing amount that helped to not just develop my programming skills in general, but that also gave me an insight into the production environment behind Ensembl's approximately quarterly release schedule. I am particularly grateful to Dr Paul Flicek for arranging the visit in the first instance, to Dr Glenn Proctor the former Ensembl software development team leader for liaising with the team and discussing possible project ideas, to Dr Stephen Keenan my primary mentor whilst undertaking the research project, and a source of a wealth of Perl programming knowledge, and to Andy Yates the current Ensembl software development team leader, who took over the role during my period at Hinxton, and who gave me a great deal of advice and encouragement. The other team members were also extremely helpful in providing me with advice and a friendly, relaxed working environment in which I could complete the project; these were Dr Rhoda Kinsella, Monika Komorowska, Ian Longden, and Andreas Kahari. I am also extremely grateful towards Dr Bert Overduin on the Ensembl Helpdesk for advice and assistance throughout my PhD and to Dr Ewan Birney for being an inspirational scientist and bioinformatician in whose footsteps I'm sure all would like to follow.

No acknowledgments section would be complete without the obligatory mention of one's friends and family. This is by no means however, included through traditional requirement. Without my friends and family I wouldn’t have made it through my undergraduate studies, let alone my postgraduate ones. My Mum and Step-Dad and my Dad and Step-Mum have always been supportive of my decision to return to education and have provided motivation, personal advice and the occasional kick up the arse that has allowed me to maintain my sanity through the last 7 years. My sisters; Stephanie, Shelli, Sophie and Louise have provided love, affection and constant support, that has allowed me to keep things in perspective and not lose sight of my goals. They have also provided thoroughly enjoyable discussion, relaxation and much more than mild intoxication (perhaps closer to inebriation) during our sparsely distributed family gatherings. Additionally, throughout my life, my Uncle, Dr Andrew Davidson has been a huge inspiration to me. He has fuelled my desire to be a better person and to learn all that I can about life and to more importantly, question all that I have learnt. He is a passionate, intelligent, and diverse individual that has always set the bar high for me. He has provided a great deal of input and advice on all of my work throughout the last 7 years and I feel I would have accomplished far less without him. In addition to the academic and personal support, he has also been an amazing friend. I love him dearly and hope we can maintain a close relationship, wherever life might take us, for the rest of our days.

Penultimately, to my Grandma, Dorothy, who unfortunately succumbed to cancer in February 1991. She was my best friend as a child. I would spend every waking minute I could with her and it was an unbelievable loss to me when she died. Everything I do now in life, I do with her in mind, and although she is gone, I still hope that my actions would make her proud if she were still here. Likewise, undertaking a PhD, in addition to my passion for research, is something I have undertaken in order to provide a better future for my children and to make them proud. My daughter Ashleigh will be almost 5 when I submit and has been a beautiful inspiration to me and the driving force behind my motivation across the final year of my undergraduate degree and throughout my PhD. She shares many of my character traits and passions and fills me with love and life. She has been the primary force in maintaining my sanity and has saved me from the pits of despair and worse on many occasions. Equally, my son James, who recently saw his 1st birthday, has dragged me kicking and screaming from an often-visited realm of negativity and has provided me with strength, compassion and resolve.

Finally, to my wife, Kelsey, who has had to put up with my idiosyncrasies, depressions and stresses for more than the last 7 years. She has been a beautiful symbol of strength and determination for me and inspired me to keep chipping away in the face of internal turmoil and adversity. Were it not for her, I would likely have given up on everything in my life and I would certainly be in a much darker place right now. As it happens, we are building a fruitful future together and raising our own wonderful family. It is with unbelievable gratitude and the delights of happiness and love that I thank her for being there for me and standing by me through these difficult few years. From now on, things can only get better.

CONTENTS

LIST OF FIGURES

Supplementary Figure 5.3 – Inset of frequency distribution of intron count in all 61 species in Ensembl release 70. Cut-off at 12.5 to emphasize mode count in the distribution.

Supplementary Figure 5.4 – Inset of frequency distribution of intron density in all 61 species in Ensembl release 70. Cut-off at 0.01 to emphasize mode density in the distribution.

Supplementary Figure 5.5 – Inset of frequency distribution of intron size in all 61 species in Ensembl release 70. Cut-off at 750 bp to emphasize mode size in the distribution.

LIST OF TABLES

LIST OF ACRONYMS

API	Application Program Interface
BCE	Before Common Era
bit	Logic state in computer systems architecture equal to 0 or 1 (off or on)
bp	Base pair
byte	8 bits
cDNA	Complementary DNA
CDS	CoDing Sequence
CE	Common Era
CORBA	Common Object Request Broker Architecture
CNV	Copy Number Variation
CPAN	Comprehensive Perl Archive Network
CRAN	Comprehensive R Archive Network
CSV	Comma Separated Values
DNA	Deoxyribose Nucleic Acid
DSBR	Double-Stranded Break Repair
EDA	Exploratory Data Analysis
FTP	File Transfer Protocol
Gb	Gigabase
Gbps	Gigabits per second
GB	Gigabyte
GNU	GNUs Not Unix
GOLD	Genomes OnLine Database
GPL	General Public License
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IDE	Integrated Development Environment
Indel	Insertion or Deletion
JSON	JavaScript Object Notation
Kb	Kilobase
Kbps	Kilobits per second
KB	Kilobyte

LUCA	Last Universal Common Ancestor
Mb	Megabase
Mbp	Megabits per second
MB	Megabyte
MRCA	Most Recent Common Ancestor
MVC	Model-View-Controller
N	Population Size
N_e	Effective Population Size
NHEJ	Nonhomologous End Joining
Pbps	Petabits per second
PB	Petabyte
pg	Picogram
RAM	Random Access Memory
REST	REpresentational State Transfer
RNA	Ribose Nucleic Acid
RPC	Remote Procedure Call
SOAP	Simple Object Access Protocol
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
TSV	Tab Separated Values
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
UTR	UnTranslated Region
W3	World Wide Web
W3C	World Wide Web Consortium
WWW	World Wide Web
WWWC	World Wide Web Consortium
XML	eXtensible Markup Language
YAML	YAML Ain't Markup Language

CONSTANTS, DEFINITIONS AND EQUATIONS

picogram to bp conversion

Number of base pairs = mass in pg x 0.978 x 109

1_{pg} = 978 Mb

CAFE lambda definition

CAFE (De Bie *et al.*, 2006) defines lambda (λ) as the probability of both gene gain and loss per gene per unit time in the phylogeny - put more simply, it describes the rate of change as the probability that a gene family either expands (via gene gain) or contracts (via gene loss) per gene per million years.

“It has often and confidently been asserted, that man's origin can never be known: but ignorance more frequently begets confidence than does knowledge: it is those who know little, and not those who know much, who so positively assert that this or that problem will never be solved by science.”

— Charles Darwin, The Descent of Man, 1871.

“The most erroneous stories are those we think we know best - and therefore never scrutinise or question.”

— Stephen Jay Gould, Full House, 1996.

“Evolution has meant that our prefrontal lobes are too small, our adrenal glands are too big, and our reproductive organs apparently designed by committee; a recipe which, alone or in combination, is very certain to lead to some unhappiness and disorder.”

— Christopher Hitchens, God is Not Great, 2007.

"Increasingly, the real limit on what computational scientists can accomplish is how quickly and reliably they can translate their ideas into working code."

— Greg Wilson, Where's the Real Bottleneck in Scientific Computing?, 2006.

ABSTRACT

The last four decades have seen the development of a number of experimental methods for the deduction of the whole genome sequences of an ever-increasing number of organisms. These sequences have in the first instance, allowed their investigators the opportunity to examine the molecular primary structure of areas of scientific interest, but with the increased sampling of organisms across the phylogenetic tree and the improved quality and coverage of genome sequences and their associated annotations, the opportunity to undertake detailed comparisons both within and between taxonomic groups has presented itself. The work described in this thesis details the application of comparative bioinformatics analyses on inter- and intra-genomic datasets, to elucidate those genomic changes, which may underlie organismal adaptations and contribute to changes in the complexity of genome content and structure over time. The results contained herein demonstrate the power and flexibility of the comparative approach, utilising whole genome data, to elucidate the answers to some of the most pressing questions in the biological sciences today.

As the volume of genomic data increases, both as a result of increased sampling of the tree of life and due to an increase in the quality and throughput of the sequencing methods, it has become clear that there is a necessity for computational analyses of these data. Manual analysis of this volume of data, which can extend beyond petabytes of storage space, is now impossible. Automated computational pipelines are therefore required to retrieve, categorise and analyse these data. Chapter two discusses the development of a computational pipeline named the Genome Comparison and Analysis Toolkit (GCAT). The pipeline was developed using the Perl programming language and is tightly integrated with the Ensembl Perl API allowing for the retrieval and analyses of their rich genomic resources. In the first instance the pipeline was tested for its robustness by retrieving and describing various components of genomic architecture across a number of taxonomic groups. Additionally, the need for programmatically independent means of accessing data and in particular the need for Semantic Web based protocols and tools for the sharing of genomics resources is highlighted. This is not just for the requirements of researchers, but for improved communication and sharing between computational infrastructure. A prototype Ensembl REST web service was developed in collaboration with the

European Bioinformatics Institute (EBI) to provide a means of accessing Ensembl's genomic data without having to rely on their Perl API. A comparison of the runtime and memory usage of the Ensembl Perl API and prototype REST API were made relative to baseline raw SQL queries, which highlights the overheads inherent in building wrappers around the SQL queries. Differences in the efficiency of the approaches were highlighted, and the importance of investing in the development of Semantic Web technologies as a tool to improve access to data for the wider scientific community are discussed.

Data highlighted in chapter two led to the identification of relative differences in the intron structure of a number of organisms including teleost fish. Chapter three encompasses a published, peer-reviewed study. Inter-genomic comparisons were undertaken utilising the 5 available teleost genome sequences in order to examine and describe their intron content. The number and sizes of introns were compared across these fish and a frequency distribution of intron size was produced that identified a novel expansion in the Zebrafish lineage of introns in the size range of approximately 500-2,000 bp. Further hypothesis driven analyses of the introns across the whole distribution of intron sizes identified that the majority, but not all of the introns were largely comprised of repetitive elements. It was concluded that the introns in the Zebrafish peak were likely the result of an ancient expansion of repetitive elements that had since degraded beyond the ability of computational algorithms to identify them. Additional sampling throughout the teleost fish lineage will allow for more focused phylogenetically driven analyses to be undertaken in the future.

In chapter four phylogenetic comparative analyses of gene duplications were undertaken across primate and rodent taxonomic groups with the intention of identifying significantly expanded or contracted gene families. Changes in the size of gene families may indicate adaptive evolution. A larger number of expansions, relative to time since common ancestor, were identified in the branch leading to modern humans than in any other primate species. Due to the unique nature of the human data in terms of quantity and quality of annotation, additional analyses were undertaken to determine whether the expansions were methodological artefacts or real biological changes. Novel approaches were developed to test the validity of the data including comparisons to other highly annotated genomes. No similar expansion was seen in mouse when comparing with rodent data, though, as assemblies and annotations were updated, there were differences in the number of significant changes, which brings into question the reliability of the underlying assembly and annotation data. This emphasises the importance of an understanding that computational predictions, in the absence of supporting evidence, may be unlikely to represent the actual genomic structure, and instead be more an artefact of the software parameter space. In particular, significant shortcomings are highlighted due to the assumptions and parameters of the models used by the CAFE gene family analysis software. We must bear in mind that genome assemblies and annotations are hypotheses that themselves need to be questioned and subjected to robust controls to increase the confidence in any conclusions that can be drawn from them.

In addition functional genomics analyses were undertaken to identify the role of significantly changed genes and gene families in primates, testing against a hypothesis that would see the majority of changes involving immune, sensory or reproductive genes. Gene Ontology (GO) annotations were retrieved for these data, which enabled highlighting the broad GO groupings and more specific functional classifications of these data. The results showed that the majority of gene expansions were in families that may have arisen due to adaptation, or were maintained due to their necessary involvement in developmental and metabolic processes. Comparisons were made to previously published studies to determine whether the Ensembl functional annotations were supported by the de-novo analyses undertaken in those studies. The majority were not, with only a small number of previously identified functional annotations being present in the most recent Ensembl releases.

The impact of gene family evolution on intron evolution was explored in chapter five, by analysing gene family data and intron characteristics across the genomes of 61 vertebrate species. General descriptive statistics and visualisations were produced, along with tests for correlation between change in gene family size and the number, size and density of their associated introns. There was shown to be very little impact of change in gene family size on the underlying intron evolution. Other, non-family effects were therefore considered. These analyses showed that introns were restricted to euchromatic regions, with heterochromatic regions such as the centromeres and telomeres being largely devoid of any such features. A greater involvement of spatial mechanisms such as recombination, GC-bias across GC-rich isochores and biased gene conversion was thus proposed to play more of a role, though depending largely on population genetic and life history traits of the organisms involved. Additional population level sequencing and comparative analyses across a divergent group of species with available recombination maps and life history data would be a useful future direction in understanding the processes involved.

[This page is left intentionally blank]

CHAPTER ONE: INTRODUCTION

The need for automated bioinformatics pipelines

The first free-living organism to have its genome sequenced was the Gram-negative coccobacilli bacterium *Haemophilus influenzae*, an opportunistic pathogen known to cause a variety of diseases in humans (Fleischmann *et al.*, 1995). This study marked the beginning of the so called "genomics era", with the genomes of numerous organisms across the phylogenetic spectra being sequenced in increasing frequency over the subsequent 2 decades. Prior to this time the sequencing and analyses of the chemical components of the cell (i.e. DNA, RNA, and proteins) was a long, expensive and arduous process, which was in dire need of improved computational methods. The software available initially tended to focus more on complementing manual analyses of the data (e.g. Staden, 1977; Staden, 1978; Staden, 1979), but the increasing production of genome scale data spurred the development of the bioinformatics discipline, and a wide variety of automated tools (Staden, 1996; Piast *et al.*, 2007).

Although computational infrastructure has developed in parallel to sequencing advances, the volume of data produced through genome sequencing projects, over the past decade in particular, have posed a considerable challenge to their analyses (Mardis, 2011; Pagani *et al.*, 2012). The increase in numbers of genomes sequenced doesn't necessarily pose a problem on its own, in terms of individual analytical requirements however; it is the increasing quantities of genomic data and associated metadata, produced as a result of improvements in sequencing technologies, as well as redundancy due to inherent sequencing error (Bouck *et al.*, 1998; Green, 2007; Milinkovitch *et al.*, 2010) that makes it impossible for manual analyses to be undertaken, and thus necessitating the need for increasingly efficient bioinformatics algorithms and automated computational pipelines (Lathe *et al.*, 2008).

The genome of *H. influenzae* is 1,830.14 Kb in size (Fleischmann *et al.*, 1995). For the computer systems available in 1995; with a maximum hard disk capacity of approximately 1 GB, processor clock speed of around 33 MHz and RAM of roughly 8 MB, this project would have posed a considerable problem, as it was much larger than the maximum 5.4 Kb (bacteriophage ϕX174) and 48 Kb (bacteriophage λ) DNA genomes previously sequenced (Sanger *et al.*, 1977; Sanger *et al.*, 1982). It is useful to remember here that the size of a genome in base pairs (bp) doesn't equate to amount of storage in bytes, due to the intricacies of computer systems architecture design and file system formats, but also due to the amount of metadata produced as part of the sequencing process. The redundant reads produced are just one source of this surplus, which require vast computational resources alone, to be ultimately disposed of. In addition, the tools required for analyses (e.g. Staden, 1996), which were novel algorithmic methods (and therefore, perhaps, not optimal) in the early days of genome analyses, had computational overheads of their own. In short, the development of robust and efficient computational protocols was a necessity.

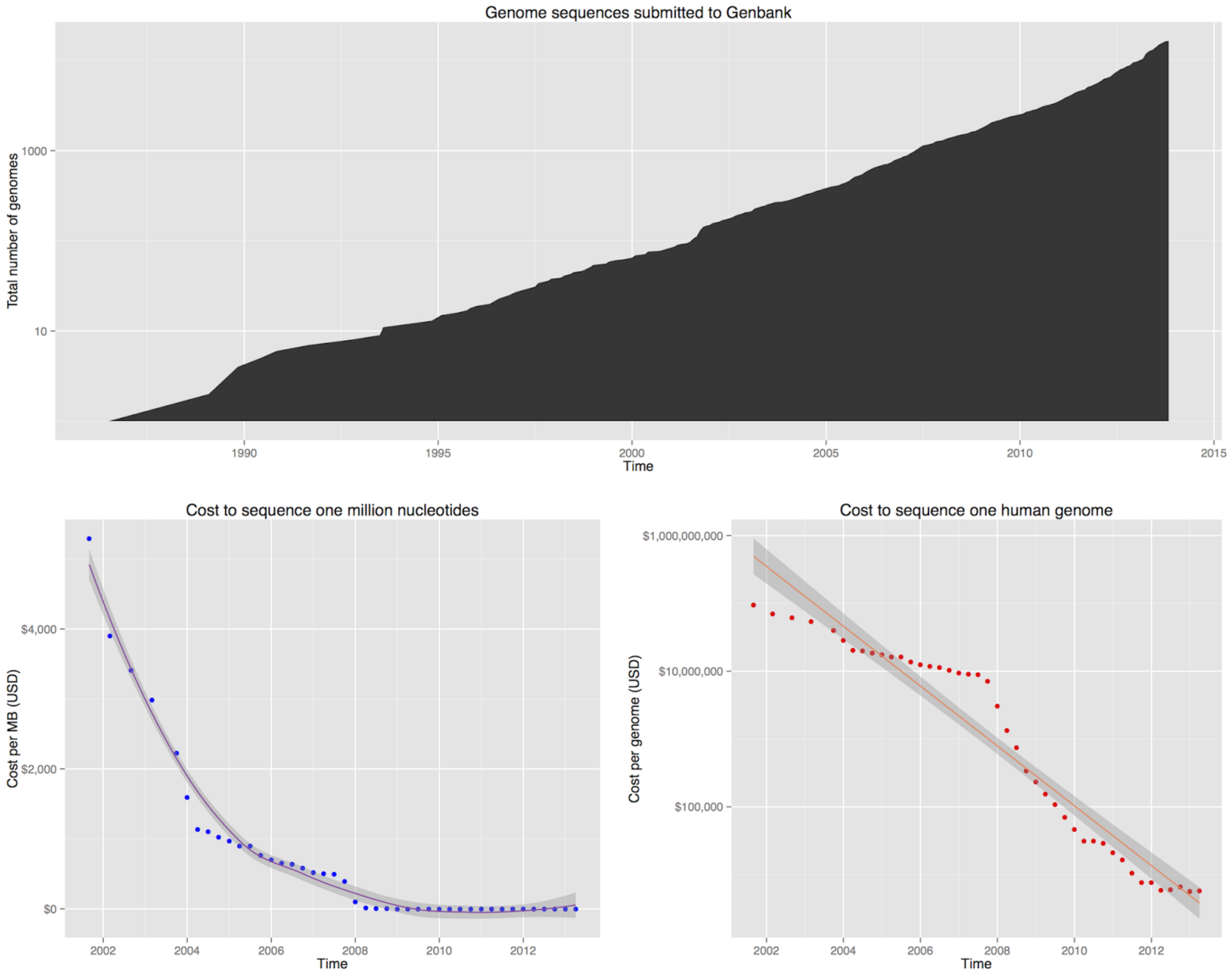
The impact of increasing sequence data on bioinformatics analyses

On the 12th September 2013, the number of genome projects listed in the Genomes OnLine Database (GOLD) stood at 30,377 (Pagani *et al.*, 2012). This number is in stark contrast to the 1 published bacterial genome in 1995, and has seen an exponential growth from then onwards. Although the number of genomes sequenced aren't a direct computational problem, the different quantities and qualities of those sequenced data are. The differences in sizes of genomes, even just in animals, is widely variable (Dufresne and Jeffery, 2011); with the smallest animal genome attributed to the plant-parasitic nematode *Pratylenchus coffeae,* at approximately 19.56 Mb (0.02 pg) (Leroy *et al.*, 2007), and the largest being the marbled lungfish *Protopterus aethiopicus* at approximately 129.91 Gb (132.83pg) (Pederson, 1971). Sequence coverage and read length are even more variable between individual projects and sequencing technologies, making for increasingly complex computational analyses in order to converge on the most likely genome model.

The increase in the number of available genomes, coupled with improvements in sequencing technologies has resulted in a massive increase in the volume of data stored in public servers at continuously decreasing cost (see Figure 1.1). Although read lengths and coverage have increased, however, the objective quality of the data, in terms of the confidence in individual base calls, are still unclear (Ye *et al.*, 2011; Earl *et al.*, 2011; Bradnam *et al.*, 2013). In addition to the underlying sequencing and base calling, there are problematic regions within genomes that compound issues further including gene families and pseudogenes, regions of high GC content, known structural variants, repeat sequences, homopolymers, and compressions. There is a great deal of variation in the frequency and size of these problematic regions within the genomes of all organisms, though there is some bias within particular clades, which are largely due to population genetic factors (Lynch and Conery, 2003; Lynch and Katju, 2004; Perry *et al.*, 2008). These problems with the sequence data, on top of the issues with the underlying sequence itself make it very difficult to converge on an accurate genome assembly. There has been a great deal of effort invested in producing effective bioinformatics algorithms for genome assembly (Miller *et al.*, 2010; Finotello *et al.*, 2011; Zhang *et al.*, 2011; Wajid and Serpedin, 2012), but in the absence of validation methods and metrics, it is difficult to interpret the data accurately.

Lagging computational power influences algorithm development

Computational power, measured in terms of the density of transistors in microprocessor units, has increased at a rate of roughly double every 18 to 24 months since 1971 (see Figure 1.2), a phenomenon known as Moore's Law (Moore, 1965). Likewise, the storage capacity of hard drives has increased at a similar rate (see Figure 1.3) and RAM capacity has increased exponentially by a factor of 10 every 4 years (Buttazzo, 2000). The number of genomes and amount of data being produced by sequencing technologies however, is increasing even faster. This means that the computational resources necessary for processing these data are always lagging behind. As many problems in bioinformatics are hard in terms of computational complexity (Jones and Pevzner, 2004; Chor and Tuller, 2005; Roch, 2006), this has resulted in most methods focusing on the development of approximate or parsimonious algorithms, which apply probabilistic models and statistical approaches in order to reach the most likely outcome (e.g. Eddy, 1998; Huelsenbeck and Ronquist, 2001; Enright *et al.*, 2002; Edgar, 2004; Huang *et al.*, 2005; Wehe *et al.*, 2008; Van Dongen, 2008). The focus on approximation within bioinformatics algorithms in order to reduce the computational burden, results in changes to the sensitivity and specificity of the methods, which can have a big impact on the outputs. The assumptions that are made to allow for scaling with available computational infrastructure, therefore don't give completely reliable results. Genome assembly and annotation is just one of the areas where these issues are inherent (Brenner, 1999; Devos and Valencia, 2001). The effect of different approaches to genome assembly algorithm development can be seen in the large discrepancies in the assemblies produced when varying individual parameters, even within the same software (Earl *et al.*, 2011; Bradnam *et al.*, 2013).



Genome sequences submitted to Genbank

Cost to sequence one million nucleotides

Cost to sequence one human genome

Figure 1.1 - Increase in the number of genome sequences deposited in GenBank since 1986 (top) and the change in cost of sequencing a genome since 2001 (bottom).

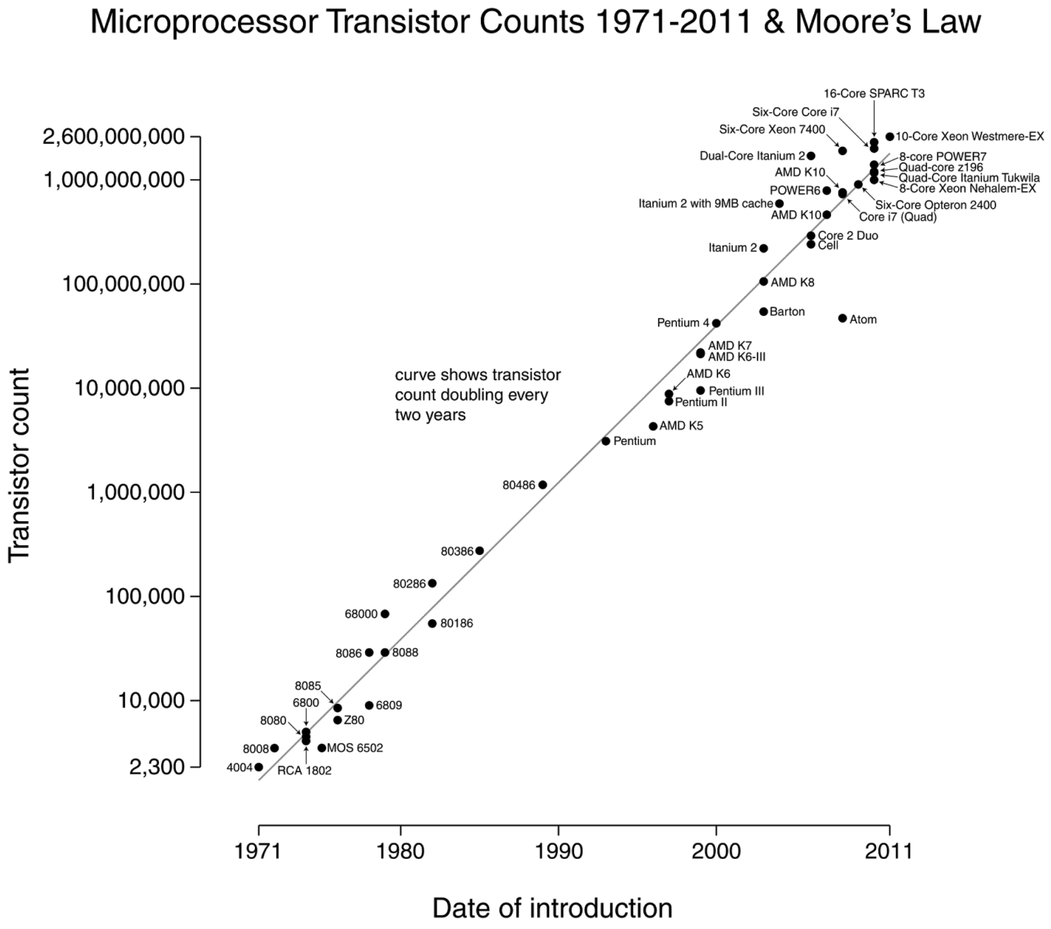


Figure 1.2 - Increase in microprocessor transistor count since advent of Moore's Law. Taken from http://en.wikipedia.org/wiki/File:Transistor_Count_and_Moore%27s_Law_-_2011.svg (CC BY-SA 3.0).

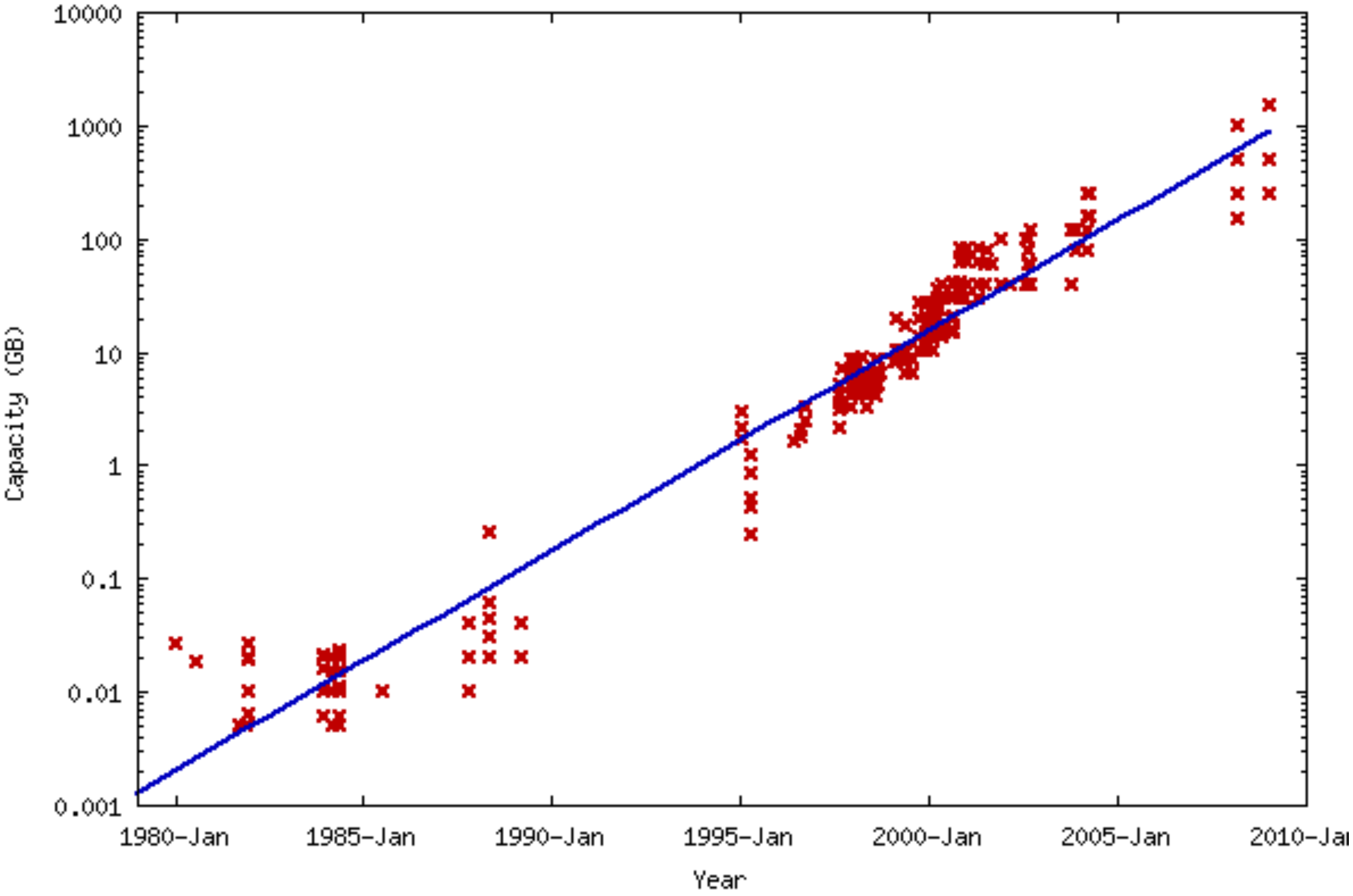


Figure 1.3 - Increase in hard drive storage capacity over time. Taken from http://en.wikipedia.org/wiki/File:Hard_drive_capacity_over_time.png (CC BY-SA 3.0).

The development of centralised resources to overcome data management and sharing issues

An approach that has been widely adopted to deal with the issues surrounding the storage and analyses of sequence data, is the development of centralised biological databases. These database projects bring together vast resources, not least in terms of funding, to integrate the sequencing platforms, and computational infrastructure with the processing and analytical pipelines necessary to make sense of the sequence data. In the early days of sequencing this involved small repositories of data relevant to the sequencing of specific components or organisms that were often distributed on physical media (Bernstein, *et al.*, 1977; Courteau, 1991). This grew into central data repositories independent of sequencing centres (Bilofsky *et al.*, 1986; Hamm and Cameron, 1986) that provide tools for querying and mining the data (Altschul and Lipman, 1990; Altschul *et al.*, 1990). This has scaled with the increase in number of active genome projects however, to incorporate data on vast numbers of species including Bacteria, Fungi, Metazoa, Plants and Protists alongside complex mining and analytics interfaces (Kersey *et al.*, 2011; Flicek *et al.*, 2012; Meyer *et al.*, 2013).

The development of these data resources has allowed for the increased automation of processes that were previously undertaken manually (Curwen *et al.*, 2004; Potter *et al.*, 2004). The sequencing, assembly, annotation, comparison and analytics of these genome data can be undertaken within relatively self-contained sites, or integrated via the Semantic Web (Berners-Lee and Hendler, 2001; Berners-Lee *et al.*, 2001) to distribute the computational burden across pooled computational infrastructure (Brooksbank *et al.*, 2010; Meyer *et al.*, 2012; Flicek *et al.*, 2012). These automated pipelines provide a wealth of metadata by integrating various computational tools, and linked data that can be used to highlight the underlying patterns in structure and content of these genomes, such as SNPs, duplications, and differing levels of gene expression (Rios *et al.*, 2010; Chen *et al.*, 2010; McLaren *et al.*, 2010; Vilella *et al.*, 2009; Ballester *et al.*, 2010). The most well-known of these resources are the UCSC genome browser (Meyer *et al.*, 2012) and the EMBL-EBI based Ensembl genome browser (Flicek *et al.*, 2012). The infrastructure alone that these kind of services aggregate is vast (Cuff *et al.*, 2004; Schadt *et al.*, 2010; Stein, 2010). They provide central hubs for the collation and dissemination of petabytes (PB) of data that are the lifeblood of global research communities (Brooksbank *et al.*, 2010; Smith *et al.*, 2013), however, there is still more that can be done to utilise the Semantic Web in integrating the various sources of biological data (Stein, 2002; Stein, 2008; Stein, 2010).

The production of bioinformatics tools to facilitate access to biological data

The UCSC genome browser and the tools it provides are an excellent resource for researchers (Kuhn *et al.*, 2012; Meyer *et al.*, 2013; Karolchik *et al.*, 2014), however it could be argued that Ensembl makes it much easier for less computationally proficient scientists to undertake research using their data. The services they provide are numerous and more importantly user-friendly (Birney *et al.*, 2004; Kasprzyk *et al.*, 2004; Stalker *et al.*, 2004; Spudich *et al.*, 2007; Spudich and Fernández-Suárez, 2010; Kinsella *et al.*, 2011). Their Perl-based API in particular provides an excellent programmatic interface to their data (Stabenau *et al.*, 2004), albeit requiring some investment in learning (though see Hubbard *et al.*, 2009 and Flicek *et al.*, 2009 for details on outreach and training activities). These sorts of programmatic tools can be used to develop focused solutions to specific biological questions, utilising Ensembl's genome data. This allows for the production of automated and reproducible workflows that can retrieve and perform relevant munging of the data, execute analytical packages and algorithms relevant to hypothesis testing, and output information (e.g. descriptive statistics, inferential models, and visualisations) necessary for publication.

Being able to reproduce ones analyses is of growing importance and it has become mandatory to provide both code and data as part of the publication process in many cases over the last few years (The EMBL Data Library and GenBank staff, 1987; Kaye *et al.*, 2009). This not online benefits the research community and public (Birney *et al.*, 2009; Wicks *et al.*, 2010; Sankoh and IJsselmuiden, 2011), but has also been shown to improve the authors citation rate (Piwowar *et al.*, 2007), though attitudes have been contradictory and much is still to be done (Ceci, 1988; Schofield *et al.*, 2009; Savage and Vickers, 2009; Nelson, 2009). The focus in reproducibility is beginning to change the way we approach bioinformatics projects however, by including more stringent policies on research related code and data (Stodden *et al.*, 2013; Petre and Wilson, 2013) and have even influenced a move towards the integration of code and data with publication in the form of "active papers" (Hinsen, 2011).

Widening access to biological data using the Cloud and the Semantic Web

There is a real demand for widening access to biological data that extends beyond reproducibility (Stein, 2002; Stein, 2008; Baker, 2012). The sharing of data needn't be an arduous and costly process, however. The development of protocols and the integration of computer systems for the purposes of sharing scientific data has been in progress since the founding of the World Wide Web (Berners-Lee *et al.*, 1996; Fielding *et al.*, 1999; Berners-Lee and Hendler, 2001). These protocols have already been used extensively for the dissemination of biological data (Dowell *et al.*, 2001; Hendler *et al.*, 2002; Stein, 2003; Lord *et al.*, 2004), leading to a strong case for moving towards a Cloud-based model for both data storage and analyses (Hoffa *et al.*, 2008; Keahey and Freeman, 2008; Dudley *et al.*, 2010), particularly in genome informatics (Stein, 2010; Wall *et al.*, 2010). The Cloud (including Amazon Web Services, and Google Compute Engine for example) is especially relevant given the excessive computational specification requirements of many bioinformatics tools (Schatz *et al.*, 2010), which may not be accessible to smaller-scale research laboratories. Smaller-scale, local computational analyses will of course continue to be necessary and the development of novel tools will be required regardless of the platforms they will be used on. The Cloud and the Semantic Web provide an excellent framework for scaling these tools and for integrating them with existing data resources globally, however. Indeed, many bioinformatics resource providers, including EMBL-EBI's Ensembl, are moving to public or hybrid Cloud models to scale with the growing demands (Bateman and Wood, 2009; Arrais and Oliveira, 2010; Baker, 2010; Flicek *et al.*, 2010; Dai *et al.*, 2012).

The Semantic Web is also useful for the development of bioinformatics resources. An issue with programming language dependent tools is the investment in time required to learn a new language (either in addition to others or from scratch), a new API, or a new development interface. There have been efforts to produce APIs to access Ensembl's data for example; in Ruby (Strozzi and Aerts, 2011) and in Python (Knight *et al.*, 2007), in addition to the native Perl API. This increases the coverage for development in additional programming languages, but these efforts often lag in their comprehensiveness in comparison to the Perl API. This is because Ensembl is a full-time managed project with a large team of developers, whereas the other APIs are either community projects or the focus of short-term grants. A way of overcoming these problems is with the development of programming language independent web services, such as those provided by the REST architectural style (Fielding and Taylor, 2000; Fielding, 2000). Ensembl have developed their own REST API (Flicek *et al.*, 2012), which allows access to their data via command-line utilities, internet browsers, or any other tool capable of utilising the HTTP protocol. Numerous other web services also exist (Gilbert, 2003; Stevens *et al.*, 2003; Kasprzyk *et al.*, 2004; Labarga *et al.*, 2007; Kinsella *et al.*, 2011; Kasprzyk, 2011) that widen and improve access to biological data. These will become of increasing importance in the field of genome analyses as the data deluge continues, particularly for comparative studies (Stein, 2010; Wall *et al.*, 2010).

Comparative genomics and evolutionary synthesis

Though there are many challenges and considerations surrounding the analyses of genome sequence data, the increasing numbers of sequenced genomes has allowed powerful comparative studies to come to the forefront of scientific investigation over the last two decades. Genomes are ultimately the source of instruction for each organismal unit (i.e. virus, or cell) and allow us the opportunity to deduce both contemporary and evolutionary information about them, but the complexities involved in understanding their sequence are substantial (Gregory, 2004; Brown, 2006; Lynch, 2007). With the sequencing of the genomes of medically important organisms such as *Haemophilus influenzae* (Fleischmann *et al.*, 1995); model organisms such as *Drosophila melanogaster* (Adams *et al.*, 2000); various primate genomes (Lander *et al.*, 2001; Venter *et al.*, 2001; Mikkelsen *et al.*, 2005; Locke *et al.*, 2011; Scally *et al.*, 2012); and now several sources of population level genomics data (Begun *et al.*, 2007; Liti *et al.*, 2009; McVean *et al.*, 2012), we can begin to ask questions, build models and make predictions based on comparisons of these data, that have not before been possible. Additionally, comparative genomics provides a method of validating metadata between different species and is now integral in the annotation of genomes (Clamp *et al.*, 2003; Hillier *et al.*, 2004; Vilella *et al.*, 2008; Flicek *et al.*, 2012).

Comparative genomics has assumed a central role in both the functional annotation of genome features and in our understanding of the nature of organismal divergence, adaptation, and genome evolution (Hardiso, 2003; Miller *et al.*, 2004; Drosophila 12 Genomes Consortium, 2007; Hahn *et al.*, 2005; Hahn *et al.*, 2007; Stajich *et al.*, 2007; Moss *et al.*, 2011; Prufer, 2012; Jones *et al.*, 2012; Alföldi and Lindblad-Toh, 2013). Comparing whole genomes is an extremely powerful approach to understanding the forces shaping genome structure and content within and between groups of organisms. There has been a failure however, to merge classical molecular evolution with modern large-scale genome informatics that is only just being realised (Rokas and Abbot, 2009; Alföldi and Lindblad-Toh, 2013). The comparative method has been central to biological analyses for over 3 decades (Kimura, 1980; Felsenstein, 1985; Harvey and Pagel, 1991) and provides a robust framework that has been adopted extensively in phylogenetic analyses (Pagel, 1994; Martins and Hansen, 1997; Pagel, 1999; Martins, 2000; Blomberg and Garland, 2002; Butler and King, 2004; Hansen *et al.*, 2008; Eastman *et al.*, 2011), yet these principles are often forgotten when undertaking modern large-scale genomic analyses. For example, most of the underlying changes in genome structure and content are neutral and therefore subject to population genetic forces (Lynch, 2007), yet are often incorrectly attributed to adaptation under selection (Gregory, 2005).

By simply comparing genome data one is only able to describe the number and location of differences (or similarities) that occur and not whether they are biologically relevant or real. The challenges facing their validation are extremely complex (Chain *et al.*, 2003; Jones and Pevzner, 2004; Moore, 2010), which are compounded by the underlying issues with assemblies and annotations (Brenner, 1999; Devos and Valencia, 2001). It is necessary to utilise our existing knowledge of molecular evolution alongside the development of powerful computational methods and bioinformatics algorithms in our approach to understanding these data. There is much to be done, but steps towards improving reliability (Howison *et al.*, 2013; Ghodsi *et al.*, 2013; Rahman and Pachter, 2013; Clark *et al.*, 2013; MacManes and Eisen, 2013; Le *et al.*, 2013; Ilie and Molnar, 2013) and analytical methods in light of the error-prone data (Hubisz *et al.*, 2011; Löytynoja *et al.*, 2012; Han *et al.*, 2013) are being made. By building more robust and comprehensive approaches to the analyses of genome data we will be able to more accurately reflect the true biological signal, and thus be more confident in the information we are able to extract and the conclusions we are able to draw from it.

Thesis Overview, Goals and Intentions

The work described in this thesis will detail the application of novel comparative bioinformatics methods to inter- and intra-genomic datasets. I will describe the analyses of a diverse range of genomes in order to test specific hypotheses relating to the molecular evolution of introns, gene duplications and repetitive elements, which have previously been determined to contribute most to variation in genome size and complexity, particularly in eukaryotes (Lynch and Conery, 2003). I will undertake novel computational analyses to increase our understanding of how genome structure and content evolves across a wide range of species. In addition I will develop new tools and methods of approaching the problems inherent with genomic analyses, as discussed in this thesis, in order to provide a framework for simplified analytics of genomic data in future research.

In Chapter Two I discuss the production of a computational pipeline named the Genome Comparison and Analysis Toolkit (GCAT). The development of this pipeline was necessary to perform the data mining and analytics required throughout this thesis, but also provides an excellent resource for the wider research community. Several examples of the pipeline's utility are provided, including analyses of the structure and content of genes in *Mus musculus*; characterisation and analyses of the repetitive landscape of the Great Apes; and a large-scale comparative analyses of intron structure and content across all available species in the Ensembl database. In addition, chapter two approaches the topic of access to scientific data, which is a common issue facing biological research on the whole. A Semantic Web resource in the form of a prototype Ensembl REST API is described, which provides programming language independent access to Ensembl's genome data. A Python wrapper is also developed to utilise the REST API. Finally, analyses are performed to determine how raw MySQL queries, the Ensembl Perl API and the Ensembl REST API compare in terms of efficiency and efficacy.

Chapter Three details a comparative genomic analyses of the 5 teleost genomes available in the Ensembl databases, with a focus on understanding their intron structure and content. In addition to highlighting the descriptive statistics and distributions of the introns and their associated repetitive content across these genomes, a hypothesis testing approach is taken to determine the evolutionary cause of a novel difference in the intron size distribution in the zebrafish, *Danio rerio*. The chapter concludes the difference in intron distribution was due to an ancient expansion of repetitive elements within introns of the size class 500-2,000 bp, though increased phylogenetic sampling, particularly including an outgroup species, will allow more conclusive examination of this result. This study resulted in a peer-reviewed publication in the journal Genome Biology and Evolution entitled "Comparative Analysis of Teleost Genome Sequences Reveals an Ancient Intron Size Expansion in the Zebrafish Lineage" (Moss *et al.*, 2011) that can be accessed via http://dx.doi.org/10.1093/gbe/evr090.

Chapter Four explores the evolution of gene families in terms of their number of members and functionality, with a focus on determining significant changes in gene family size in the primates. A large expansion in gene family sizes is highlighted in the branch leading to modern humans that is unexpected in terms of the amount of temporal change relative to other branches of the phylogenetic tree. The expansion in humans is followed up with rigorous testing to determine its validity, including comparison with the available rodent genomes in Ensembl's databases. The expansion is thought to be an artefact of the extensive population and tissue specific sampling undertaken in humans, but perhaps more concerning there also seems to be a significant affect of the assumptions and parameters of the models used by the CAFE gene family analysis software. The significantly changed gene families are then examined in connection with the functional classification of their gene members to test whether they are the result of adaptive evolution. Ensembl's comparative genomics annotations and associated Gene Ontology data are retrieved for the relevant gene members in order to determine their annotated function. A hypothesis is developed that expects a bias in functions of a reproductive, immune or developmental nature and the data largely confirms this. Additionally a comparison is made with the findings of previous studies in order to provide additional means of validation, though this emphasises the problems inherent in the different approaches to genome assembly and annotation.

Chapter Five examines the impact of gene family size on the evolution of introns. Data are retrieved for all 61 species in the release 70 Ensembl databases in order to describe, visualise and test for correlations between gene family size and intron characteristics (intron size, intron count, and intron density). The chapter also examines the spatial distribution of intron characteristics at the chromosome level in humans only. These analyses are undertaken to determine whether the position of introns in the genome influences their evolution. This is an important analyses as it allows us to make progress towards determining whether it is more abstract spatially-relevant molecular processes that drive the evolution of lower level features such as introns, or whether more traditional forces such as non-homologous recombination are more involved. It is determined that gene family size has a weak impact on the evolution of genomic features such as introns, and that conversely the spatial location within the chromosomes has a larger effect. A location specific effect highlights a potential role of GC-rich isochores and epigenetics on the underlying evolution of genomic features. The most striking effect is seen by separating intron characteristics into groups of autosomes and sex chromosomes. There is a significant lower number and size of introns across the Y chromosome in particular highlighting how lack of recombination can impact on the intron landscape.

[This page is left intentionally blank]

CHAPTER TWO: THE DEVELOPMENT OF COMPUTATIONAL TOOLS FOR COMPARATIVE BIOINFORMATICS ANALYSES: THE GENOME COMPARISON AND ANALYSIS TOOLKIT (GCAT) AND AN ENSEMBL RESTFUL WEB SERVICES FRAMEWORK

Literature Review

The power of bioinformatics software pipelines in genomics analyses

It was obvious, even with the sequencing of the very first genome, that of the bacteriophage phi-X174 at a size of 5,386 bp (Sanger *et al.*, 1978) that automated computational software was needed to make sense of the volumes of biological data (for example Staden *et al.*, 1999). The application of computer hardware and software wasn't just limited to its ability to aid in the assembly of genomes in order to deduce their primary structure however, but extended to their annotation, such as the identification of genes, and to more complex modelling of change in their structure over time. Tasks such as assembly and annotation simply aren't possible, or are intensely time consuming to undertake manually.

The ability to assemble and annotate genomes is an essential requirement of any genome sequencing project, as we would otherwise be left with a random assortment of nucleotides that could serve no further purpose. Being able to perform comparisons between genomes however is even more powerful, as it allows us to use the genome sequence and annotated metadata to ask specific biological questions, such as how phenotypic traits, inter-specific relationships, and ultimately life came into being and have changed over time.

As more divergent genomes were sequenced the ability to undertake comparisons within and between species presented itself and the field of comparative genomics was born. Comparative analyses, by their very nature, require the consideration of the underlying relationships between species in order to account for the phylogenetic signal inherent in their sequence. This has, in part, driven the sequencing of more species to increase the sampling across all areas of the tree of life. Due to the differing sizes of these genomes and the complexity involved in their assembly and annotation, there still exists a bias in sampling towards smaller genomes, particularly of bacterial species (Pagani *et al.*, 2012).

The sequencing, analyses and storage requirements of genome sequencing initiatives has fuelled the development of database projects and large scale data warehouses such as DDBJ (Tateno *et al.*, 2002), EMBL's ENA (Leinonen *et al.*, 2010) and GenBank (Benson *et al.*, 2005). These kind of projects have matured into unified genome service providers such as The Ensembl Database Project (Hubbard *et al.*, 2002) and The UCSC Genome Browser (Fujita *et al.*, 2010), which provide assembly, annotation, analyses, storage, and sharing of data all built on powerful bioinformatics software pipelines.

Existing computational software and the need for a broader solution

In general researchers need free and simple access to genomic data from their chosen repositories. Comparative molecular evolutionary analyses necessitates the retrieval of homologous structures or features of a uniform type such as orthologous genes or introns. There is a need to analyse these both quantitatively and by the extraction of individual components for example UTRs (untranslated regions) or splice sites. The methodological approach should be generic and efficient across data sources and analyses types facilitating the extension to a range of different questions.

Few solutions exist that are designed to undertake such broad-scale comparative analyses of whole genomes (Knight *et al.*, 2007; Yandell *et al.*, 2006). Most software focuses on providing solutions to the precise requirements of individual projects; commonly represented by a collection of source code files, which require some degree of manual configuration and manipulation in order to be executed. Attempts have been made to develop Application Program Interfaces (APIs) that provide a common framework for the bespoke development of scientific software (Kent *et al.*, 2002; Stabenau *et al.*, 2004; Strozzi and Aerts, 2011; Marygold *et al.*, 2012), but again these often result in the production of software that focuses on solving the specific problems at hand. Generic, accessible and flexible comparative bioinformatics toolkits that provide simple means of access to genomic resources are essential. The development of the Genome Comparison and Analysis Toolkit (GCAT) attempts to solve this problem.

The Ensembl databases (Flicek *et al.*, 2011; Flicek *et al.*, 2012) and Ensembl Perl API (Stabenau *et al.*, 2004) currently provide the richest, simplest and most powerful method of accessing genomic data for a variety of taxa. The Ensembl genome database project was launched in 2002 (Hubbard *et al.*, 2002) and required the design and implementation of both the underlying infrastructure, and software pipelines to deal with the processing (Cuff *et al.*, 2004; Potter *et al.*, 2004), as well as a programmatic means of sharing and accessing that data. The Ensembl Perl API was published in 2004 (Stabenau *et al.*, 2004) to meet the broader access requirements of the scientific community. In October 2012 (release 69) Ensembl housed 61 annotated genome assemblies in its main vertebrate genomes databases, with several others available in preview assembly format. Additionally they held data from 359 diverse invertebrate species in their Ensembl Genomes databases (Kersey *et al.*, 2011), which include taxa from the Metazoa, Protists, Bacteria, Plants, and Fungi. The UCSC genome browser (Fujita *et al.*, 2010) also houses a number of annotated genomes and is highly regarded as a source of data for comparative analyses, however, in the scope of this study the ease of access to Ensembl data, particularly in relation to their Perl API, was a key aspect in the decision to design the GCAT software pipeline around Ensembl's existing architecture.

The decision to develop a bespoke solution on top of Ensembl was not taken lightly. PyCogent (Knight *et al.*, 2007) is a powerful comparative genomics toolkit that would have been ideal for use in these studies, however its ability to connect to remote genomic data sources was limited. The PyCogent project didn't have the dedicated resources to maintain development alongside the fast paced Ensembl release schedule for example, and therefore lagged behind in its support for some of their cutting edge features. There was a clear need for a solution that could complement the Ensembl Perl API (Stabenau *et al.*, 2004) and yet expand on it to provide a comprehensive solution to comparative genomic analyses.

Problems with access to biological data

It is possible to access the Ensembl data via the Ensembl Genome Browser (Stalker *et al.*, 2004), via EnsMart (Kasprzyk *et al.*, 2004), BioMart (Kinsella *et al.*, 2011) or even directly via a MySQL client, however there is a learning curve associated with these methods of access that lends towards their use by the more computationally literature research scientists. There is, of course, an investment in time required to become familiar with any methodological technique, however complex computational skills are often required to mine data effectively. This is one of the drawbacks of developing APIs and scriptable software pipelines, as they may require the user to learn a new programming language, or even present a larger barrier to those that have never programmed before.

In addition to the issues with using and interacting with the different data access methods, there are other problems faced by research scientists when analysing data. 1) A large number of different file formats exist; some can be plain text files and others proprietary binary files, more still are minor variants causing a great deal of trouble with parsing the data in the first place. 2) These data are often kept using a variety of different storage solutions from flat files, through integration with various flavours of freely available SQL databases, to proprietary commercial storage solutions. The raw data, particularly in the case of flat files or SQL dumps, may be accessible over the Internet via an FTP server, or other file serving protocol, or it may not. 3) If APIs exist for access to these data, which they rarely do, then they are often programming language dependent. The Ensembl APIs for example are developed using the Perl programming language, creating a hurdle for those people that aren’t familiar with that language. These issues decrease the likelihood of being able to reproduce experiments effectively. It has been suggested that open-source distributed web services might be the most appropriate solution to this (Stein, 2002) since they use standardise formats for data exchange (e.g. FASTA or JSON), they abstract away from the storage layer and provide easy access to required data, and they are programming language agnostic; providing simple URI endpoints for access to data that can be utilised via a web browser, or any language with appropriate HTTP libraries.

The Semantic Web as a solution to data access and sharing

The Semantic Web, a collaborative project led by the main international standards organisation for the World Wide Web (WWW), the World Wide Web Consortium (W3C); aims to improved access to data, not just for humans, but also for computers. Improving the way that computers communicate increases the automated sharing of data between different data service providers. This in turn allows for data to be easily synchronised and updated, providing simplified access to the latest versions of datasets. A number of solutions exist that allow data to be shared between providers (Jenkinson *et al.*, 2008; Barsnes *et al.*, 2009; Brooksbank *et al.*, 2010; Gross, 2011) and projects such as ELIXIR (Croswell and Thornton, 2012) are pushing to increase this to a global scale.

One of the simplest semantic web service protocols is REST. REST stands for Representational State Transfer and is a software architectural style for designing and implementing distributed network applications (Fielding, 2000). REST is used for designing distributed network applications, primarily on top of the HTTP protocol (Fielding *et al.*, 1999). It provides a simple alternative to the other more complex mechanisms such as CORBA, RPC or SOAP. All that is needed in order to implement a REST interface is 1) knowledge of a resource, for example a uniform resource identifier (URI), 2) a method, for example the HTTP GET method (see Appendix 2.1) and 3) a data format, for example JavaScript Object Notation (JSON). When this is implemented over HTTP, it is known as a RESTful web service.

By utilising the Model-View-Controller (MVC) software architecture pattern (Reenskaug, 1979) it is possible to integrate existing data sources with a REST interface to provide programming language agnostic access to data. The model component allows one to wrap a database within a program object allowing the program code to interact with it at runtime. The controller component centralises the program logic and creates API calls that wrap SQL queries, simplifying the overall code. The view component allows data to be returned in a particular format depending on the user requirements. For example, if the user places a call with a `Content-Type` of `application/json` in the header, then the data will be returned in JSON format. The integration of the Ensembl Perl API with a REST software library (such as Perl's Catalyst API) to provide an Ensembl RESTful web service will allow access to Ensembl data using web browsers, command line tools, a variety of programming languages, or any other resource that is capable of utilising the standard HTTP protocols.

Aims of this chapter

In this chapter I develop a simple, flexible, and adaptable open-source software pipeline called GCAT, which allows for large-scale comparative analyses of the genome data available in Ensembl's MySQL databases. Although GCAT is flexible enough to undertake comparisons utilising any type of genome data, this thesis focuses primarily on analyses of vertebrate genomic data taken from introns, repetitive elements, and duplicate genes. I document the design, implementation and key attributes of GCAT, along with a demonstration of the power of its approach towards comparative genomics analyses of a number of datasets from Ensembl. In this chapter I focus on comparisons of gene structure in *Mus musculus*, repetitive element content in the Great Apes, and intron sizes across 52 divergent vertebrate genomes. I then describe the development of a RESTful web services framework that extends the Ensembl Perl API, allowing for programming language independent access to Ensembl's genome sequence data and annotations. I compare the efficiency of different methods of access to Ensembl's data by comparing Ensembl's Perl API against the REST API. I also review the most appropriate ways forward in relation to semantic web services and genomic data analysis.

Implementation

GCAT

The GCAT pipeline is implemented using the Perl programming language, allowing for efficient access to a large number of openly available community resources. The software primarily utilises the Ensembl Core Software Libraries (Stabenau *et al.*, 2004), a collection of Perl-based APIs allowing access to the Ensembl genome databases. BioPerl is also heavily utilised (Stajich *et al.*, 2002) as is R (R Core Development Team, 2012) for statistical analyses via the `Statistics::R` Comprehensive Perl Archive Network (CPAN) package. The use of the BioPerl and other open-source community libraries allows us to provide parsing routines for processing data in a variety of common bioinformatics formats, extending the application of the pipeline beyond Ensembl alone.

GCAT stores its output in Comma Separated Value (CSV) flat-files, whilst also working with local copies of sequence data which it stores in FASTA format. By using these common file formats it makes creating downstream analysis scripts far more efficient. GCAT has a number of ready-made plugin scripts implemented for processing CSV and FASTA data and a detailed API for additional development.

In order to reduce network latency it is recommended to install a local copy of the Ensembl MySQL data, and GCAT provides a support script to retrieve and install data from the current release (see https://github.com/gawbul/gcat/blob/master/support_files/get_ensembl.py). Network latency can be a problem when accessing data across the Internet, especially at peak times, or from poorer quality network connections. The decrease in runtime is often several-fold, even when using high-bandwidth connections, such as those available to universities. The difference between a gigabit local area network connection for data transfer and the limited (<10 megabit) connection across the rest of Internet, is substantial enough to warrant this process, especially in larger-scale studies.

GCAT (see Figure 2.1) provides a front-end script named `gcat.pl` which can be invoked without any arguments to process the default workflow file, `gcat-pipeline.txt` or instead, by giving the `-f` input flag, a custom workflow file can be used. The processing position in the workflow file is tracked, allowing the user to restart from where they left off should the pipeline be stopped for any reason. The `-c` input flag can be given to execute a shell environment that the user can use to run scripts manually. A number of built-in commands are available, which can be investigated using the help command. Additional scripts can be developed using the GCAT APIs and placed in the Scripts folder, which then allows them to be automatically identified by the GCAT pipeline. We provide several ready-made scripts that have been used in our previous studies, which can be listed using the `scripts` command. Detailed documentation and example code is available on the GitHub repository wiki at https://github.com/gawbul/gcat/wiki.

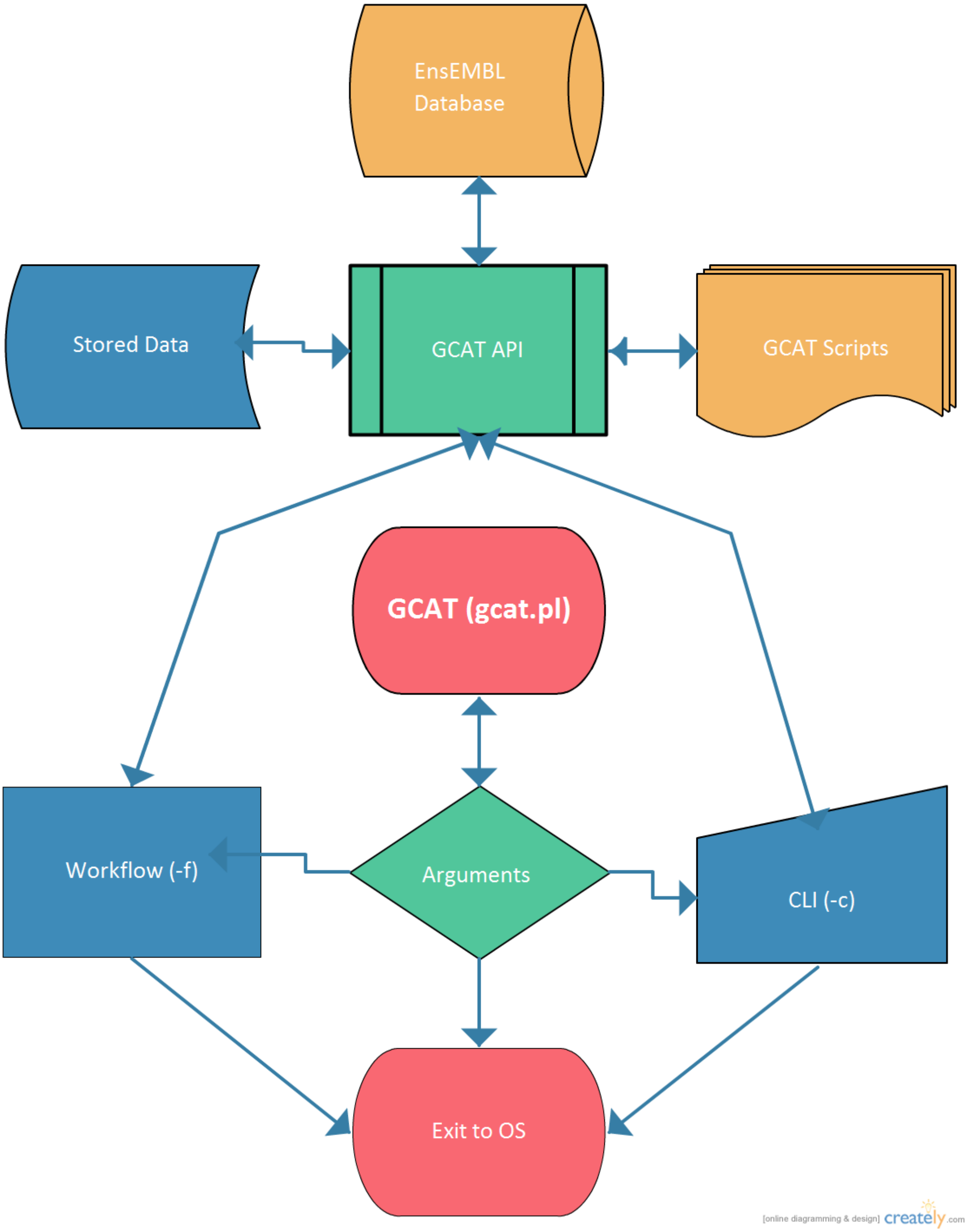


Figure 2.1 -A flowchart detailing the design and internals of the GCAT pipeline.

REST

The Perl Programming Language was used along with several open-source Comprehensive Perl Archive Network (CPAN) modules. The main library was the Catalyst Framework, a set of modules that enables the Model-View-Controller (Reenskaug, 1979) architectural pattern and REST (Fielding, 2000) architectural styles of distributed network application development in Perl. The Eclipse integrated development environment (IDE) version Classic 3.7 along with the EPIC plugin version 0.5.33 was used for management of this project. Python version 2.7.1 and Ruby version 1.8.7 were used to implement examples. JSON was used for the view component in the programmatic examples, but YAML Ain't Markup Language (YAML) was the format returned via the browser window. Additional formats such as eXtensible Markup Language (XML) are automatically supported and are accessible via setting the `Content-Type` HTTP header.

The `Bio::EnsEMBL::Registry` module was integrated into the model component of the framework and extended all the Ensembl Perl API functionality. A single controller was implemented with the base path part termed `/get_adaptor/` taking three arguments in the form `/species/database/adaptor/`. This retrieved the relevant adaptor object via the registry, for example the URI `/get_adaptor/human/core/gene/` would retrieve the gene adaptor to the *Homo sapiens* core database. Null-argument methods could be called on an adaptor using a single additional path part such as `/list_stable_ids/` to get all the human gene stable IDs from the database, based on the previous example. Methods that took single arguments could be called by supplying two additional path parts, for example `/fetch_by_stable_id/ENSG00000139618/` to retrieve the BRCA2 gene object. It is also possible to call methods on the retrieved object such as `/description/` to list the gene description, or methods that take additional arguments such as `/slice/name/` to get the name of the corresponding slice (see Appendix 2.2 for URI examples).

The RESTful web services framework was tested under Ubuntu 11.04 in both the Mozilla Firefox and Google Chrome web browsers. Perl version 5.10.1, Python version 2.7.1 and Ruby version 1.8.7 were used to run the example scripts from the Linux Terminal emulator. The *Homo sapiens* genome data were used in the examples described in this chapter, however, a wide variety of disparate organisms were used during the testing stage, to ensure no unexpected behaviour.

Comparison of the different methods of access to Ensembl data

In order to compare the different methods of access to Ensembl data a number of benchmarking test scripts were implemented to determine the efficiency of each method. The time to execute the script (measured in UNIX wall clock time) and memory usage (measured in total allocated bytes) were recorded. A "*simple dataset query*" was designed to retrieve the full list of gene IDs for *Homo sapiens* and a more "*complex dataset query*" was designed to retrieve all protein coding gene sequences in FASTA format for *Saccharomyces cerevisiae*. The comparisons consisted of testing a bash script executing a raw MySQL query; a Perl script using the Ensembl BioMart API, a Perl script using the Ensembl Perl API, a bash script calling the alpha Ensembl REST API, a pyEnsemblRest Python script calling the public Ensembl REST API, and a Python script using the PyCogent Ensembl API. The code for these benchmarks is available in https://github.com/gawbul/gcat/support_files/ensembl_benchmark.

pyEnsemblRest

A more comprehensive Python wrapper was developed around the public beta release version of the Ensembl REST API built by Ensembl (Yates *et al.*, 2014), which I called pyEnsemblRest. This software was developed using Python version 2.7.5 and is freely available under the GNU GPLv3 license. The source code is available to download at https://github.com/pyOpenSci/pyEnsemblRest.

Results

GCAT

GCAT is able to wrap any part of the Ensembl Perl API and therefore access the full extent of data available in the Ensembl genome databases. I developed functionality to retrieve commonly analysed genomic components such as CDSs, UTRs, introns, repeat types, and orthology groups. In addition GCAT can provide information on the start and end positions and hence length, relative order and ordinal position of genomic features. GCAT can also return information on nucleotide bias, relative rate, dN/dS etc. via script plugins. I have a comprehensive range of plotting and summary statistics available due to GCAT's easy integration with R (including the plyr, dplyr, reshape2, and ggplot2 packages). Together these functions give many diverse possibilities to biologists carrying out comparative genomics. Example workflows of only a few of the available features were implemented using the GCAT pipeline's functionality. These involved investigating the structure of genes in the widely used model organism *Mus musculus*; analysing the repeat content in the Great Ape genomes, and reporting on the intron frequency distribution of all 52 genomes available in the Ensembl as of February 2011.

Description of gene features in Mus musculus

To demonstrate GCAT's functionality, details of the gene structure for the house mouse, *Mus musculus*, were retrieved from Ensembl using a custom workflow (see `example-gene_structure-workflow.txt` in the examples directory on GitHub). We identify the 5'- and 3'-UTR lengths, coding region length, and total intron length for 19,327 genes. GCAT allows extensive data filtering, allowing us to retrieve only the 88.3% (19,327) of the annotated protein coding genes in the *Mus musculus* genome annotated with both a 3'- and 5'-UTR. Similarly we could filter to exclude especially small or large CDS sizes or by other genome feature criteria. GCAT, by utilizing R, can output a broad range of descriptive statistics on these data, such as the mode sizes of each group (101 bp, 684 bp, and 130 bp for the 5'-UTR, coding region and 3'-UTR lengths respectively) and additionally creates highly customisable plots of the different regions of the gene (see Figure 2.2).

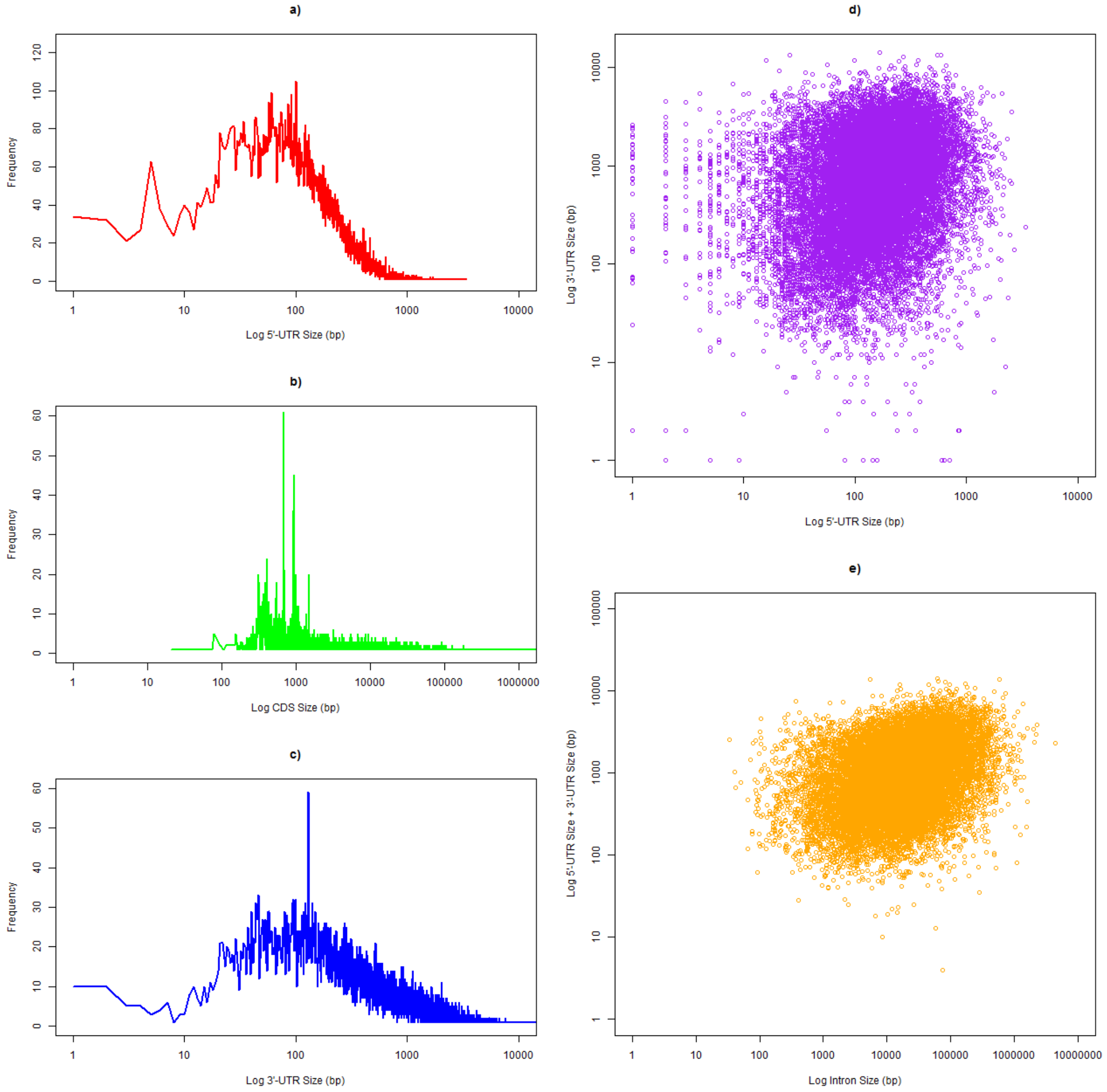


Figure 2.2 - Plots showing the frequency distribution of common gene structure components in 19,327 protein coding genes of the house mouse, Mus musculus. a) Frequency distribution plot of 5'-UTR length. b) Frequency distribution plot of coding region length. c) Frequency distribution plot of 3'-UTR length. d) Scatterplot of 5'-UTR length vs 3'-UTR length. e) Scatterplot of combined UTR lengths vs intron length.

Classification of repeats in the Great Apes

A GCAT workflow (see `example-repeats-workflow.txt` in the examples directory on GitHub) was used to identify between 6,321,250 and 7,077,761 repeat elements in *Homo sapiens* and *Pongo abelii* respectively (see Table 2.1). These repeats account for 1,513,523,923 bp (*Homo sapiens*) to 1,742,400,866 bp (*Gorilla gorilla*) of the overall genome sequence, or between 47.85% and 58.73% in *Pongo abelii* and *Nomascus leucogenys* respectively.

GCAT can additionally break the sequences down into different classes of repetitive element, as identified by the RepeatMasker (Smit *et al.*, 2010) program, by parsing its output (see Figure 2.3). Ensembl also utilises RepeatMasker, so it is possible to use GCAT to retrieve these data directly via their API (Potter *et al.*, 2004), though by providing a wrapper around RepeatMasker one can also undertake their own replications to validate their data. It is then possible for us to use GCAT's functionality to visualise the proportion of the genome that is contained within the respective classes of repetitive elements and make inferences on the processes that resulted in this structure. We can then investigate the relationships in more detail, by outputting these data in CSV output and utilising GCAT's interface with the R statistical language in order to undertake more complex statistical analyses to test our hypotheses.

Table 2.1 - A summary of the repeat elements retrieved using our example repeat elements workflow for the five available *Hominoidea* genomes.

	Homo sapiens	***Pan troglodytes***	***Gorilla***	***Pongo abelii***	***Nomascus leucogenys***
Genome Size (Gbp)	3.10	3.30	3.04	3.44	2.93
Number of repeats (millions)	6.32	6.88	6.71	7.08	6.49
Total genomic repeat length (Gbp)	1.51	1.72	1.74	1.65	1.72
Total genomic repeat percentage (%)	48.79	51.99	57.30	47.85	58.73

52 genome intron frequency

GCAT allows the retrieval and analyses of data in a large-scale comparative manner. In this case I was able to retrieve and summarise the intron sizes across all 52 genomes available in the Ensembl databases as of February 2011 (see Figure 2.4). This analyses highlighted a relatively uniform distribution of intron sizes in most eukaryote species, with the exception of the unicellular eukaryote *Saccharomyces cerevisiae*, which has had its intron content widely studied (Wolfe and Shields, 1997; Spingola *et al.*, 1999; Bon *et al.*, 2003; Neuvéglise *et al.*, 2011; Hooks *et al.*, 2014). The rest of the species have a mode intron size of 60 to 120 bp, with the majority of introns being <200 bp in length. The intron distribution is generally unimodal with a monotonically decreasing slope from 60 to 120 bp onwards. However, in the sea squirt *Ciona intestinalis* and the zebrafish *Danio rerio*. There is also a second, less pronounced peak of intron size at ~150-450 bp in the sea quirt and 500-2,000 bp in the zebrafish.

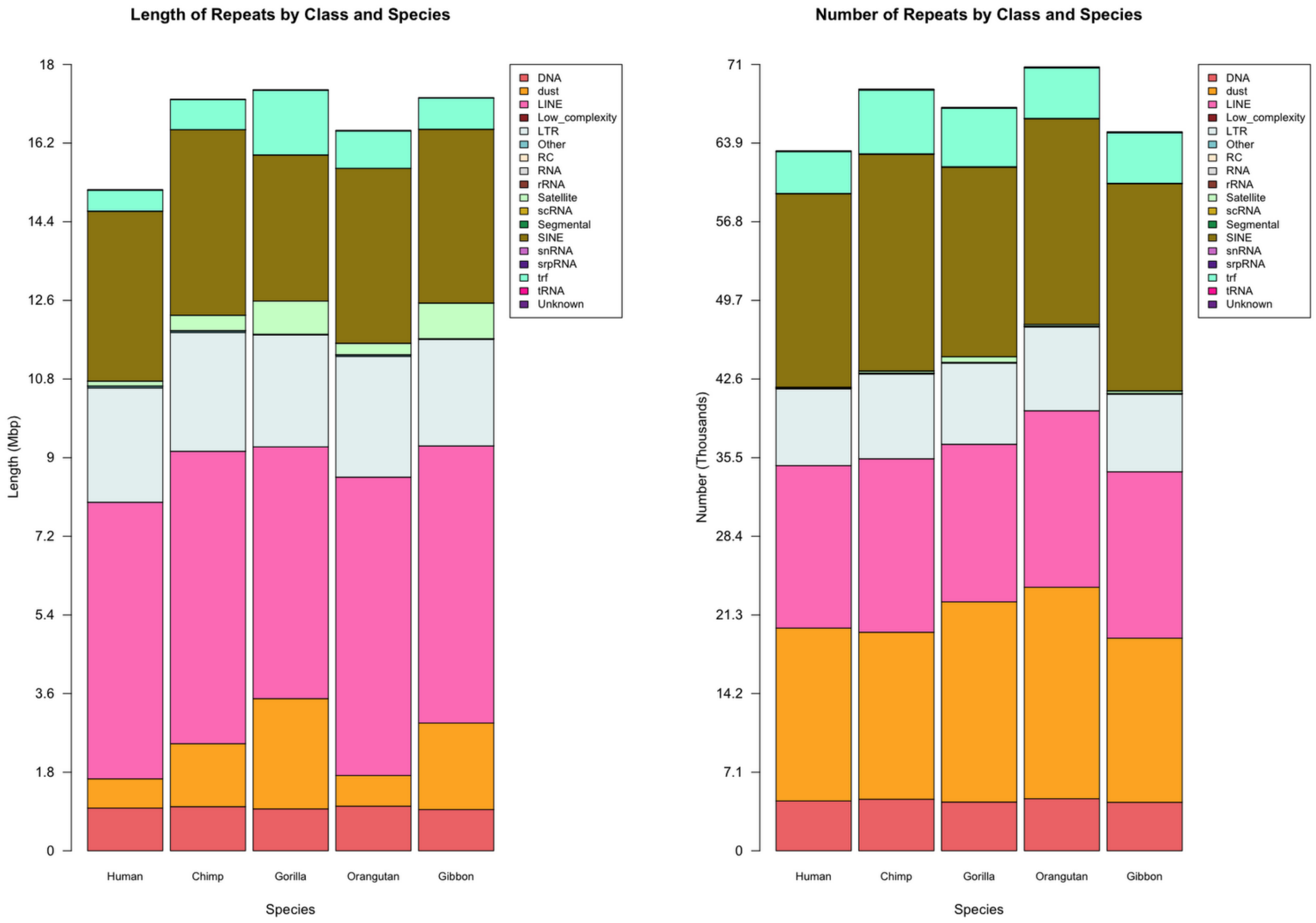


Figure 2.3 - a) Classes of repetitive element by length across the genomes of five primate species. b) Classes of repetitive element by number across the genomes of five primate species.

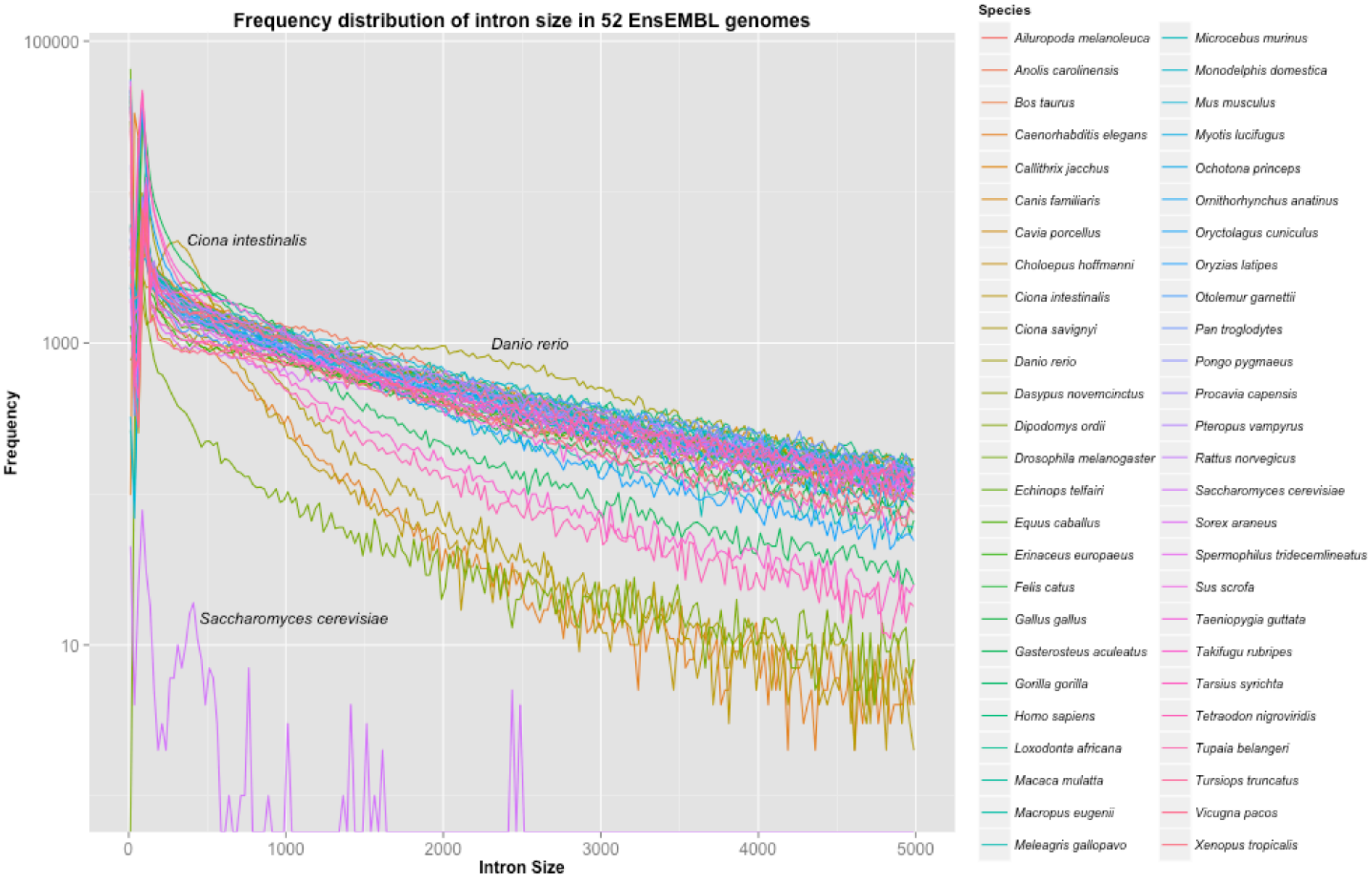


Figure 2.4 - Frequency distribution of intron sizes in the 52 genomes available in the main Ensembl genome databases as of February 2011. Interesting and unexpected differences are highlighted in the Sea Squirt *Ciona intestinalis* and the Zebrafish *Danio rerio*.

REST

A fully working prototype of the Ensembl RESTful web service was implemented allowing for programmatically independent access to the main Ensembl Core API functionality. The Ensembl Compara (comparative genomics), Functional Genomics (transcriptional regulation) and Variation (polymorphisms and structural variants) APIs were also available to the web service, but their functionality wasn't extensively tested.

The API was tested by retrieving the full human gene list and the BRCA2 gene object with endpoints accessed using the Mozilla Firefox web browser. Output was returned in YAML format. Individual scripts were also created in Perl, Python and Ruby that returned the BRCA2 gene object in JSON format and also displayed a 1,000 bp segment of the DNA sequence (see examples in Appendix 2.3). These examples show the power, flexibility, and simplicity of the Ensembl RESTful web service. In comparison to the Ensembl Perl API, retrieving a list of human genes becomes as easy as calling a single URL for example, as opposed to several lines of Perl code (see Table 2.2)

Table 2.2 – Comparison of the different methods of accessing the Ensembl API. a) Example code in the Perl programming language, using the Ensembl Perl API, to return a list of human gene IDs. b) Example URL endpoint for accessing the same data using the Ensembl RESTful web service.

a) Example code in the Perl programming language, using the Ensembl Perl API, to return a list of human gene IDs from the Ensembl server

```
#!/usr/bin/env perl

use strict;
use warnings;
use Bio::EnsEMBL::Registry;

# setup registry object and connect to Ensembl server
my $registry = 'Bio::EnsEMBL::Registry';
$registry->load_registry_from_db(
  -host => 'ensembldb.ensembl.org',
  -user => 'anonymous',
  -pass => undef,
  -port => 5306
);

# setup gene adaptor object
my $gene_adaptor = $registry->get_adaptor('Human', 'Core', 'Gene');

# retrieve list of stable gene IDs
my @gene_ids = @{$gene_adaptor->list_stable_ids()};

# output gene IDs to the screen
foreach my $gene_id (@gene_ids) {
      print "$gene_id\n";
}
```

b) Example URL endpoint using the Ensembl RESTful web service to return a list of human gene IDs from the Ensembl server

```
http://localhost:3000/get_adaptor/human/core/gene/list_stable_ids
```

Comparison of the different methods of access to Ensembl data

A comparison of the efficiency (in terms of memory usage and time to retrieve data) of the different methods of access to Ensembl data was undertaken with execution time equal to wall clock time in seconds and memory usage in bytes (see Table 2.3 and Table 2.4). In both cases a) "MySQL" is the use of standard SQL syntax to retrieve data directly from the Ensembl databases, b) "BioMart" is the use of the existing BioMart Perl API to retrieve data from the Ensembl databases, c) "Perl API" is the use of the existing Ensembl Perl API to retrieve data from the Ensembl databases, d) "Alpha REST API (curl)" is the use of a novel alpha version of the Ensembl REST API to retrieve data from the Ensembl databases, e) "Public REST API (pyEnsemblRest)" is the use of a novel wrapper script around the existing public Ensembl REST API to retrieve data from the Ensembl databases, and f) "PyCogent" is the use of the existing PyCogent API to retrieve data from the Ensembl databases.

Table 2.3 – Simple dataset query execution time and memory usage.

	Execution time (seconds)	Memory usage (bytes)
MySQL	0.995	5,768
BioMart	6.699	550,300
Perl API	23.353	44,096
Alpha REST API (curl)	2.120	3,968
Public REST API (pyEnsemblRest)	N/A	N/A
PyCogent	1.275	41,448

Table 2.4 – Complex dataset query execution time and memory usage.

	Execution time (seconds)	Memory usage (bytes)
MySQL	N/A	N/A
BioMart	11.123	552,012
Perl API	562.205	69,776
Alpha REST API (curl)	4,237.392	54,028
Public REST API (pyEnsemblRest)	461.318	74,528
PyCogent	809.871	108,412

Discussion

This chapter describes the development of open source bioinformatics pipelines for the comparative analysis of Ensembl genome data. The GCAT software can utilise the genomic features annotated by Ensembl and integrates powerfully with the R statistical language to implement a large range of biologically relevant analytical output and visualisations. Its design enables users to easily add modules for bespoke features in addition to those provided in the package and together permit extensive comparative analyses of genome structure and content.

GCAT

GCAT is developed specifically to undertake large-scale comparative analyses of genome data available from the Ensembl genome databases. It provides an all-round solution for the retrieval, analysis, description and visualisation of multiple genome annotations with a particular focus on understanding genome evolution. In this chapter I document the design, implementation and key attributes of GCAT, and additionally demonstrate a few of its features by employing the pipeline to carry out comparative genomic analyses of a number of datasets from Ensembl. These analyses highlight some interesting biological trends that warrant deeper analyses.

Description of gene features in Mus musculus

GCAT can also be used to highlight errors in annotation that require further investigation. In Figure 2.2 above; in addition to detailing the genome-wide structure of the genes in the common house mouse, I highlight annotated UTR regions as small as a single nucleotide. If carrying out an analysis of UTRs one might wish to re-validate these UTRs before including them in a larger dataset of more typically sized UTRs. There is no certainty that the features annotated by Ensembl are homogeneous from a single biological class, and further investigation of this idea would be a wise first step in any statistical analysis. For example, one might decide to test the nucleotide composition bias (GC:AT) of 3’UTR regions. A bimodal distribution of these values could indicate that at least two different classes of 3’UTR had been annotated the same way, and different analytical designs would be required to investigate the true nature of their variation. GCAT is especially powerful at accomplishing these types of data exploration and hypothesis generating analyses.

Classification of repeats in the Great Apes

Ensembl automatically executes RepeatMasker (Smit *et al.*, 2010) as part of its annotation pipeline (Curwen *et al.*, 2004; Potter *et al.*, 2004) to identify repetitive elements within the genes or intergenic regions of a genome, allowing it to associate this information with the relevant sequences in its database. GCAT can retrieve these data, analyse them and produce high-quality visualisations by calling a simple workflow script, which is included in the pipeline's example documentation. Using this workflow it is possible to retrieve a comprehensive repeat element dataset for *Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla*, *Pongo abeli* and *Nomascus leucogenys*; the five apes (superfamily Hominoidea) available in Ensembl.

In Table 2.1 a descriptive focus is taken to reporting on the repetitive content of the genomes of the Great Apes, however by breaking the repeat datasets down into their component annotations it is possible to produce visualisations that highlight the differences in the datasets more easily (see Figure 2.3). This exploratory analyses of repeat content was able to highlight similarities in Alu content between the genomes of these apes that had previously been dismissed (Locke *et al.*, 2011). Locke *et al* identified approximately 250 lineage-specific Alu retroposition events in the Orang-utan genome, which was much lower than that of the other sequenced primates, including humans. They used Allele-Specific Alu PCR to amplify the Alu insertions from novel Orang-utan genomic DNA, followed by BLASTZ (an independent implementation of the Gapped BLAST algorithm designed for local alignment of two long genomic sequences) to identify the repeat types. This is in contrast to the method used here, which although is built upon the same genomic DNA sequence, has undergone re-annotation and additional validation via the Ensembl annotation pipeline (Curwen *et al*., 2004; Potter *et al*. 2004). My method also uses the RepeatMasker program (which is based on BLASTZ) to identify repeats across the entire genomic DNA sequence, followed by mining of the individual elements using GCAT's integration with R.

Although further, phylogenetically controlled analyses are required to determine the validity of the Orang-utan Alu conclusions, they corroborate the findings of other studies that highlight areas where the Orang-utan genome is more similar to the genome of Humans than are Chimpanzees (Hobolth *et al.*, 2011) and also lend to further investigation of claims of ancient lineage specific expansions of Alu elements in the Orang-utan genome (Walker *et al.*, 2012).

52 genome intron frequency

Comparative genomics requires more than reporting simple summary statistics, stating for example the mean and mode sizes (Moss *et al.*, 2011). GCAT allows the exploration of more sophisticated patterns in the data and its power in large-scale whole-genome analyses is demonstrated with the retrieval and visualization (see Figure 2.4) of the entirety of the intron size classes in the 52 genomes available in Ensembl's Core database as of February 2011 (release 61). This analysis not only highlights key features of the distribution of intron sizes across a divergent range of species, but also pinpoints some meaningful differences in the distributions between the species, in addition to expected outcomes.

The mode intron peak is clearly visible at between 60 and 120 bp in all species with the majority of data points existing in this region. The distribution is right skewed however, with a number of lower frequency data points towards the maximum intron size classes for each species. The distributions are, largely, monotonically decreasing from tens of thousands of introns at the mode, too often only 1 or 2 in the larger size classes. Figure 2.4 has a cut-off at 5,000 bp as an upper limit on its x-axis, however it isn't uncommon to have introns ranging to hundreds of thousands or millions of base-pairs in size. The biological reason behind this right skewed distribution isn't clear, but it is likely that the metabolic burden of removing the larger introns from these genes has contributed to this as a result of purifying selection. The requirements to remove these elements in terms of ATP utilisation by the cell would have been significant and an unnecessary demand on the early eukaryote (Wagner, 2005; Castillo-Davis *et al.*, 2002). However, as many of these species have relatively small effective population sizes and longer generation times, the ability of selection to remove these from the population is greatly reduced (Lynch, 2002).

Some interesting anomalies also present themselves in the form of a bimodal distribution in the Sea Squirt *Ciona intestinalis* and the Zebrafish *Danio rerio*. The reason behind this requires further detailed comparative analyses. As of February 2011 only two Sea Squirt genomes existed, in contrast to five Teleost fish genomes. This analyses is taken further in Chapter 3 (Moss *et al.*, 2011) by comparing the genomes of the five available teleost fish in order to understand the cause of these differences in greater detail. The outlier, and expected result in this case is the budding yeast *Saccharomyces cerevisiae*, which exhibits the lowest frequency and smallest of intron sizes, confirming previous findings (Spingola *et al.*, 1999). *S. cerevisiae* has been extensively studied, but annotation error still likely exists here, with higher frequency peaks in intron size being seen at regular intervals along the x-axis. These may well be real data, but would require strict validation before further analyses were undertaken.

While comparative genomics is held to be increasingly important, the tools for its implementation are currently poorly adapted. The GCAT pipeline was designed to overcome this and specifically to exploit the rich annotation provided by Ensembl. A large number of tools for data exploration, visualisation and quantitative analysis are already incorporated, although its open design allows the addition of other analytical modules without difficulty. The use of this library however, requires knowledge of programming principles and in particular knowledge of the Perl programming language. This creates a barrier the usage of GCAT in comparative genomics analyses for those scientists that lack prior programming experience. The development of the Semantic Web and programmatically independent access to data resources has been discussed as a means of overcoming these issues, and additionally in improving the integration and sharing of data worldwide (Stein, 2002; Stein, 2008; Stein, 2010).

REST

The development of programmatically independent data access is of the utmost importance, not just in the scientific arena, but for any organisations dealing with large volumes of data. A uniform format or database schema for the distribution and storage of data would aid the realisation of this greatly. This would, however, require a great deal of collaboration and additionally a considerable change in the organisation of the existing data assembly and annotation pipelines, not to mention the conversion of existing data. By building an abstract layer on top of the existing infrastructure, such as those provided by web service frameworks, one can retrieve these data with limited additional investment in time and resources.

The development of a prototype RESTful web service framework, built using the REST architecture standards, providing a simple, but powerful means of accessing Ensembl's genome data, provides an ideal means of non-programmatic data access for the wider scientific community. Being able to access these data over standard HTTP via a web browser, independent of any programming language, makes the information much more accessible, especially to those researchers with limited or no programming experience. This can, in some cases, alleviate the need for dedicated bioinformatics support, in smaller scale studies.

The aim of developing a REST API was not to eliminate the need for bioinformatics support, but to complement them. By enabling wider, less stringent access to genomics data and their annotations, it can increase the throughput of a research project considerably. Rather than having to learn a complex API, one can simply learn the different path parts required to construct a URL to retrieve the data they need. As the majority of scientists will at least have a basic knowledge of using the World Wide Web, through an Internet browser, this would require very little training. Additionally, those that have a knowledge of one or more programming languages, will have a reduced overhead in learning how to access the data using standard HTTP libraries.

One of the issues that was highlighted in the testing process, was memory management by the browser, when executing particularly large data queries. The browser would first download the data into memory, before rendering it to the browser window. When the data received exceeded several hundred megabytes and certainly above one gigabyte, this would cause considerable performance related issues. In some cases the browser, or computer would become unstable and require a restart. Being able to cache queries to reduce this overhead would be preferable. Of course, most users wouldn’t need to execute large queries via the browser and it is likely, in fact, that larger queries would be run via a terminal based script, integrated into an analysis pipeline. However, performance and efficiency of the web service is still a concern and would require some investment of time in order to ensure both the user and the server don’t experience any unnecessary latency.

Issues such as these were highlighted during the prototype development and it was clear a number of additional steps would be required before the web service was functional enough for use by the wider public. These included caching of the data locally, implementation of separate base parts for the controller, adaptation of the `Bio::Ensembl:Registry` module to return an object rather than a hash reference, and ensuring optimal security considerations. Registration of an API key or standard rate-limiting steps to limit the number of queries was also determined to be necessary, in order to ensure no inadvertent denial of service by particularly high-throughput users. These considerations were addressed in the development of the public Ensembl REST API (Flicek *et al.*, 2012; Yates *et al.*, 2014).

Comparison of the different methods of access to Ensembl data

In order to compare the different methods of access to Ensembl data, particularly from the perspective of efficiency of the web service implementations in relation to programmatic APIs, a number of scripts were developed (see Implementation). A baseline raw SQL query was designed to retrieve the data with minimal overheads in each case. Comparisons were made using a simple and more complex query. The "simple query" involved retrieving all the gene stable IDs for *Homo sapiens*. The "complex query" was to retrieve all protein coding gene sequences for *Saccharomyces cerevisiae* in FASTA format.

At the time of analysis the public REST API didn't have a function for retrieving lists of gene stable IDs and hence this benchmark could not be run. The complexity involved in writing the SQL for the complex query made it unfeasible to implement. This is because there is a great deal of normalisation of the Ensembl databases to reduce redundancy by splitting fields across multiple tables and defining a number of interwoven relationships. The sequences themselves are stored as unassembled contigs that require recursive functions and multiple selects to join all the individual coordinates together, whilst also taking into account the orientation of the strand, in order to compile the full gene sequence. This is beyond the scope of the structured query language (SQL).

Simple query

It is clear from these comparisons (see Table 2.3) that the most efficient method of access to the Ensembl data is through the construction of raw SQL queries (0.995 seconds). Surprisingly this is only marginally faster than the PyCogent Python API (1.275 seconds), with the alpha REST API taking twice as long again (2.120 seconds). The BioMart API still performed relatively well (6.699 seconds), with the Perl API taking the longest at 9.976 seconds. The Perl API taking almost 10 times longer than the raw SQL query was very surprising. Running `Devel::NYTProf` on the code showed that this time was spent on displaying the gene IDs to STDOUT via the `main::CORE::print` method and in the `DBI::st::execute` method. A total of 197 individual calls were made to the latter method, which is likely where SQL statements have been prepared and subsequently executed against the backend database. The overheads involved in setting up the connections to the Ensembl servers and performing initial caching to populate the database adaptor objects seem to take their toll for smaller queries that can be executed in a fraction of the time by other methods. The benefit of caching is that on subsequent runs of the code, this runtime improved by nearly 50%.

In terms of memory usage the most optimal strategy for retrieving data was via the alpha REST API. The small memory footprint here (~3.9Kb) is likely due to the use of the `curl` program to submit the `GET` request to the REST web server. `curl` is a terminal based UNIX package developed in the C programming language and can therefore have greater memory control via lower level system calls. The memory footprint of the REST web server is likely to be much greater, as this does all the leg work, but as this was a test of standalone methods for retrieval of Ensembl data, those measurements weren't taken into account. The worst performer here was the Perl script utilising the BioMart API, which used ~550Kb of memory. This was determined to be due to the numerous caching steps undertaken on the first run of the script. Additional runs of the script load the cached data from locally stored XML files into memory to improve execution and retrieval of data. This has the benefit of improving the runtime over the Ensembl Perl API, but at the cost of extra memory utilisation.

Complex query

The complex query involved retrieving all the protein coding gene sequences for *Saccharomyces cerevisiae*. Due to database normalisation it wasn't possible to retrieve these data using a single SQL query. This is because Ensembl stores the raw unassembled sequences in its database, and assembles the appropriate fragments into the contigs that form the relevant gene sequence via a number calls to the Core Perl API that abstracts some of that complexity away from the end-user. It is expected that this would have been a relatively fast method of accessing the Ensembl data, with low memory consumption in particular. Due to the overheads involved in building a shell script that would call the individual queries involved in assembling the sequences via the MySQL command line client, however, this would have been unfeasible to assess and compare in the same manner as the other methods, introducing an unfair bias to the benchmark. The MySQL benchmark was therefore not included for the complex query (see Table 2.4).

In contrast to the simple query, the most efficient method of accessing Ensembl's data when constructing more complex queries was the Ensembl BioMart API (Kinsella *et al.*, 2011) clocking in at 11.124 seconds. It is expected that raw MySQL queries would have performed at least as well as this, though due to the caching steps undertaken by the BioMart API, additional queries would likely have performed even better. This is because they would have been loaded from the local disk, and therefore not be subject to network latency. A simple method of reducing network latency is to install a local copy of the raw data in a local MySQL server installation, however, there is still additional overhead involved in making a loopback network connection, in comparison to reading from disk. These kinds of differences weren't assessed, though it is expected that there would be only millisecond variations in the overheads involved. This can add up considerably over 10s of thousands of sequences, however. The PyCogent API, which performed very well in the simple query, fell behind the other methods at 809.871 seconds in this benchmark. This is likely due to inefficient query construction and a lack of appropriate caching steps, though the performance of the Python `SQLAlchemy` library was also highlighted as a possible cause during profiling. The Perl API and Pubic REST API performed approximately the same, with slightly better performance being seen via the Public REST API. The greater performance of the latter is no doubt due to caching steps and optimisations undertaken server side, including load balancing provisions. The worst performing at more than 5X slower than the PyCogent API, was the alpha REST API. This discrepancy in the performance of the alpha REST API can be put down to the lack of local caching and server side optimisations.

In terms of memory usage there isn't a great deal of difference from the simple query. The increase in memory utilisation doesn't appear to correlate with the increase in data returned, with 64,785 gene IDs with an average length of 14 characters returned via the simple query, and 6,692 DNA sequences with an average length of 1,369 bp returned in the complex query. The largest increase was seen via the alpha REST API with a ~13.5X increase in memory utilisation, followed by a ~2.5X increase with the PyCogent code. The increase in memory utilisation with the alpha REST API was unexpected, as the data retrieval was undertaken using the `curl` command line application, which is particularly light weight. This may have been due to buffering of data before being written to disk, or oversights in the `curl` program code with respect to the proper release of memory. The increase in memory utilisation with PyCogent is likely due to the latter. The Perl API and BioMart methods were about the same as the simple query approximately 70Kb and 550Kb usage respectively. This highlights that memory utilisation is relatively well optimised in these code, though there is still some room for improvement. The Public REST API couldn't be contrasted with the simple query usage, though it performed relatively well in comparison to the other methods used in the complex query execution.

One would expect that as additional software layers are added; memory usage and access times would increase. This is understandable, as method calls must be translated through the respective layers to the lowest level database queries, all of which have their own overheads in terms of use of time and memory. This isn't always the case, however, as we see with the implementation of caching steps in some of the methods that instead retrieve some data from disk, or preload some objects into memory. For simple queries the differences are negligible, but for more complicated queries, especially for those needed in larger-scale comparative genomics studies, these allow for valuable improvement in runtime and memory utilisation. In general, the most efficient method must be chosen, which will in most cases result in the decision to use the least time consuming approach, though this depends on the available infrastructure. One might find that a workflow can be parallelised to reduce runtime by distributing across multiple machines therefore perhaps favouring optimal memory usage, though there should always be a focus on optimising code to achieve maximum performance on a single node before considering distributing across many.

Conclusions

It is clear that there is a real need for wider access to genomic sequence data. The creation of programming language independent APIs by large, data hungry web platforms such as Twitter in order that users, including academics, can efficiently mine their data reinforces this position (Lin and Ryaboy, 2013). Having access to such web service frameworks, in addition to the existing, powerful methods of programmatic access provided by data providers, can only serve to strengthen the open-access nature of scientific data. If designed and implemented effectively, with the lowest level of abstraction possible; web service frameworks represent a significant solution to data access problems. These sort of services can be run both on a standalone computer, or scaled out across a server farm or significant Cloud-based platform, allowing for even greater performance. The case for the development of this kind of bioinformatics infrastructure is already well defined (Stein, 2002; Stein, 2008; Stein, 2010). By allowing scientists from a broader range of backgrounds to be able to retrieve, analyse and make inferences from genomic data, it can only improve our understanding of genome biology and aid future scientific discovery.

Of course, improving code standards and training for researchers is an additional area that must be focused on, and the push to improve computational literacy and knowledge of computer science *en masse* as part of government initiatives is a welcome trend (Berges *et al.*, 2013; Golberg *et al.*, 2013). In general more time needs to be invested and considerations given to the development of software in the sciences. The development of software as a method for experimentation must be held in the same regard as wet lab experiments, which are subject to much scrutiny through robust experimental design. Training in the development of software and understanding of computational methods for data analyses should be as important as gaining a thorough understanding of the scientific method and associated statistical methodologies used in data analyses. There are a number of different groups focusing on these areas already that are making good progress towards ensuring that code is produced to a high standard in the sciences (Dubois, 2005; Dudley and Butte, 2009; Noble, 2009; Wilson *et al.*, 2012; Petre and Wilson, 2013; Crusoe and Brown, 2013; Wilson, 2014). In the meantime, tools like GCAT and the development of RESTful web service frameworks such as the Ensembl REST API will provide much needed wider access to Ensembl data in particular that reduces the learning curve considerably.

[This page is left intentionally blank]

CHAPTER THREE: COMPARATIVE ANALYSIS OF TELEOST GENOME SEQUENCES REVEALS AN ANCIENT INTRON SIZE EXPANSION IN THE ZEBRAFISH LINEAGE

Introduction

Introns are a major component of metazoan genomes, comprising ~24% of the human genome compared to only 1.1% for exons (Venter *et al*., 2001). Even in species with genomes considerably smaller than humans, and representing taxonomically diverse lineages, introns can account for a substantial proportion of the genome. The nematode *Caenorhabditis briggsae*, for example, has introns containing 1.3 times as many nucleotides as do exons, which together account for ~30% of the entire genome sequence (Stein *et al*., 2003). Intron sequence in general evolves at a high rate, close to that of fourfold degenerate sites, pseudogenes, and non-coding regions (Hughes and Yeager, 1997; Chamary and Hurst, 2004; Gaffney and Keightley, 2006). Despite this, introns may also contain gene regulatory elements (Majewski and Ott, 2002; Gaffney and Keightley, 2006), and their impact on translation, via alternative splicing, can also be substantial (Mironov *et al*., 1999; Kim *et al*., 2007). Even without the presence of regulatory elements within introns they may still contribute strongly to the deleterious mutation rate. Correct splicing requires the maintenance of specific splicing signals at the start and end of each intron, an interior branch point adenine, and a number of other sequences imperfectly conserved across eukaryotes involved in the recruitment of the spliceosome (Schwartz *et al*., 2008). Together these sequences increase the mutational load of intron-containing genes since mutations in any of the required splicing signals can lead to non-functionalization of the locus. The several hundred thousand introns in a vertebrate genome are therefore a considerable mutational burden and it has been estimated that perhaps a third of all human genetic disorders involve mutations affecting splice-site recognition (López-Bigas *et al*., 2005; Frischmeyer and Dietz, 1999). The study of introns can therefore greatly aid in our understanding of the genome’s mutational dynamics and in the selectively maintained regulation of surrounding coding regions.

There are diverse mechanisms by which introns may be gained including reverse splicing, local duplications, transposable elements, and transfer from paralogs by unequal crossing over (Roy and Gilbert, 2006; Iwamoto *et al*., 1998; Rogers, 1989; Sharp, 1985; Hankeln *et al*., 1997). Subsequent to its origin, introns will change in size due to the accumulation of small insertions and deletions, non-homologous recombination and the action of transposable elements. Repetitive sequences such as transposable elements occupy from 33% to 52% of sequenced vertebrate genomes, and it has been shown that 20% to 60% of vertebrate introns contain transposable elements (Mills *et al*., 2007; Sela *et al*., 2010). Intron frequency and mean intron size are known to vary considerably across animal taxa (Lynch and Conery, 2003; Roy and Gilbert, 2006; Gazave *et al*., 2007; Yandell *et al*., 2006; Zhu *et al*., 2009; Deutsch and Long, 1999) though few investigations have been able to compare the intron composition of entire genomes within and between closely related taxa. There are a small number of previous whole genome studies of introns although these have often been limited to one-to-one comparisons or groups of phylogenetically very divergent organisms (Coghlan and Wolfe, 2004; Yandell *et al*., 2006; Gazave *et al*., 2007; Sharpton *et al*., 2008; Stajich *et al*., 2007; Marais *et al*., 2005; Li *et al*., 2009). A full understanding of the processes shaping intron diversity and evolution will require a large-scale comparative genomic approach making full use of the rapidly increasing number and diversity of whole genome sequences. Such evolutionary comparative genomic studies however are slowed by substantial analytical technical challenges presented to most biologists in dealing with these huge amounts of data. In this study we present a bioinformatics pipeline that can be used to compare the size distribution and content of introns in a comparative genomics study. We investigate the potential of such a genomic approach by comparing introns in the genomes of five teleost fish available at Ensembl (Flicek *et al*., 2010) - the zebrafish (*Danio rerio*), three-spined Stickleback (*Gasterosteus aculeatus*), Medaka (*Oryzias latipes*), Fugu (*Takifugu rubripes*) and Tetraodon (*Tetraodon nigroviridis*). These fish have been used as model organisms in the laboratory for a number of years and a great deal of research has been undertaken focusing on their anatomical and physiological structure (Roest Crollius and Weissenbach, 2005; Haffter *et al*., 1996; Kimura *et al*., 2004; Aparicio *et al*., 2002; Jaillon *et al*., 2004). We feel the information that can be elucidated from their genomes in relation to the biological processes driving or constraining their genomic evolution is therefore of particular interest. Our pipeline (GCAT: Genome Comparison and Analysis Toolkit) has allowed us to characterise, in detail, the composition and diversity of approximately 1 million introns in these teleost genomes and provides a valuable open source extensible platform for comparative genomics of introns and other genomic components.

Materials and Methods

Sequences used

The intron data were retrieved from the Ensembl Core online database, release number 61. The individual fish database versions used were `danio_rerio_core_61_9a`, `gasterosteus_aculeatus_core_61_1n`, `oryzias_latipes_core_61_1m`, `takifugu_rubripes_core_61_4o` and `tetraodon_nigroviridis_core_61_8f`.

Method of access

The data were accessed using a novel bioinformatics pipeline, built using the Perl Ensembl Core Software Libraries (Stabenau *et al*., 2004), along with BioPerl (Stajich *et al*., 2002) and several open-source Comprehensive Perl Archive Network (CPAN) libraries. Some information was verified manually using Ensembls' BioMart website and the Ensembl MySQL databases. The pipeline code is available at http://github.com/gawbul/gcat.

The software pipeline used for this project was developed on an Intel Mac Pro machine running Mac OS X Snow Leopard version 10.6.7. The specifications included dual 2.8GHz quad-core Xeon processors and 8GB RAM. The Perl programming language, version 5.12.3 was used for development. BioPerl 1.6.1 was installed using the CPAN command line client, alongside its dependencies. The release 61 Ensembl Core Perl API was retrieved from the Ensembl website and installed from source following the site instructions. R version 2.13.0 ERROR: requested citation index out of range was used for statistical analysis. The Ensembl Core MySQL databases were installed locally, to reduce network latency, using MySQL version 5.5.11. The Python programming language version 2.6.6 was used for parsing the WindowMasker results. Run-time for all analyses performed was ~8 hours.

Intron sequence retrieval

Intron sequences were retrieved using the canonical transcript for each gene, as defined by the Ensembl Core database. The database and application programming interface (API) are designed in such a way that the intron sequences can only be retrieved automatically via their associated transcript, but because there can be multiple transcripts per gene, this can result in redundant intron data. Introns aren't explicitly defined in the database, and are instead implicitly defined from the exon coordinates by the Ensembl Perl API, and our pipeline was used to automate the intron retrieval process. Since we anticipated that annotation of non-protein-coding genes would vary with genome annotation quality we restricted our analyses to introns in genes matching the biotype ‘`protein_coding`’, which represented greater than 98% of all introns in all fish.

Frequency distributions

The frequency distributions were built for each of the five fish using our pipeline via the `Statistics::Descriptive` CPAN package and plotted using custom-made R scripts. The Comprehensive R Archive Network (CRAN) package gdata, was used to provide functionality for concatenating multiple columns of csv data, but all other calculations were made using novel R code, built on top of the core R functionality. The calculations for the sliding window means and confidence intervals were calculated from a subset of the intron frequency data, consisting of successive 25 bp windows between 1 bp and 5,000 bp. This resulted in 200 points being plotted and reduced any noise due to variation in intron size frequency within each window, but did not affect the overall shape of the distributions.

Determining repeat element content and unique intron size

Ensembl explicitly defines repeat elements, as determined by the RepeatMasker, DUST and TRF software (Smit *et al*., 2004; Morgulis *et al*., 2006; Benson, 1999), as annotation features in its database, and these were retrieved by our pipeline for the canonical transcript of each gene matching the ‘protein_coding’ biotype. We also used WindowMasker (Morgulis *et al*., 2006) to check for repeats, as the quality and coverage of the RepBase repeat libraries (Jurka *et al*., 2005) used by RepeatMasker has previously been questioned (Bergman and Quesneville, 2007). A novel bioinformatics script (see count_wm_repeats.py in the git repository) was developed to parse the WindowMasker results, in order to determine the unique sequence length of each intron by removing its total repeat element length.

Intron position and type

Intron frequencies per gene region (5’-UTR, CDS, 3’-UTR) were calculated according to Ensembl annotations. We corrected for size differences between regions by calculating introns per bond, where the number of phosphodiester bonds in each region is equal to the nucleotide count for the UTRs, and the nucleotide count minus one for the CDS, since UTRs are defined by reference to the CDS coordinates in our pipeline.

Additional to this we calculated the intron type based on explicit splice site nucleotides, matching 5’ GU-AG 3’ and 5’ AU-AC 3’ for the U2 and U12 intron categories respectively. Any introns not matching these definitions were placed in an ‘other’ category.

Results

Intron retrieval and characterization

Table 3.1 presents intron size and frequency data as provided by Ensembl. We retrieved between 185,494 (*O. latipes*) and 221,589 (*D. rerio*) introns per genome totalling 982,544 introns between these five fish. The smallest total intron lengths were those of the pufferfish *T. rubripes* and *T. nigroviridis* at 90,447,562 bp and 108,524,412 bp respectively. *D. rerio* has the largest at 622,476,590 bp as well as the lowest intron density of all the teleost genomes with 8.93 introns per gene compared with 9.80 to 10.51 introns per gene for the other four fish. Despite this, at 622 Mb of intronic DNA, *D. rerio* has from 2.8 to 6.9 times more intronic sequence than the other fish.

Table 3.1. The summary statistics for the five teleost fish. We include total genome size, total number of genes and total number of transcripts, but our study focuses on the introns found within the genes matching Ensembl's 'protein_coding' biotype.

	Danio rerio	***Gasterosteus aculeatus***	***Oryzias latipes***	***Takifugu rubripes***	***Tetraodon nigroviridis***
Genome size	1,412,464,843	461,533,448	868,983,502	393,312,790	358,618,246
Number of genes	32,312	22,456	20,422	19,388	20,562
Number of transcripts	51,569	29,245	25,397	48,706	24,078
Protein coding genes	24,803	20,109	18,920	17,876	18,872
Canonical transcripts	24,803	20,109	18,920	17,876	18,872
Introns per gene	8.93	9.93	9.80	10.51	9.96
Number of introns	221,589	199,624	185,494	187,962	187,875
Maximum intron length	378,145	175,269	295,125	93,537	631,227
Total intron length	622,476,590	151,619,269	219,591,667	108,524,412	90,447,562
Mean length	2,809	760	1,184	577	481
Median length	984	219	252	143	118
Mode length	84	85	77	78	76
25th Percentile length	138	104	90	84	80
75th Percentile length	2,563	615	1,026	450	350
GC Content	50.58%	50.48%	47.10%	40.39%	49.21%
Percentage of genome	44.07%	32.85%	25.27%	27.59%	25.22%

Frequency distributions of teleost intron size

Figure 3.1a shows a frequency plot of intron size class in all five fish, with 5% and 95% confidence intervals. The ordinate is a log scaled count and the abscissa represents the mean of 25 bp sliding windows of intron size. We observe a change in the shape of the intron distribution in *D. rerio* that is not present in the other fish. The minimum, mode, and maximum intron sizes for each fish are given in Table 3.1. Above the 5,000 bp cut-off in Figure 3.1a the number of instances of each individual size class is very low, causing a great scatter in values, although the trend does not differ.

Repeat element content and unique intron size

The length of repeat elements determined by RepeatMasker (see Methods) ranges from 942,285 bp (*T. nigroviridis)* to 13,406,652 bp (*D. rerio)* comprising between 0.66% (*O. latipes)* and 2.15% (*D. rerio)* of total intronic sequence (Table 3.2). A summary of the subsequent WindowMasker analysis is shown in Table 3.3, giving a breakdown of the repeat elements and the unique intron sizes calculated. WindowMasker calculated between 20,313,082 bp (*T. nigroviridis*) and 291,676,913 bp (*D. rerio*) with from 2.71 to 20.69 repeats per intron. This accounts for between 22.46% (*O. latipes)* and 46.86% (*D. rerio)* of total intronic sequence. We used the WindowMasker results to re-plot intron size frequency, as shown in Figure 3.1a, using the unique intron sequence frequency distributions of all introns after repeat element trimming (Figure 3.1b). We also calculated the frequency of individual and cumulative repeat element sizes within introns of length 500 to 2,000 bp and plotted the mean of 25 bp windows across the intron (Figure 3.2a and 3.2b respectively) in order to highlight the contribution of repetitive elements to those particular intron size classes.

Table 3.2. A summary of the intronic repeat element content of the five teleost fish genomes, using the Ensembl RepeatMasker annotations.

	Danio rerio	*Gasterosteus aculeatus*	*Oryzias latipes*	*Takifugu rubripes*	*Tetraodon nigroviridis*
Number of repeat elements	2,927,753	622,107	446,619	297,226	268,530
Length of repeat elements	13,406,652	1,266,916	1,449,300	1,045,873	942,285
Number of repeat elements per intron	13.21	6.35	2.41	1.58	1.43
Percentage of repeat elements	2.15%	0.84%	0.66%	0.96%	1.04%
Length of unique introns	609,069,938	150,352,353	218,142,367	107,478,539	89,505,277

Table 3.3. A summary of repeat element content in the five teleost fish, determined using the WindowMasker software.

	Danio rerio	*Gasterosteus aculeatus*	*Oryzias latipes*	*Takifugu rubripes*	*Tetraodon nigroviridis*
Number of repeat elements	4,583,943	891,753	1,498,499	591,789	509,271
Length of repeat elements	291,676,913	31,910,164	74,289,913	20,701,619	20,313,082
Number of repeat elements per intron	20.69	4.47	8.08	3.15	2.71
Percentage of intron length	46.86%	21.05%	33.83%	19.08%	22.46%
Length of unique introns	330,799,677	119,709,105	145,301,754	87,822,793	70,134,480

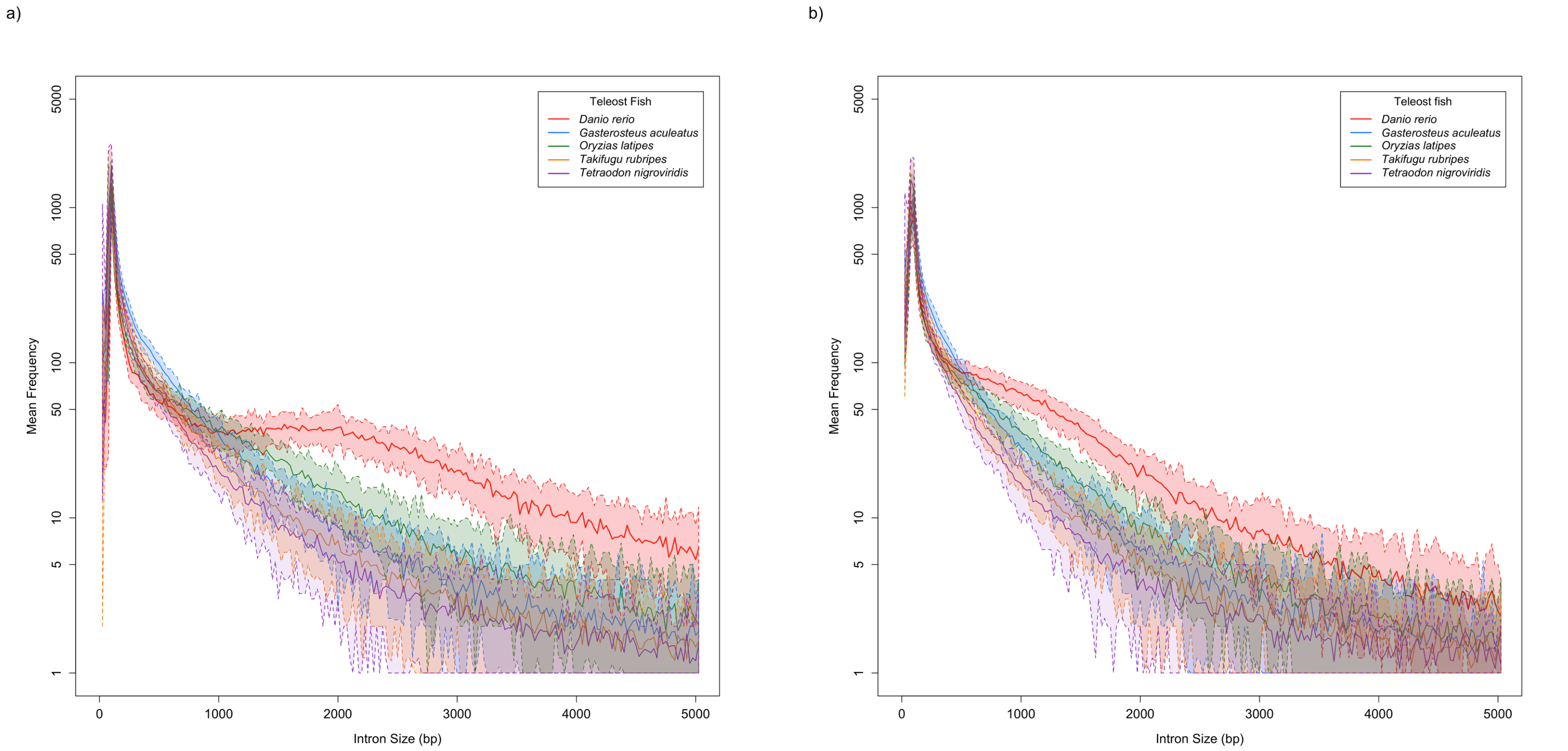


Figure 3.1 a. A frequency distribution plot of intron size in the five teleost fish. Each point represents the mean of intron sizes within a 25 bp sliding window. The lower and upper dashed lines represent the 5% and 95% confidence intervals respectively. All fish present an initial peak of ~80 bp and then decay in a similar pattern, with the exception of *D. rerio*, which has a second peak between ~500 bp and 2,000 bp and subsequently decays parallel to the others. B. A frequency distribution plot of unique intron size in the five teleost fish, representing the intron sizes after removal of repeat sequences.

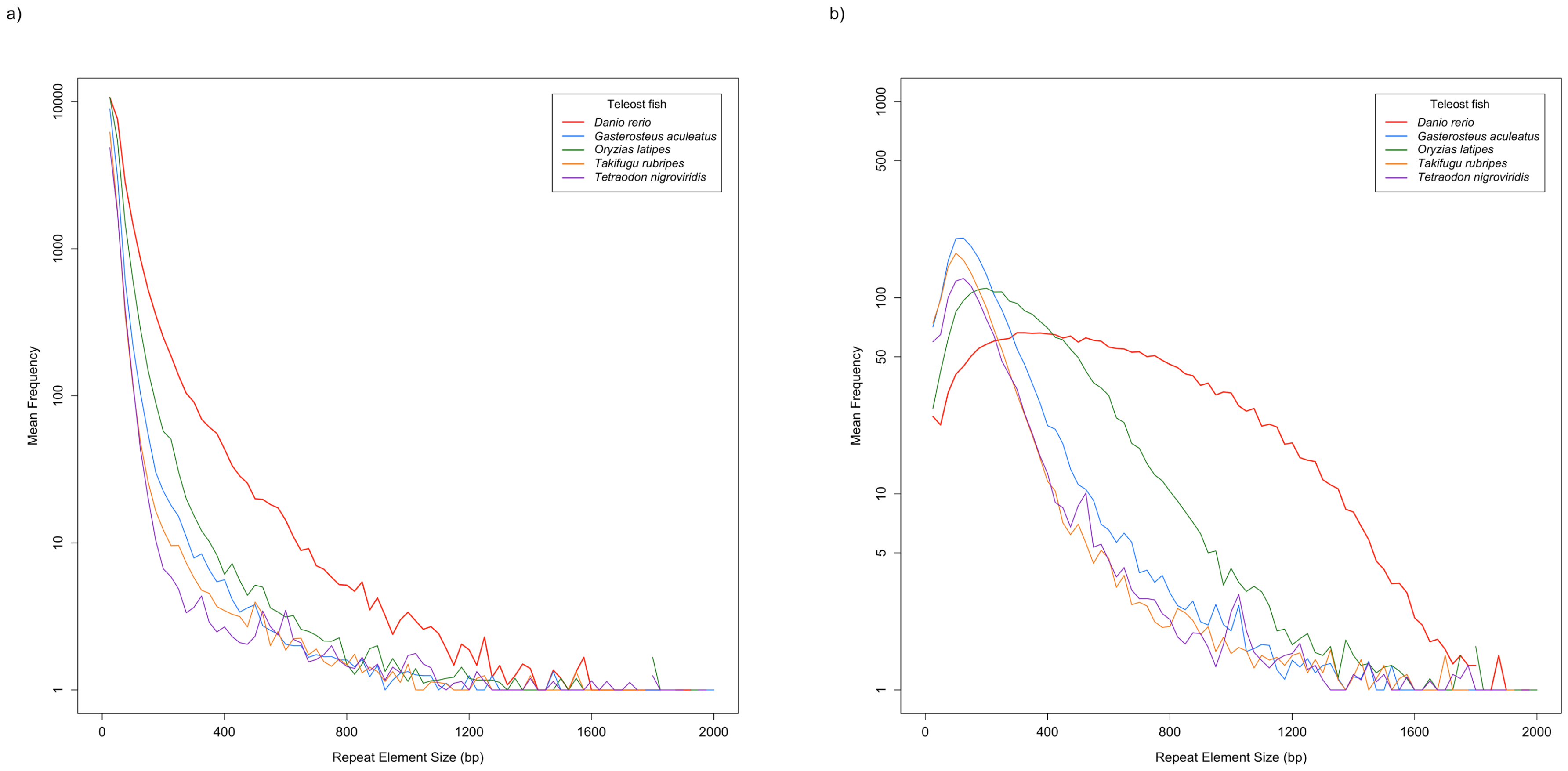


Figure 3.2 a. A frequency distribution of individual repeat element sizes in introns between 500 bp and 2,000 bp in size. Each point represents the mean of intron sizes within a 25 bp sliding window. B. Frequency distribution of cumulative repeat element size produced by pooling all repeat elements within individual introns.

Large introns

The maximum intron size found in each genome is presented in Table 3.1. These are not solitary outliers however, with 1,228 (0.6%) *D. rerio* introns greater than 50,000 bp in size (here referred to as 'large introns' after Shepard *et al* (2009). There are between 16 and 221 introns in the other fish (Table 3.4) accounting for between 0.9% (*T. nigroviridis)* and 17% (*D. rerio)* of total intron length. Our figure for *D. rerio* large introns is different from the 756 reported by Shepard *et al* (2009), perhaps because their data was retrieved from a custom database and represents an earlier version of the *D. rerio* genome. However, our teleost large intron values do fall within the range of 7 (Mosquito) to 3,473 (Human), previously reported for metazoan (Shepard *et al*., 2009).

Table 3.4. A summary of large introns in the five teleost fish. We define this class as those introns of greater than 50,000 bp in length.

	Danio rerio	*Gasterosteus aculeatus*	*Oryzias latipes*	*Takifugu rubripes*	*Tetraodon nigroviridis*
Large intron number	1,228	63	211	16	22
Large intron length	107,485,505	5,353,830	18,210,315	1,000,337	3,568,282
Large intron percentage	17.27%	3.53%	8.30%	0.92%	3.95%

Small Introns

We refer to 'small introns' as those less than 80 bp, which approximates the mode of the pooled teleost dataset. These comprise from 11,473 (*D. rerio*) to 44,755 (*T. nigroviridis*) introns accounting for between 0.12 and 2.97% of the total intronic sequence respectively (Figure 3.3, Table 3.5).

Table 3.5: A summary of small introns in the five teleost fish. We define this class as those introns of less than 80 bp in length.

	Danio rerio	*Gasterosteus aculeatus*	*Oryzias latipes*	*Takifugu rubripes*	*Tetraodon nigroviridis*
Small intron number	11,473	16,415	28,480	35,589	44,755
Small intron length	740,482	906,853	1,818,832	2,283,344	2,685,694
Small intron percentage	0.12%	0.60%	0.83%	2.10%	2.97%

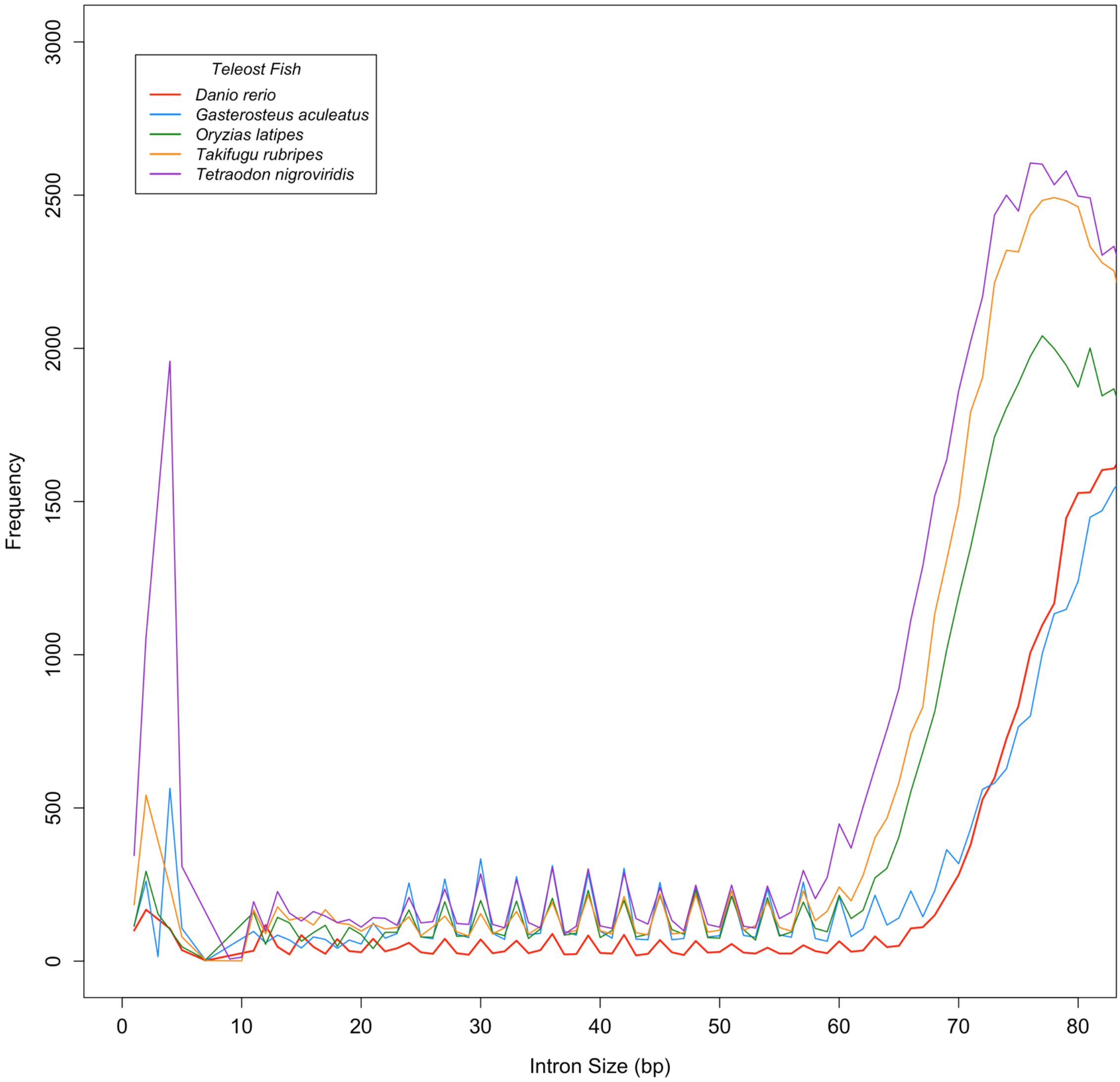


Figure 3.3. Small intron frequency distribution in the five teleost fish showing the 3bp periodicity of peaks between 24 bp and 57 bp.

Intron location

Within protein coding transcripts introns may occur in the coding region (CDS), or either of the terminal un-translated regions (5'-UTR or 3'-UTR). Of those 24,803 *D. rerio* transcripts containing all three regions, 2.08% of introns were in 5'-UTR, 0.57% in 3'-UTR, and 97.35% in the CDS. Similar percentages were found in the other fish (Table 3.6). Correcting for the sizes for these three regions we find 3.4 x10^{-4} introns per bond in the CDS, 1.5 x10^{-4} introns per bond in the 5'-UTR and 0.6 x10^{-4} introns per bond in the 3'-UTR.

Table 3.6. A summary of the intron locations within the pre-mRNA transcripts of the five teleost fish. Introns are classified as either 5'-UTR if they occur before the first nucleotide of the CDS, 3'-UTR if they occur after the last nucleotide of the CDS, or CDS if they fall within this range.

	Danio rerio	Gasterosteus aculeatus	Oryzias latipes	Takifugu rubripes	Tetraodon nigroviridis
5'-UTR introns number	4,611	3,572	2,830	831	1,000
5'-UTR introns length	27,540,112	6,252,772	8,421,615	1,016,292	790,697
5'-UTR introns percentage	2.08%	1.79%	1.53%	0.44%	0.53%
CDS introns number	215,707	195,370	182,309	187,085	186,562
CDS introns length	588,484,433	143,536,499	209,376,669	107,485,857	89,208,382
CDS introns percentage	97.35%	97.87%	98.28%	99.68%	99.30%
3'-UTR introns number	1,270	682	355	46	313
3'-UTR introns length	6,452,045	1,829,998	1,793,383	22,263	448,483
3'-UTR introns percentage	0.57%	0.34%	0.19%	0.03%	0.17%

Splice Signals

This teleost introns dataset contained introns bounded by the typical GU-AG splice signal (U2-type), AU-AC splice signal (U12-type) and those employing other splice signals. *T. nigroviridis*, at 82.51%, has the lowest percentage of typical GU-AG introns and *D. rerio*, at 93.57%, the highest. All fish have a similar number of U12-type introns, with some variation in “other” introns (Table 3.7).

Table 3.7. A summary of the different types of introns in the five teleost fish, as determined by our GCAT pipeline. U2-type introns are defined using the classic GU-AG splice signals and U12-type by the AU-AC splice signals. All other splice signals are placed in to the ‘other’ category. Although this may include some mis-identified introns, the error is likely very small.

	Danio rerio	***Gasterosteus aculeatus***	***Oryzias latipes***	***Takifugu rubripes***	***Tetraodon nigroviridis***
U2-type (major) introns	207,336	170,137	156,282	161,778	155,012
U2-type percentage	93.57%	85.23%	84.25%	86.07%	82.51%
U12-type (minor) introns	184	139	161	105	103
U12-type percentage	0.08%	0.07%	0.09%	0.06%	0.05%
Other introns	14,069	29,348	29,051	26,079	32,760
Other percentage	6.35%	14.70%	15.16%	13.87%	17.44%

Discussion

We have employed a novel comparative genomic pipeline to perform detailed comparison of the intron characteristics of five teleost fish genomes. This allowed us to identify the diversity of intron content and characteristics across the whole genome and to partition these data into biologically relevant categories. Previous approaches to such characterisation have typically either restricted themselves to single comparisons or else incorporated exceptionally divergent organisms (Coghlan and Wolfe, 2004; Yandell *et al.*, 2006; Gazave *et al.*, 2007; Sharpton *et al.*, 2008; Stajich *et al.*, 2007; Marais *et al.*, 2005; Li *et al.*, 2009). Since our bioinformatic pipeline has been designed to build on the high quality genome annotations present at Ensembl, and use open source software libraries such as BioPerl, this approach can be easily integrated into more general studies in comparative genomics. For the analysis of teleost genome data presented here our pipeline has proved itself to be highly automated, yet flexible, fast and to lend itself to evolutionary and statistical approaches to comparative genomics.

Intron Size Distributions

Our characterisation of teleost introns shows that *D. rerio*, the species with the largest total genome size, has more and larger introns than any of the other fish genomes. Although simple summary statistics such as 'average intron length' are commonly applied to the description of a genome's intron content in the literature, these can be significantly influenced by outlier values and miss many of the important differences between taxa. The mean intron length for *D. rerio* is 2,809 bp, yet 50% of all introns are found below 985 bp in length with the modal size only 84 bp. Figure 3.1a also shows the shape of intron frequency for each fish up to intron sizes of 5,000 bp. *Oryzias latipes* has more than twice the mean intron size of *T. nigroviridis* and *T. rubripes*, yet the distribution of intron sizes in Figure 3.1a shows them to be remarkably similar. In contrast to the mean, modal intron size is relatively tightly grouped among these five fish, in the range 76 bp to 85 bp, despite ~150 million years divergence (Benton and Donoghue, 2007) (Table 3.1, Figure 3.1a). For the pooled set of teleost introns the mean size is 1,214 bp (range 481 to 2,809 bp) yet the mode intron size is a mere 81 bp with up to 37% of introns within 20 bp of this mode value. The zebrafish *D. rerio* has a modal intron size only 1 bp different from the stickleback *G. aculeatus*, yet contains 4.1 times as much intronic DNA, an extra 471 Mb. Most introns across fish are small and similar in length, yet introns much bigger than this mode size vary and contribute extensively to the differences between fish. Although 50% of all introns in *D. rerio* are less than 985 bp these account for only 4.8% of all intronic nucleotides.

The comparisons of intron size frequency distributions generated here highlight the unique pattern present in the *D. rerio* genome. The multi-modal distribution we see with zebrafish contrasts with the monotonically decreasing pattern in the other fish (Figure 3.1a). The shape of this curve represents separate genomic processes generating an intron size distribution with a broad peak of ~500 bp to 2,000 bp in addition to the usual teleost ~80 bp mode size.

Our analyses emphasise that over-reliance on simple summary statistics, such as mean or mode intron size, can obscure real biological trends and differences that would be revealed with much more detailed investigation of the distribution of the data as a whole.

Repeat element content as an explanation of intron size differences

Zebrafish has both more and larger introns than the other fish (Figure 3.1a, Table 3.1), accounting for between 402 and 532 million extra nucleotides compared to the other fish genomes. Repetitive elements are known to be the major cause of genome size variation (Mills *et al*., 2007; Sela *et al*., 2010) and we were interested to see if they also accounted for the difference in intron size between these teleosts, in particular the increased intron content of *D. rerio*. We took two different approaches to determine this. The first relied on the annotations available at Ensembl, which uses the RepeatMasker software and compares data against a curated library of repeats using local alignment methods. The standard repeat libraries however may not have optimal quality and coverage for some taxa (Bergman and Quesneville, 2007; Morgulis *et al*., 2006). The second approach used the WindowMasker program, which compares the genome against itself to identify repeats and is therefore independent of previous repeat curation in closely related taxa. It implements the DUST and WinMask algorithms to identify low-complexity regions and global repeats respectively, by identifying and scanning for repetitive regions within the genome sequence.

Using the Ensembl annotations we detected repeat elements accounting for from 0.66% to 2.15% of the total intronic length. A much larger proportion of intronic sequence was characterised as repetitive using WindowMasker (Table 3.3) with *D. rerio* introns containing

46.86% repeat sequences. This result for *D. rerio* agrees with the values obtained by Sela *et al* (2010). WindowMasker doesn't annotate the repeats however, thus one can't determine the class of repetitive elements they belong to.

The increased percentage of repeat elements within the *D. rerio* intron sequences accounts for some of the difference in its frequency distribution (Figure 3.1b). It is possible that the additional proportion of this sequence was formerly repetitive, and has since decayed beyond our ability to recognize it as such. Since repetitive elements are likely to be the origins of the majority of all non-coding DNA (Lander *et al*., 2001; Smit, 1999), we propose that the *Danio* lineage experienced an early burst of repeat element expansion that has been decaying for many millions of years. Figures 3.2a and 3.2b show the frequency distribution of repeat elements within the major class of introns (500 bp to 2,000 bp), which includes the region comprising the second intron size peak in *D. rerio* (Figure 3.1a). If there had been a recent expansion of particular repeat elements Figure 3.2a would be expected to show peaks in the frequency of specific size classes. Contrary to this, our analysis reveals a gradual decline in the repeat element size frequency distribution, indicating no recent large-scale repeat expansions. Figures 3.2a and 3.2b also show that the frequencies of both individual and cumulative repeat element sizes are greater in *D. rerio* within the size range expected to contribute to the second zebrafish peak in Figure 3.1a. We consider it likely therefore that repeat elements have contributed importantly to the second *D. rerio* intron size peak, but that this striking repeat expansion was an ancient rather than recent genomic change.

The differences in the distributions may also represent a continuum that with increased sampling within the teleostei infraclass, particularly of those species intermediate to those presented here, would fill the gap. *O. latipes* exhibits a very subtle difference in its intron size class distribution and in being more closely related to *D. rerio* than any of the other fish, adds some weight to this argument.

Large Introns

Large introns can present several problems for organisms, including the expense of transcription and the difficulty of splicing large introns (Shepard *et al*., 2009). The 1,228 large introns in *D. rerio* consist of 107,485,505 nucleotides, which is 17.3% of all *D. rerio* intronic nucleotides and 7.6% of the entire genome sequence. Such large introns may be very costly with regard to both the time and energy required for synthesis (Wagner, 2005). Intronic nucleotides are removed from the mRNA before its export from the nucleus and the synthesis and subsequent degradation of introns has a cost approximately proportional to the length of those introns multiplied by the frequency of transcription. Large introns account for 15.8% of the transcribed section of the genome in *D. rerio*, and while we do not know the transcription rate of the large intron containing genes, they have the potential to account for a significant metabolic cost to the cell.

In addition to metabolic costs, splicing large introns may also introduce conformational problems. A key step of intron splicing is the formation of the loop-like "lariat" structure as the recently cleaved 5' end of the intron is attached to the branch point sequence close to the 3' intron junction. Since a 100 Kb intron may extend out over 30 microns, its size may become a problem for the ~5 micron cell (Shepard *et al.*, 2009). It has been proposed that especially large introns require different splicing mechanisms than standard introns, and that these recursively splice the intron at a series of internal "ratcheting points" rather than in one piece (Shepard *et al.*, 2009; Hatton *et al.*, 1998; Burnette *et al.*, 2005). It is as yet unclear to what extent this large intron ratcheting also occurs in fish.

Wagner (2005) discusses the cost of gene duplication in yeast in terms of extra energy expenditure from increased nucleotides transcribed and finds a significant cost to duplication in terms of extra transcription. We can therefore infer that there must also be a significant cost to large introns. It is possible that these large introns are recent recipients of extensive repetitive sequence expansions and selection has not had time to favour their reduction in size. Our analyses support this, revealing that greater than 70.61% of all large *D. rerio* intron sequence is repeat DNA, also reducing the number of introns greater than 50,000 bp to 426. It is possible that these remaining 426 introns also contain a portion of decayed repeats that cannot be recognised using the novel identification algorithms.

Small introns as a proxy for annotation quality

The minimum intron size reported in a previous Ensembl release (version 59) of *D. rerio* was zero nucleotides, with a further 882 introns less than 5 bp. The existence of 0 bp introns is a result of the way the Ensembl API identifies introns based on the exon coordinates. Given that intron splicing requires a *minimum* of 5 nucleotides (GU-AG plus an A for the branch point) these introns cannot be real and/or functional. In practice, both for steric requirements of intron bending during splicing, and due to the need for other signal sequences, minimum intron sizes are likely to be larger (Schwartz *et al.*, 2008). Certainly in yeast (*Saccharomyces cerevisiae*), there is a conserved 8 bp branch site that is typically 18 bp to 40 bp upstream of the 3’ splice site (Zhuang *et al.*, 1989). This implied 30 bp minimum size in yeast may well be different to vertebrates where branch site sequences are not conserved, but given that the branch point must still be displaced from the intron boundaries and a 3’ polypyrimidine tract interacting with the U2 snRNP auxiliary factor of the spliceosome is common (Zhuang *et al.*, 1989; Adams *et al.*, 1996) typical introns will be considerably larger. For all these reasons we do not consider introns of 1 to 5 nucleotides to be biologically realistic. In *D. rerio* the smallest intron for either U2 or U12 is 11 bp, whereas the “other” splice site category has 412 introns smaller than this. We suggest that since these introns have non-standard splice signals and a different size range to standard introns they should be treated with caution until they are experimentally validated. Although we included all introns annotated by Ensembl in our analyses, “small introns” comprise less than 0.19% of all introns and do not influence our conclusions.

D. rerio is widely considered to be a reasonably high-quality genome annotation, though it undoubtedly contains intron annotation errors, as indeed will all genomes. We note that the extreme intron size outliers in the *D. rerio* genome have changed considerably with releases 59 to 61 of Ensembl. Not only have the two zero-size introns been removed but also an ~2 Mb intron that was previously the largest. It is likely that automated intron annotation errors can particularly skew the extremes of the intron size distribution since these have relatively few members. As an example of an additional source of error in the annotation of genomic introns we can envisage that if a gene was annotated by comparison to cDNA from a paralog containing a small coding indel, or to a transcript that had spliced out a small exon, the extra sequence present in the genomic copy would likely be identified as intronic. Since these coding regions must necessarily be a multiple of 3 bp they will lead to a 3 bp size periodicity of any coding region mis-annotated as intronic and we would expect introns present in the CDS but not 5’-UTR, or 3’-UTR to show such a periodicity. Figure 3.3 shows exactly this 3 bp pattern of periodicity for small introns between ~11 bp and 60 bp. This pattern was present in CDS introns but could not be detected in 5’-UTR or

3’-UTR introns. This indicates that CDS introns smaller than ~60 bp have a significant quantity of mis-annotated coding region.

U2 and U12 introns

Given the difficulties of studying the interaction of the spliceosomes with identified introns, we have based our determination of U2 and U12 introns on the splicing signals they contain. Although this may contain errors since the U12 spliceosomes can interact with U2 type splicing signals (Lin *et al*., 2010) this is not the normal situation, and our error is likely to be very small. The frequencies of intron type are shown in Table 3.7 and reveal that, as expected, the vast majority of introns are of the U2 type. For all fish except *D. rerio* there are 13.9 to 17.4% of introns that we classify as “other”, since they do not possess the classical splicing signals encountered with either U2 or U12 type introns. *D. rerio,* the highest quality genome, has considerably fewer of these “other” introns (6.4%) and a similarly higher percentage of the major U2 type introns, suggesting that the “other” category is dominated by poorly annotated regions.

Conclusions

Understanding the diversity of genome variation using comparative genomics requires a bioinformatics approach that can be tailored and modified by the end user. We have developed a comparative genomics pipeline based on the well-tested and open-source code of the Perl Ensembl Core Software Libraries and BioPerl APIs (Stabenau *et al*., 2004; Stajich *et al*., 2002). Our analysis of the five currently available fish genomes indicates that although the intron content of these genomes is very similar in many respects, different genomic processes appear to be shaping the genomic intron content. The five fish differ not only in scale (number and total amount of intronic sequence) but also the frequency distribution of different intron size classes. The zebrafish *Danio rerio* in particular does not have monotonically decreasing intron frequency with size from an ~80 bp mode, as the other fish appear to have, but rather has a second peak of introns in the 500 bp to 2,000 bp range. Repetitive DNA including transposable elements, satellites sequences and simple repeats are known to be largely responsible for the differences in genome size between species that do not vary in ploidy (Neafsey and Palumbi, 2003; Boulesteix *et al*., 2006; Hawkins *et al*., 2006; Bosco *et al*., 2007) and it is likely therefore that much non-coding DNA will have this origin, even if it has accumulated so many mutations that its previous repetitive nature can no longer be recognised. Our diverse approaches to characterising repetitive elements in *D. rerio* introns revealed that ~47% of intronic sequence could be identified as repetitive. Repeating our analyses only with non-repetitive intron sequences still revealed a unique size distribution for *D. rerio* introns, indicating that this has not been caused by a recent expansion of repetitive sequences, as these would have been readily recognisable as repetitive. Instead we suggest that a more ancient expansion of repeats has created this intronic pattern and little signal of their repetitive origins still remains. As *D. rerio* is the outgroup in this case, a broader sampling of teleost genome sequences in a robust phylogenetic design, with species ancestral to *D. rerio* and intermediate between *D. rerio* and the other fish, would help to locate such an event and better clarify the origins of intron expansion across these lineages.

[This page is left intentionally blank]

CHAPTER FOUR: COMPARATIVE BIOINFORMATICS ANALYSES OF GENE FAMILY SIZE AND FUNCTION IN PRIMATES

Introduction

The impact of duplication on genome structure and content

Duplications contribute a great deal towards the variety in both size and structure of genomes within and between species (Lynch and Conery, 2000; Lynch and Conery, 2003a). There have been many studies that focus on the role of selection in producing differences in sequence composition and phenotype (Sella *et al.*, 2009; Oleksyk *et al.*, 2009; Schluter *et al.*, 2010; Simonson *et al.*, 2010; Bigham *et al.*, 2010; Arnold *et al.*, 2012; Barsh and Andersson, 2013), however duplication is thought to contribute far more to changes in the structure and function of genomic features over evolutionary time scales (Ohno, 1970; Zhang, 2003). Past research suggests that there is much more complexity in terms of variation between genomes due to duplications than due to sequence substitutions (Demuth *et al.*, 2006; Zhang *et al.*, 2009; Sudmant *et al.*, 2013). The difference between humans and chimps alone is approximately 1.5% sequence divergence between orthologous nucleotides and yet at least 6% of genes have species specific in-paralogs (Demuth *et al.*, 2006). There is a fair amount of research examining single nucleotide polymorphism (SNP) data, primarily in model species, in contrast with insertions or deletions. High R_u values - the ratio of unpaired nucleotides attributable to insertions or deletions (indels) to those attributable to substitutions - are seen in thale cress (*Arabidopsis thaliana)*, purple sea urchins (*Strongylocentrotus purpuratus)*, common fruit flies (*Drosophila melanogaster)*, and likely most other species, including bacteria (Britten *et al.*, 2003; Iafrate *et al.*, 2004; Sebat *et al.*, 2004; Freeman *et al.*, 2007). Research examining copy number variation (CNV), however, has focused primarily on humans (Perry *et al.*, 2008), with limited data available for non-human species (Perry *et al.*, 2006; Dopman and Hartl, 2007; Egan *et al.*, 2007; Graubert *et al.*, 2007; Guryev *et al.*, 2008; Lee *et al.*, 2008), meaning that within-species CNVs and between-species copy number differences (CNDs) haven’t yet been fully examined (Locke *et al.*, 2003; Fortna *et al.*, 2004; Demuth *et al.*, 2006; Goidts *et al.*, 2006; Wilson *et al.*, 2006; Dumas *et al.*, 2007).

In addition to genome structure, genome sizes are also greatly influenced by duplications. The amount of repetitive DNA in the genomes of organisms, whose characteristics promote the proliferation and retention of duplications, can be more than 50% of the total amount of DNA. The amount of repetitive DNA in humans for example has recently been estimated at 2,234 Mb, accounting for 78.1% of the total genome size (de Koning *et al.*, 2011). This is greater than the generally accepted value of ~45-50% of the total DNA content, however (Lander *et al.*, 2001). Likewise, 441 Mb in the cotton *Gossypium raimondii*, accounts for 57% of total genome size (Wang *et al.*, 2012). In particular, these repetitive elements also tend to be located within non-coding DNA (Moss *et al.*, 2011), which points to neutral forces of evolution playing a role in their propagation, particularly genetic drift.

Duplication as a population genetic and life history process

Gene duplications are very frequent in all types of genome (Ophir and Graur, 1997; Petrov, 2002; Witherspoon and Robertson, 2003; Zhang and Gerstein, 2003; Johnson, 2004; Blumenstiel *et al.*, 2012), though some more so than others. In plants for example, the number of duplications, especially whole-genome duplications are far greater than in other species (Wendel, 2000; The Arabidopsis Genome Initiative, 2000; Cui *et al.*, 2006; Paterson *et al.*, 2010; Proost *et al.*, 2011). As with most molecular phenomena, their survival are subject to population genetic forces, and the ability of duplications to survive within a lineage are therefore directly impacted by effective population size (N_e) and species life-history characteristics (e.g. generation time, or variation in number of offspring) (Lynch, 2007; Charlesworth, 2009; Brougham, 2011). This is likely the reason why the observation of duplications occurs more often in species with longer generation times and smaller effective population sizes. In unicellular eukaryotes for example, where N_e can vary between 10^7 and 10^8 in ideal conditions (Lynch and Conery, 2003b; Shiu *et al.*, 2005), duplications are constantly fighting against an ever changing genomic landscape. In these circumstances, the power of drift is reduced, and selection will be more effective, resulting in a reduction in the number of duplications being fixed (Lynch and Conery, 2003a). In those species at the opposite extreme, however, duplications are likely to persist far longer within any individual, and because they face less competition are more likely to drift to fixation within a population.

The differences in gain and loss of duplications between taxa at the extremes of N_e can be seen when we compare gene birth rates (*B)* that are relatively stable across extremely divergent species groups, with death rates (*D*) that are quite variable. *D* is much lower, and thus the birth/death ratio (*B/D*) is much higher in plants, such as *Arabidopsis thaliana* where *B* = 0.0032, *D* = 0.033 and *B/D* = 0.0970; in contrast to unicellular species such as *Saccharomyces cerevisiae* where these values are *B* = 0.0025, *D* = 0.324 and *B/D* = 0.0077 (Lynch and Conery, 2003a).

The functional impact of gene duplication

In addition to the quantitative influence that duplications exert on the genome, they also have a qualitative impact. Duplications are very important for a large number of functional reasons and they provide great power in understanding many facets of biology (Korbel *et al.*, 2008; Stapley *et al.*, 2010). They are important in understanding health and disease in humans (Conrad and Antonarakis, 2007) and other species. The evolution of pathogenicity in the Chytrid fungus *Batrachochytrium dendrobatidis* for example, which is linked with a worldwide decline in amphibian populations due to chytridiomycosis can be directly linked to chromosomal copy number variation (Ruiz and Rueda-Almonacid, 2008; Fisher *et al.*, 2009; Langhammer *et al.*, 2013; Farrer *et al.*, 2013; Olson *et al.*, 2013). Differences in the expression and epigenetic landscape of duplicate genes is thought to be the reason behind why ants and other social insects exhibit such a wide variety of morphologically different caste phenotypes from the same genotype (Gadagkar, 1997; Weinstock *et al.*, 2006; Bonasio *et al.*, 2010; Wurm *et al.*, 2010; Schwander *et al.*, 2010; Gadau *et al.*, 2012; Libbrecht *et al.*, 2013; Lattorff and Moritz, 2013). Cichlid fish also exhibit extensive phenotypic variation, albeit from different genotypes, yet gene duplications are postulated to have played an important role in cichlid adaptive radiations (Spady *et al.*, 2006; Watanabe *et al.*, 2007; Seehausen *et al.*, 2008; Hofmann and Carleton, 2009; Fan *et al.*, 2011; Weadick and Chang, 2012).

Adaptation to different environmental conditions are often due to neofunctionalization or subfunctionalization of duplicate genes. Genes that become duplicated, either as a result of speciation (orthologs) or duplication (paralogs), can become fixed into families, with potentially related functions. As the duplicate copies are often under less selective pressure they are able to accumulate mutations without impacting the fitness of the host, and as a result can diverge to take on new functionality, or allow existing functionality to be shared across the duplicates with subtle differences in s. Indeed, we see examples where cold tolerant grasses have emerged due to the evolution of cold stress associated gene families through such processes (Sandve *et al.*, 2008), along with similar adaptations to the cold in Antarctic notothenioid fish (Chen *et al.*, 2008). Many species of plants and fish have evolved tolerance to high levels of salinity (Wu *et al.*, 2012; Chao *et al.*, 2013; Jiang *et al.*, 2013; Norman, 2013), and plants have proved resilient to long periods of drought (Fischer *et al.*, 2011; Sheik *et al.*, 2011) and different soil types (Turner *et al.*, 2010), as well as various other extremes of environment (Oh *et al.*, 2012) due to sub- or neo-functionalization of gene copies. There are also examples from modern human evolutionary history, with an increase in amylase copy number being seen, which is postulated to be due to increased starch in our diets following the advent of agriculture (Pronk *et al.*, 1982; Perry *et al.*, 2007). The two tissue specific groups of amylase (salivary - AMYIA, AMYIB, and AMYIC; and pancreatic - AMY2A, and AMY2B) can be traced back to a single copy during primate evolution (Samuelson *et al.*, 1990). In addition adaptations surrounding domestication of plants and animals thought to be due to duplications are numerous (Liu *et al.*, 2010; Swanson-Wagner *et al.*, 2010; Campos *et al.*, 2011; Nicholas *et al.*, 2011; Sakudoh *et al.*, 2011; Paudel *et al.*, 2013; Olsen and Wendel, 2013).

Computational complexities relating to gene duplications

From a computational perspective; gene duplications, as with repetitive elements, make it extremely difficult for computational biologists to construct an accurate representation of the structure of the genome (Martin, 1999; Bao and Eddy, 2002; Bansal and Eulenstein, 2008; Wehe *et al.*, 2010). This isn't in the least because of CNV both within and between species, which can lead to bias in genome assembly and annotation due to the choice of individual (Levy *et al.*, 2007; Wang *et al.*, 2008; The 1000 Genomes Project Consortium, 2011), but because duplications can be so similar in sequence composition, especially if they have only recently diverged that telling them apart in the first place becomes a real challenge (Bailey *et al.*, 2002; Jiang *et al.*, 2008; Han *et al.*, 2009; Ricker *et al.*, 2012). This similarity may mean that duplications are either collapsed into a single loci by bioinformatics software, thus resulting in an under-representation of gene family size, or in contrast, aren't collapsed sufficiently, resulting in an over-representation. If the quality of the underlying sequencing is poor, resulting in incorrectly called bases, then this error may also result in over-representation due to an artifactual divergence between gene copies. Variable single nucleotide polymorphisms (SNPs) within species, whether an artefact of sequencing, or due to real variability between populations of cells confounds this and requires an appropriate level of sequence coverage to reach a reliable consensus. Projects that sequence single individuals, or individuals that may not accurately reflect the wild-type genotype, such as lab or zoo kept animals (Lindblad-Toh *et al.*, 2005; Liti *et al.*, 2008; Boyko, 2011; Husby *et al.*, 2011; Chen *et al.*, 2013; Alföldi and Lindblad-Toh, 2013), may well benefit from population level re-sequencing to account for variation across the genome in terms of both substitutions and indels, however the quality and coverage of such sequencing needs to be sufficient not to introduce additional copy number variants in error.

A genome and its associated metadata (annotations) are a best guess prediction that can in many parts be confirmed by experimental evidence, but often still subject to inference errors. Areas of genomes are often highlighted as problematic when, as with computer source-code version control, "bugs" are raised due to additional research highlighting inaccuracies in the consensus sequence. Assemblies are therefore updated over time and different genome assemblies released that are more likely to represent the true genome structure. Projects like Ensembl (Hubbard *et al.*, 2002; Flicek *et al.*, 2012) take these updated assemblies and aggregate other "fixes", often from external data sources, to incorporate them into their new gene build and annotation pipelines. A new version of Ensembl, with updates for a number of species, is released approximately every 3 months relegating preceding releases to their Archives website (see http://www.ensembl.org/info/website/archives/index.html) and thus improving the accuracy of the theoretical genome structures over time.

Ensembl uses comparative genomics (Chinwalla *et al.*, 2002; Clamp *et al.*, 2003; International Chicken Genome Sequencing Consortium, 2004; Vilella *et al.*, 2009) to improve the accuracy of genome assemblies and annotations by providing a phylogenetic context to underpin their construction. Increased sampling across the phylogenetic tree aids in filling in gaps in the genome sequence between many taxa where the divergence is too great to provide an effective sequence alignment. Indeed many lower coverage sequencing projects have been undertaking to provide intermediate scaffolds between more divergent species to assist in their assembly, annotation and comparison. Comparisons can also be made between species to highlight areas of the genome that tend not to vary or remain static across vastly divergent phylogenetic distances. Identifying portions of the genomes that vary very little between species provides some manner of quality control. There are tools available that attempt to determine the quality and coverage of genome assemblies in this manner, but the number of variables to consider are vast, and these checks are often limited to identified coding genes and other functional areas of the genome, which constitute much less than 10% of the total genome content (Mikkelsen *et al.*, 2005; Green *et al.*, 2006; Noonan *et al.*, 2006; Sea Urchin Genome Sequencing Consortium *et al.*, 2006; Taft *et al.*, 2007; Velasco *et al.*, 2007; Jaillon *et al.*, 2007; Gibbs *et al.*, 2007; Merchant *et al.*, 2007; Organ *et al.*, 2007; Dieterich *et al.*, 2008; Thomas, 2008; Schmutz *et al.*, 2010).

Without a template to compare to, the accuracy of the genome is largely unknown. New technologies promise to provide single molecule sequencing essentially reading the linear primary sequence of strands of DNA a base at a time (Clarke *et al.*, 2009; Eid *et al.*, 2009; Pushkarev *et al.*, 2009; Steinbock *et al.*, 2012; Mason and Elemento, 2012). This may allow us more certainty in determining the true primary structure of the genome, however error rates are still high (Wang *et al.*, 2012; Simpson, 2013), though at least better estimated (Carneiro *et al.*, 2012; Roberts *et al.*, 2013; Powers *et al.*, 2013). There are a number of programs that can be used to assemble the raw sequence reads into the contigs and ultimately scaffolds that represent the linear sequence of a genome (Warren *et al.*, 2006; Zerbino and Birney, 2008; Simpson *et al.*, 2009; Zerbino *et al.*, 2009; Boisvert *et al.*, 2010; Li *et al.*, 2010; Boisvert *et al.*, 2012; Leo *et al.*, 2012), though varying the parameters used when executing these programs can have vastly different outcomes on the inferred assembly (Earl *et al.*, 2011; Zhang *et al.*, 2011; Bradnam *et al.*, 2013). There are attempts at implementing software that utilises probability theory to build models of the genome assemblies, estimating log-likelihood values in order to find the assembly that maximises the likelihood given the raw sequence reads (Medvedev and Brudno, 2009; Clark *et al.*, 2013; Hunt *et al.*, 2013; Ghodsi *et al.*, 2013), however this is still an immature field of research and the algorithmic considerations are so complex that they are unlikely to ever be solved in polynomial time (Cook, 1971; Levin, 1973; Jones and Pevzner, 2004). These tools can be useful in comparing the assemblies from varying the parameters of the same genome assembly program, or from assemblies created using different programs, in order to highlight the most optimal, though this is extremely time consuming and computationally expensive, requiring specialist, high-specification hardware to run.

As with genome assembly; genome annotation is also an extremely complex computational task (Schadt *et al.*, 2010; Fernald *et al.*, 2011; Mak, 2011; Yandell and Ence, 2012; Ward *et al.*, 2013; Yip, 2013) that is fraught with error (Brenner, 1999; Devos and Valencia, 2001). The volumes of data required to be aggregated for the purposes of providing a relatively accurate picture of the components of the genome are simply not feasible to undertake manually in a tractable time frame (Searle *et al.*, 2004; Loveland, 2005; Ashurst *et al.*, 2005; Wilming *et al.*, 2007; Amid *et al.*, 2009). The infrastructure required to undertake automated genome annotation is vast (Birney *et al.*, 2004; Cuff *et al.*, 2004; Curwen *et al.*, 2004; Potter *et al.*, 2004), and as with automated genome assembly is subject to error, though steps are taken to minimise this by incorporating various sources of evidence (Curwen *et al.*, 2004) along with manual annotation and quality control (Wilming *et al.*, 2007). Any errors in assembly and annotation are likely to impact on downstream analyses, such as the correct identification of gene family members (Ames *et al.*, 2012; Han *et al.*, 2013).

Computational tools for identifying gene family size change

The tools developed to identify change in gene family size are limited (De Bie *et al.*, 2006; Ames *et al.*, 2012; Liu *et al.*, 2011; Librado *et al.*, 2011). The most robust and widely used of these programs is CAFE, which has been applied extensively to determine expansions and contractions or gene families across many taxa (Hahn *et al.*, 2007; Sharpton *et al.*, 2009; Nygaard *et al.*, 2011; Shapiro *et al.*, 2013; Vogel and Moran, 2013; Wu *et al.*, 2013), however there has been some question over its accuracy in light of changing parameters and model assumptions (Ames *et al.*, 2012; Han *et al.*, 2013). In addition to making numerous assumptions at an algorithmic level (such as inferring a constant birth/death rate across the phylogenetic tree), these programs also make the assumption that the annotated data they are provided with are correct. Any inaccuracies resulting from the assembly or annotation will impact on the downstream results, as previously discussed, meaning that the output from these programs needs to be assessed in context.

Previous work has highlighted that gene family sizes approximate a power law distribution (Huynen and van Nimwegen, 1998; Koonin *et al.*, 2002; Karev *et al.*, 2003; Barabási and Oltvai, 2004), where one quantity varies as a power of another. Explicit probabilistic graph models of Birth-Death in gene families have been described as being able to accurately highlight significant changes in gene family size (Hahn *et al.*, 2005; De Bie *et al.*, 2006), though variations on these models that introduce parameters such as "innovation" are perhaps more robust as they don't assume balanced birth and death rates remaining asymptotically closer to the estimated distribution curve (Karev *et al.*, 2003; Karev *et al.*, 2004; Karev *et al.*, 2005; Novozhilov *et al.*, 2006).

By necessity, the determination of change in gene family size requires incorporating the phylogenetic relatedness of the species, so that gene family sizes at internal nodes can be reconstructed and significance values estimated based on the likelihood that change at a particular node varies from what is expected. CAFE takes a maximum likelihood approach, though other programs provide weighted parsimony (Ames *et al.*, 2012) and Bayesian (Liu *et al.*, 2011) implementations that perform at least equally well. It is clear however that the different approaches to gene family size estimation, along with variation in parameters and *a priori* assumptions made by the programs can produce different results (Ames *et al.*, 2012; Liu *et al.*, 2011; Vieira and Rozas, 2011; Han *et al.*, 2013). Undertaking comparisons between the different approaches and scrutinising them within their relevant biological context is paramount if we are to reach robust and reproducible conclusions.

Functional annotation of the genome sequence

Just as a collection of raw sequence reads is of no real use to biologists wanting to test specific hypotheses, an unannotated genome assembly is likewise. The steps towards annotating a genome are numerous (Karolchik *et al.*, 2003; Curwen *et al.*, 2004; Yandell and Ence, 2012) and subject to different agendas, of which assigning the function of their parts is only one. Initial stages of genome annotation focus on performing *de novo* gene prediction; highlighting open-reading frames, transcription start sites, masking repetitive elements and generally constructing a basic structure of the genome on which to build further evidence (Potter *et al.*, 2004). Ensembl, for example, calls this stage the “raw compute”, which is followed by the gene build proper and protein annotation stages (Curwen *et al.*, 2004). Different genome annotation pipelines follow similar steps (Kent *et al.*, 2002; Karolchik *et al.*, 2003) with the latter stages bringing in evidence from known protein, cDNA and EST sequences (Curwen *et al.*, 2004; Potter *et al.*, 2004). By referencing this experimental evidence, or performing *de novo* homology searches it is possible to assign function to the annotated components of the genome automatically (Kent *et al.*, 2002; Diehn *et al.*, 2003; Conesa *et al.*, 2005; Quevillon *et al.*, 2005; Hinrichs *et al.*, 2006; Flicek *et al.*, 2007; Mulder *et al.*, 2008; Schmid and Blaxter, 2008; Falda *et al.*, 2012). What constitutes function however, has been subject to much debate (Bemstein *et al.*, 2012; Eddy, 2012; Doolittle, *et al.*, 2013; Eddy, 2013; Graur *et al.*, 2013; Niu and Jiang, 2013; Hurst, 2013). In the context of this chapter, function refers to genes that are transcribed and translated, and that have the Ensembl `bio_type` classification `protein_coding`.

The Ensembl pipeline annotates function by reference to external databases such as InterPro (Apweiler *et al.*, 2001; Hubbard *et al.*, 2002; Hunter *et al.*, 2011) and provides a dedicated functional genomics database and API to its users as of release 47 (Flicek *et al.*, 2007). Different methods of assigning function can result in conflicting classifications however (Rison *et al.*, 2000). As with many other areas of bioinformatics (Stevens *et al.*, 2000; Stein, 2002; Bodenreider and Stevens, 2006; Smith *et al.*, 2007; Antezana *et al.*, 2009), a standard for the classification of gene and gene products across species and databases was developed known as the Gene Ontology (GO) (Ashburner *et al.*, 2000; Blake and Harris, 2002; Ashburner *et al.*, 2005; Plessis *et al.*, 2011). The development of this standardised approach has largely improved matters, such as conflicting naming conventions, though it still requires improvement and conformity from researchers (Rhee *et al.*, 2008; Tirmizi *et al.*, 2011).

It is interesting, in philosophical terms, to know the function of the components of the genome. The real power of functional annotation comes into its own however, when used in the context of *in silico* hypothesis testing. By utilising functional annotations such as GO terms in an analyses to determine significantly expanded or contracted gene families, it is possible to make inferences on the evolution of those gene families from the perspective of adaptation (Demuth *et al.*, 2006; Hahn *et al.*, 2007a; Hahn *et al.*, 2007b; Clark *et al.*, 2007; Sharpton *et al.*, 2009). Previous studies have discussed a bias towards functions of a reproductive, immunological and developmental nature (Hahn *et al.*, 2005; Demuth *et al.*, 2006; Hahn *et al.*, 2007a; Dumas *et al.*, 2007; Demuth and Hahn, 2009), and this makes sense from a logical standpoint. Genes of this nature will intrinsically be subject to strong positive or purifying selection due to environmental pressures, increasing the likelihood of the expansion or contraction of their constituent families within any given population. These analyses can be circular however, and it is wise to interpret them in context. The birth and death of gene families is disjointed from natural selection. The ability of duplicate genes to arise and be maintained within a population is subject to a number of variables, with a clear distinction between the processes responsible for their establishment and subsequent modification by mutation and natural selection (see Chapter One; Lynch and Conery, 2000; Lynch and Conery, 2003). Making inferences on the cause of an expansion or contraction *ex post facto* must be subject to scrutiny with a number of sources of evidence to support the inferred conclusions.

Chapter goals

This chapter will attempt to assess the power of current bioinformatics approaches in accurately reflecting the duplication landscape within genomes as part of comprehensive comparative analyses. Analyses to identify duplications presents a substantial challenge to bioinformaticians, even within such a relatively closely related group of species as the primates. Difficulties relating to the underlying assemblies and annotations will be discussed, in addition to the complexities involved with the different methodological approaches to gene family size inference. A significant expansion of gene families within the branch of the phylogenetic tree leading to modern humans is identified and within this context is proposed as likely to be an artefact of the data, rather than a real biological change. In addition functional genomic analyses of significantly expanded and contracted duplications will be discussed with a view to highlighting the role of the member genes, and to understand why they have been maintained in a population. An assumption is made that most genes with an increased rate of duplication will be involved in reproductive, immunological or development roles and the data confirms this. These analyses also confirm findings from previous studies, but reach largely different conclusions to others. This highlights the need for detailed scrutiny of any results obtained, as choice of method can lead to widely different conclusions.

Materials and Methods

Species used in this study

Primates

The primates used in this study were the common marmoset *Callithrix jacchus*, the western lowland gorilla *Gorilla gorilla gorilla*, the human *Homo sapiens*, the rhesus macaque *Macaca mulatta*, the grey mouse lemur *Microcebus murinus*, the northern white-cheeked gibbon *Nomascus leucogenys*, the northern greater galago *Otolemur garnettii*, the common chimpanzee *Pan troglodytes*, the Sumatran orang-utan *Pongo pygmaeus abelii*, and the Philippine tarsier *Tarsius syrichta*. The northern treeshrew *Tupaia belangeri* was used as the outgroup species.

Rodents

The rodents used in this study were the guinea pig *Cavia porcellus*, the Ord's kangaroo rat *Dipodomys ordii*, the thirteen-lined ground squirrel *Ictidomys tridecemlineatus*, the house mouse Mus musculus, and the Norwegian brown rat *Rattus norvegicus*. The European rabbit *Oryctolagus cuniculus* was used as the outgroup species.

Primates gene families analyses

Data retrieval

Novel scripts were developed in the Perl programming language using the GCAT API (see Chapter Two) to retrieve data on primate gene families from the Ensembl Core (Hubbard et al., 2002; Flicek et al., 2011) and Compara (Clamp et al., 2003; Vilella et al., 2009) databases. Two approaches were taken when retrieving the data to ensure that all gene families were retrieved. Data were retrieved from both release 66 and release 67 of Ensembl.

The “gene ids” approach first retrieved all the gene IDs for each species from the Ensembl Core database and then used the pooled gene IDs to get all the associated gene families from the Ensembl Compara database. The “all families” approach retrieved data on all gene families from the Ensembl Compara database and subsequently mined those results for families belonging to the required species. The former method was much quicker, but the latter more robust.

Determining expansions and contractions

Appropriate subroutines were developed to call external programs such as CAFE (De Bie *et al.*, 2006; Hahn *et al.*, 2007), BEGFE (Liu *et al.*, 2011), and DupliPHY (Ames *et al.*, 2012) that are necessary for determining the significantly expanded or contracted gene families within an input dataset. Routines to convert between required input formats and to parse the output from these programs were also developed. Only CAFE (version 2.2) was used for the analyses in this study as this was the most extensively used and robust program for determining change in gene family size at the time, and integrated well with Ensembl data.

CAFE requires a Newick format species tree, and a tabular format file as input. The tabular format input file requires a DESCRIPTION column, an ID column, and columns for each of the species within the associated species tree. The DESCRIPTION column provides a description of the gene family, for example *ALPHA AMYLASE PRECURSOR* and the ID column provides the gene family ID, for example *ENSFM00660001157182*. In this case the DESCRIPTION and ID match the Ensembl gene family description and ID as per the data retrieval step described previously. The additional columns require an integer value corresponding to the number of genes within the gene family for that particular species. CAFE can be run in interactive mode, where the commands and parameters are entered via a command line interface, however it can also be called as a shell using a shell script to pass the input commands and parameters. The latter approach was more amenable to distributed and programmatic analyses using GCAT and so this option was chosen.

A Newick format species tree with branch lengths was used, which was produced by Ensembl as part of their genome annotation pipeline (Potter *et al.*, 2004; Vilella *et al.*, 2009). The R package ape was used to extract the necessary primates tree from within the Ensembl species tree using the *extract.clade()* function. It was necessary to adopt the program r8s (Sanderson, 2003) and the ape package (Paradis *et al.*, 2003) to perform branch rate smoothing to overcome issues with CAFE where the product of lambda (see Chapter One) and the tree depth were greater than 1 ($\lambda * t < 1$) must be true; where t is the time from the tips to the root). The *chronopl()* and *is.ultrametric()* ape functions were used for this purpose.

The birth-death model of CAFE assumes at least one gene in the root of the species tree. Though CAFE allows the use of a *-filter* input flag when running the *load* command, a helper script was implemented in Python (`trim_cafe_families.py`) to prune all families with 0 gene members in the outgroup from the dataset. These scripts are available in the `support_files` directory of the GCAT source code repository.

CAFE allows one to use a fixed lambda across the tree or to estimate lambda for each branch. CAFE was run with both a fixed lambda value for each branch and also allowed to vary with different values for each branch of the tree. In both cases lambda was called with the *-s* input flag, which uses an optimisation algorithm to search for the value(s) of lambda that maximise the log likelihood of the data for all families. In addition CAFE was run with a *-e* flag to estimate the values of lambda for each of the gene families individually. When running CAFE this way it isn't possible to perform ancestral state reconstruction for all families combined and so individual runs of CAFE were performed to identify gene families that were significantly expanded or contracted. Novel Python and Perl GCAT helper scripts (`get_significant.pl`, `get_sig_freqs.pl`, and `merge_sig_files.py`) were developed to integrate all these analyses and allow for downstream processing and visualisation. These scripts are available in the `support_files` directory of the GCAT source code repository.

Data processing and visualisation

Novel R subroutines (see `ppf_freqs.R`, `sig_freqs.R`, and `visualize_goterms.R`) were developed and integrated with the GCAT API using the `Statistics::R` package to automate the visualisation of gene families identified as being significantly expanded or contracted. R packages plyr (Wickham, 2011), reshape2 (Wickham, 2007) and ggplot2 (Wickham, 2009) were used for this purpose.

Rodents gene families analyses

Data were retrieved and analysed for the rodents using the same protocol developed for retrieving the primates gene families. See primates gene families analyses above.

Assembly and annotation information

The Ensembl genome browser (Stalker *et al.*, 2004; Spudich *et al.*, 2007; Spudich and Fernández-Suárez, 2010), relevant NCBI Genome resources and respective sequencing consortium websites (see Appendix 4.1 and 4.2) were used to investigate assembly and annotation information for both the primates and rodents in an attempt to determine the current state of the genome assemblies and to question whether any potential error associated with the assemblies and/or annotations may exist.

Functional classifications of significantly expanded or contracted primates gene families

Data retrieval

Additional GCAT plugin scripts were developed to retrieve annotations from the Ensembl Compara database and the Gene Ontology Database (The Gene Ontology Consortium, 2001) using Perl Ensembl Compara API's Bio::EnsEMBL::Compara::FamilyAdaptor class and the Bio::EnsEMBL::DBSQL::GOTermAdaptor class respectively in order to determine the functional classifications of these data (`get_goterms.pl`).

Analytics and visualisation

GCAT based R scripts (`visualize_goterms.R` and `dumas_compare.R`) were developed to undertake data analytics and visualisation. Visualisation modules utilised the ggplot2 library (Wickham, 2009).

Comparison with previously identified functional classifications

Data were mined using a custom R script (`dumas_compare.R`) to compare and contrast the findings reported here with those of a previous study (Dumas *et al.*, 2007). These scripts are available in the `support_files` directory of the GCAT source code repository.

Supplementary Table S5 from the Dumas study was used to identify gene clusters with copy number changes with lineage specific (LS) changes in humans. A list of gene names corresponding to the hg18 assembly at UCSC (Kuhn *et al.*, 2007) were retrieved. These gene names were matched with their corresponding Ensembl gene IDs by using R (`dumas_compare.R`) to access the EnsMart and BioMart web service APIs (Kasprzyk *et al.*, 2004; Kasprzyk, 2011). This then allowed retrieval of their respective Ensembl gene family IDs. The retrieved gene family IDs were compared with the data from the release 67 raw gene family data and release 67 CAFE fixed lambda data using R (`dumas_compare.R`) to check for any matching hits.

Results

Primates gene families

Data retrieval

The “gene ids” approach identified 203,918 protein coding gene IDs for which it subsequently retrieved gene family data. This resulted in 9,166,908 lines of output, corresponding to annotated gene families. The “all families” approach identified 655,218 gene families, of which 226,280 (204,060 unique) gene IDs were identified as belonging to primates. The differences in these values are due to redundancy in the “all families” dataset. Further processing yielded very similar counts for gene families in the data (see Table 4.1), though the “all families” method was the most robust in retrieving the greatest number of gene families and all further analyses used the “all families” data.

Table 4.1 - Comparison of different data retrieval methods and the absolute numbers of data retrieved for the primates from release 66 of the Ensembl Core and Compara database.

	“gene ids” method	“all families” method
Raw data count	9,166,908	655,218
Raw gene family count	203,918	226,280
Primates gene family count	49,719	49,737
Gene family members	226,254	226,280
Number of genes	204,034	204,060
Number of species	11	11

Following trimming of the data, to meet CAFE’s requirements for having no 0-sized gene families in the out-group, this resulted in 11,307 gene families for further downstream analyses.

The sizes of the gene families vary between species with a minimum size of 0, meaning complete loss in that species, to a maximum size of 265 in *Homo sapiens*. The mean, median and mode sizes are very low pointing to a right skew in the data. Indeed upon calculating the value for skewness, we see a positive skew (see Table 4.2).

Table 4.2 – Primate gene family (GF) sizes in the release 66 "all families" gene family data.

Species	Min GF Size	Max GF Size	Mean GF Size	Median GF Size	Mode GF Size	Skew
Callithrix jacchus	0	219	1.38	1	1	44.84
Gorilla gorilla	0	220	1.30	1	1	53.93
Homo sapiens	0	265	1.37	1	1	56.85
Macaca mulatta	0	256	1.37	1	1	56.85
Microcebus murinus	0	125	1.09	1	1	35.44
Nomascus leucogenys	0	211	1.20	1	1	64.58
Otolemur garnettii	0	120	1.03	1	1	34.75
Pan troglodytes	0	235	1.27	1	1	61.90
Pongo abelii	0	183	1.23	1	1	51.09
Tarsius syrichta	0	146	0.91	1	1	53.05
Tupaia belangeri	1	93	1.37	1	1	28.99

By visualising the data we are able to get a graphical representation of the gene family sizes, allowing us to better understand how the data are distributed. We are able to view a frequency distribution of all gene family sizes (see Figure 4.1) and per species gene family sizes (see Figure 4.2).

Determining expansions and contractions

A total of 11,307 gene families were given as input to the CAFE software. CAFE identified 538 gene families as being significantly expanded or contracted. The minimum and maximum gene family sizes weren't changed. The frequency distribution of these data followed a similar pattern (see Figure 4.3).

In order to understand these data further, it was necessary to view them in a phylogenetic context. GCAT was used to plot these data on their associated species tree (see Figure 4.4). A large relative expansion in genes is seen in the branch leading to modern humans.

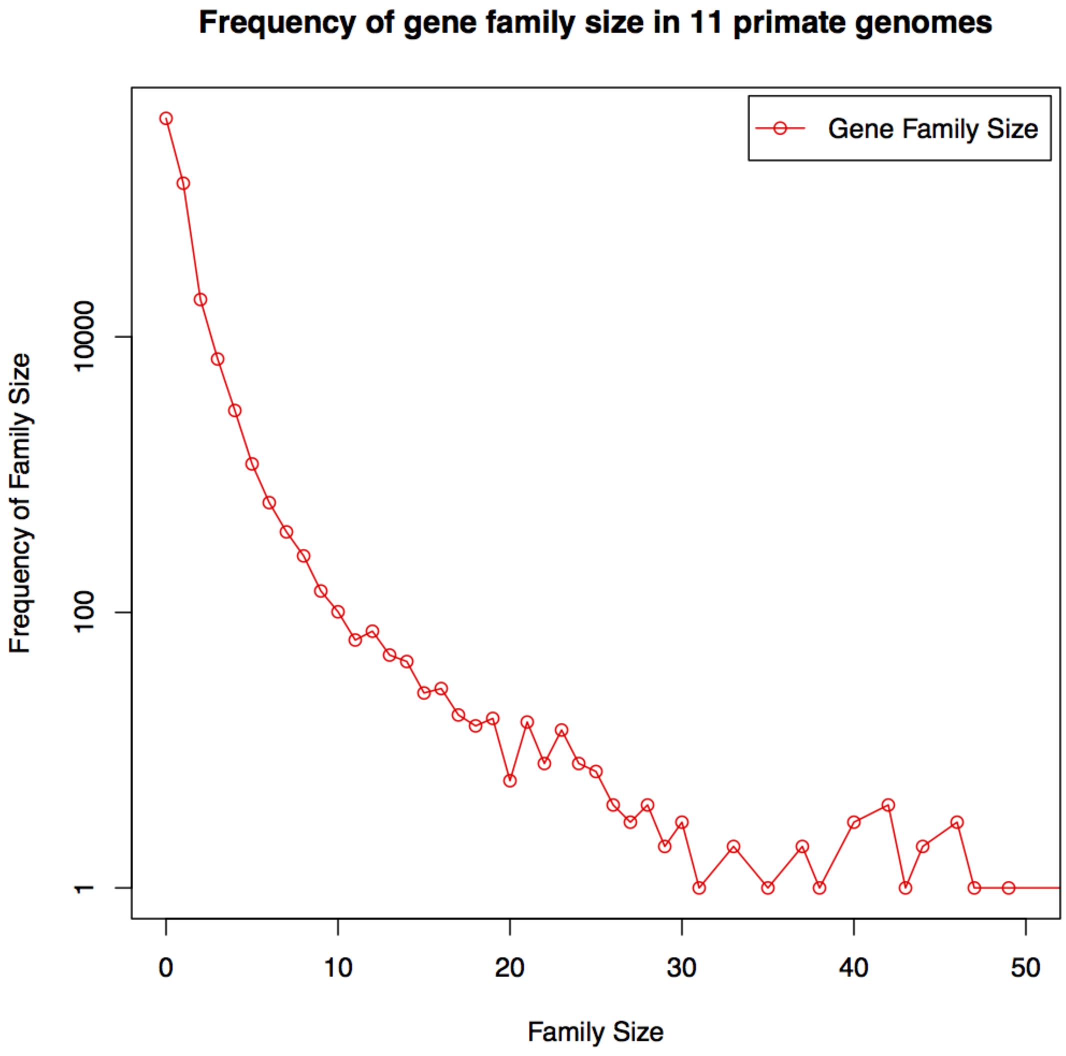


Figure 4.1 - Frequency distribution of pooled gene family sizes for release 66 of the primates gene family data. A cut-off of 50 is used as the maximum on the x-axis as this represents the majority of the data. A 0-size gene family means complete loss in that species.

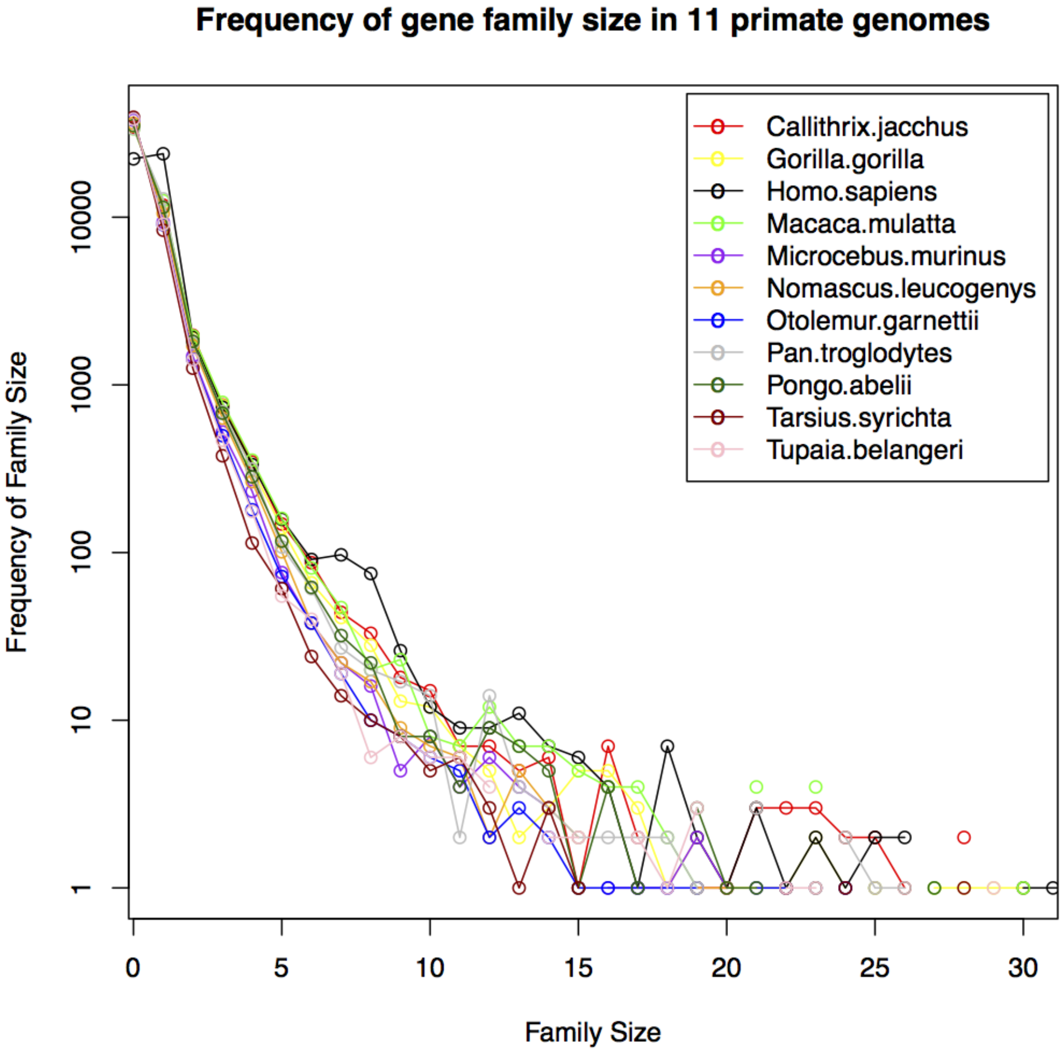


Figure 4.2 - Frequency distribution of gene family sizes in each primate for release 66 of the primates gene family data. A cut-off of 30 used as maximum on the x-axis as this represents the majority of the data. A 0-size gene family means complete loss in that species.

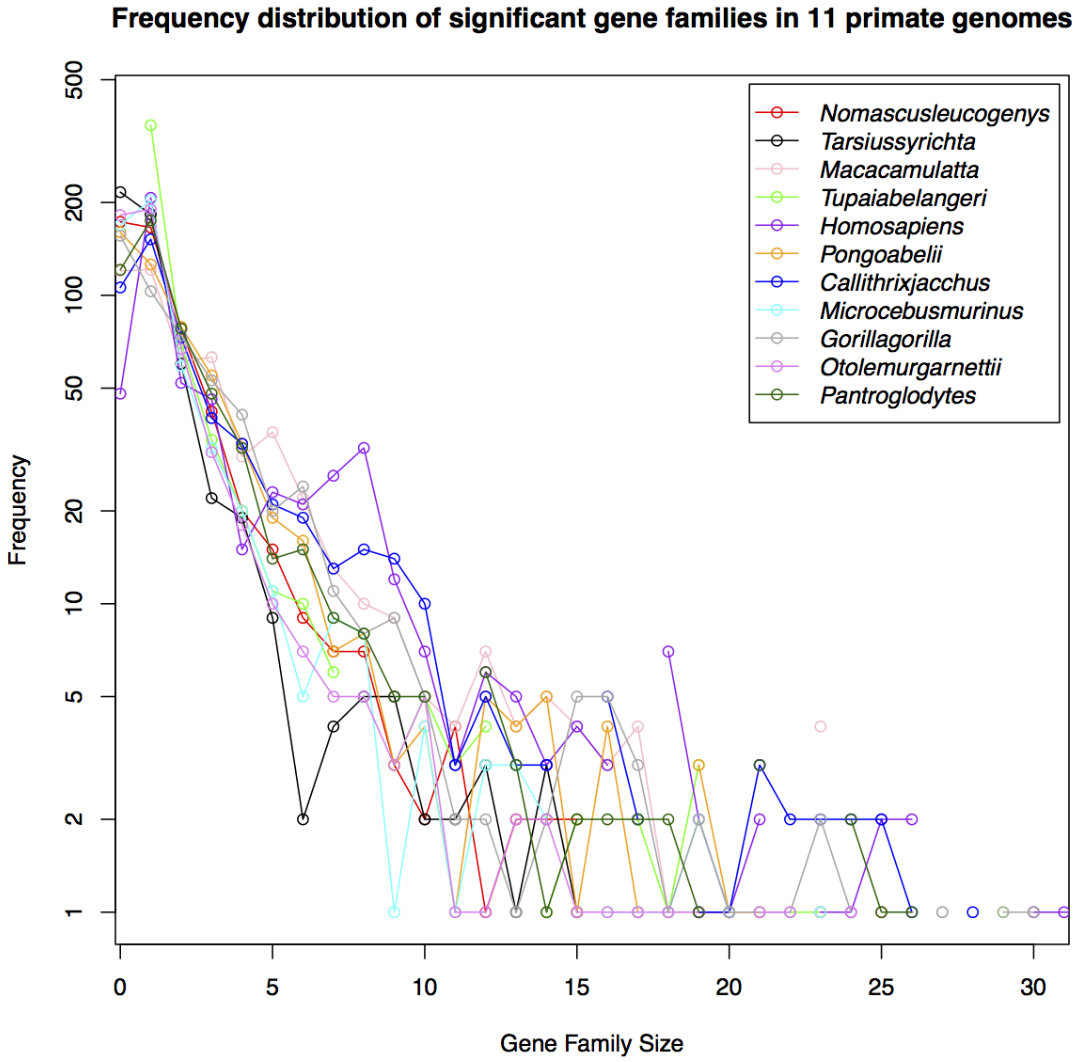


Figure 4.3 - Frequency distribution of significantly expanded or contracted gene family sizes in each primate for release 66 of the primates gene family data. A cut-off of 30 used as maximum on the x-axis as this represents the majority of the data. A 0-size gene family means complete loss in that species.

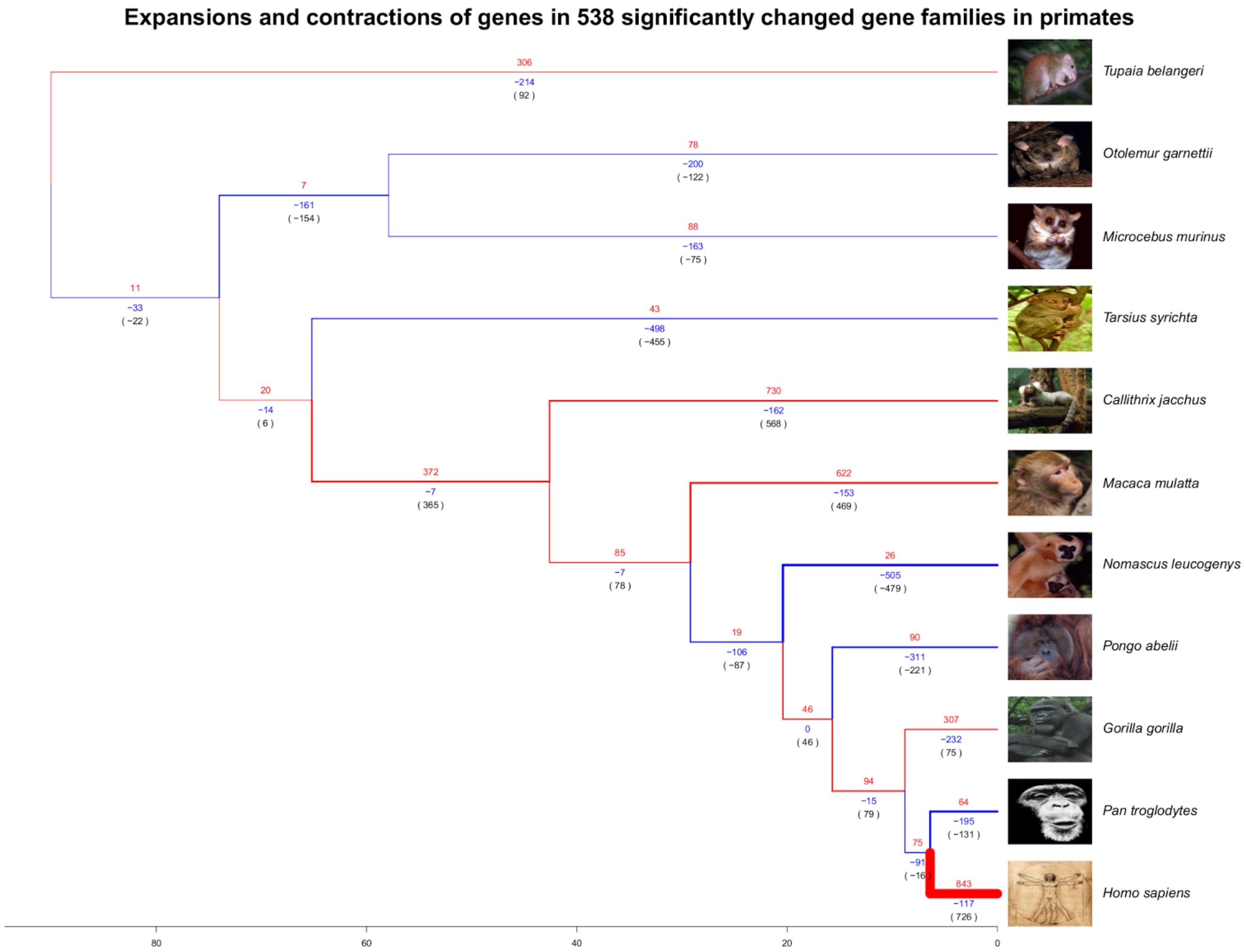


Figure 4.4 - Expansions and contractions of genes along the branches of the primate phylogenetic tree. Blue coloured branches depict overall contraction, while red coloured branches depict overall expansion. Black branches would represent equal or no change. Branch thickness represents the number of gene copy number changes weighted by the time to the ancestral node for each branch as a proportion of the time to the root node.

Rodents’ gene families

Data retrieval

The “all families” approach identified 628,351 gene families, of which 118,328 gene IDs were identified as belonging to primates. Further processing yielded very similar counts for gene families in the data (see Table 4.3), though the “all families” method was the most robust in retrieving the greatest number of gene families and all further analyses used the “all families” data.

Table 4.3 - Absolute numbers of data retrieved for the rodents from release 66 of the Ensembl Core and Compara database using the “all families” method.

	“all families” method
Raw data count	628,351
Raw gene family count	124,999
Rodents gene family count	25,610
Gene family members	124,999
Number of genes	118,328
Number of species	6

Following trimming of the data, to meet CAFE’s requirements for having no 0-sized gene families in the outgroup, this resulted in 11,947 gene families for further downstream analyses.

The sizes of the gene families vary between species with a minimum size of 0, meaning complete loss in that species, to a maximum size of 285 in *Mus musculus*. The mean, median and mode sizes are very low pointing to a right skew in the data. Indeed upon calculating the value for skewness, we see a positive skew (see Table 4.4).

Table 4.4 – Rodent gene family (GF) sizes in the release 66 "all families" gene family data

Species	Min GF Size	Max GF Size	Mean GF Size	Median GF Size	Mode GF Size	Skew
Cavia porcellus	0	66	1.40	1	1	12.86
Dipodomys ordii	0	71	1.16	1	1	14.29
Ictidomys tridecemlineatus	0	77	1.43	1	1	16.24
Mus musculus	0	285	1.57	1	1	38.86
Rattus norvegicus	0	164	1.59	1	1	21.61
Oryctolagus cuniculus	1	92	1.62	1	1	17.13

By visualising the data we are able to get a graphical representation of the gene family sizes, allowing us to better understand how the data are distributed. We are able to view a frequency distribution of all gene family sizes (see Figure 4.5) and per species gene family sizes (see Figure 4.6).

Determining expansions and contractions

A total of 11,947 gene families were given as input to the CAFE software and run using a fixed lambda. CAFE identified 414 gene families as being significantly expanded or contracted. The minimum and maximum gene family sizes weren't changed. The frequency distribution of these data followed a similar pattern (see Figure 4.7).

In order to understand these data further, it was necessary to view them in a phylogenetic context. GCAT was used to plot these data on their associated species tree (see Figure 4.8). A large relative expansion in genes is seen in the branch leading to modern humans.

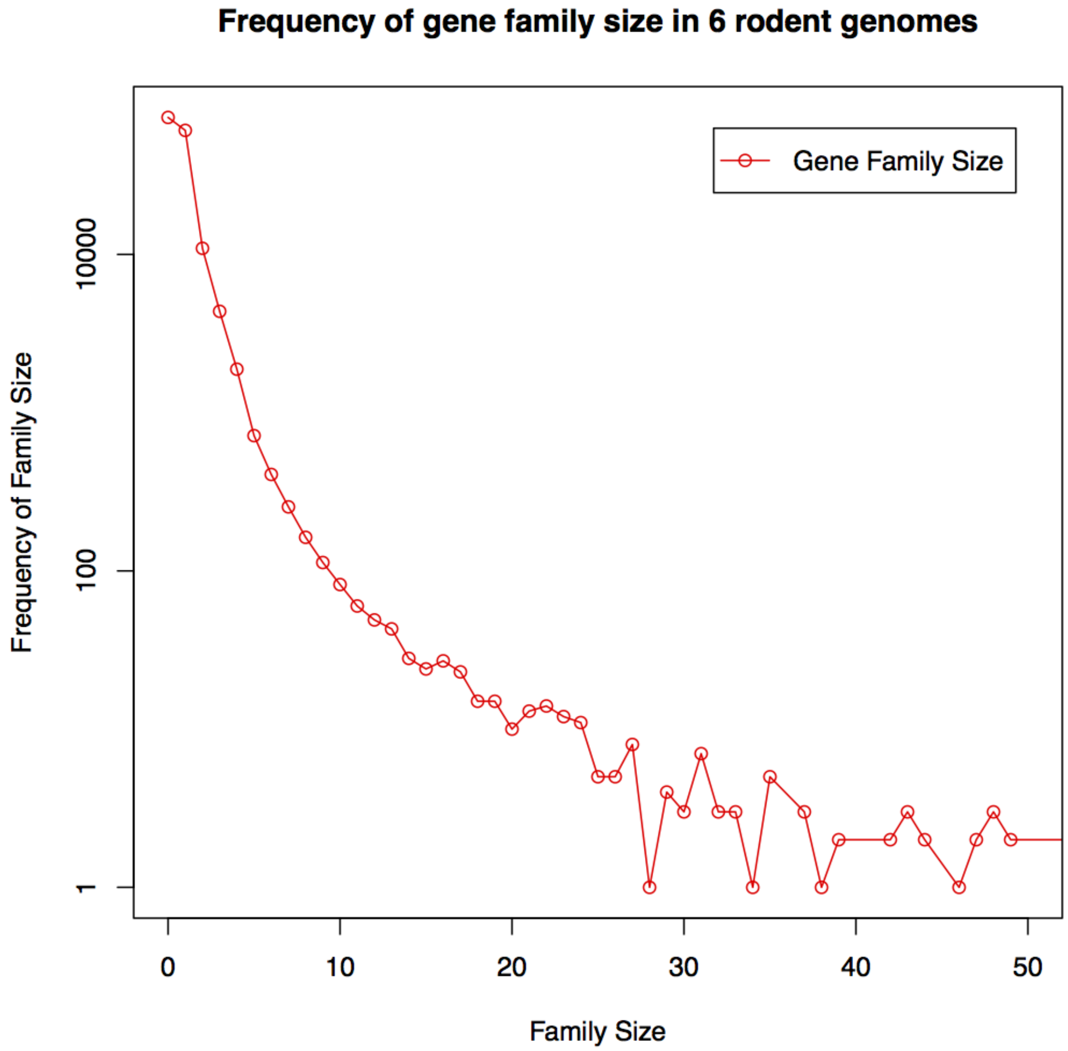


Figure 4.5 - Frequency distribution of pooled gene family sizes for release 66 of the rodents gene family data. A cut-off of 50 is used as maximum on the x-axis as this represents the majority of the data. A 0-size gene family means complete loss in that species.

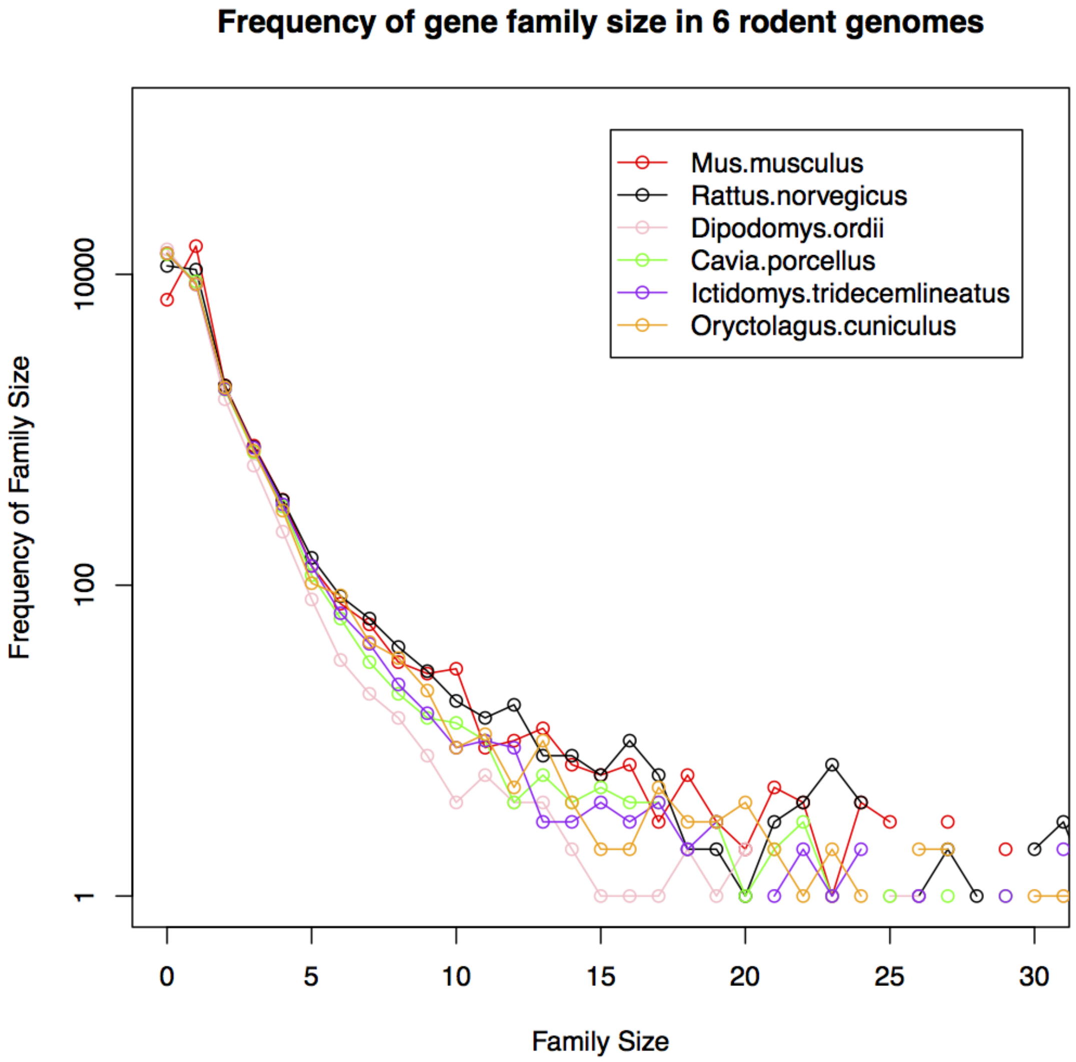


Figure 4.6 - Frequency distribution of gene family sizes in each primate for release 66 of the rodents gene family data. A cut-off of 30 used as maximum on the x-axis as this represents the majority of the data. A 0-size gene family means complete loss in that species.

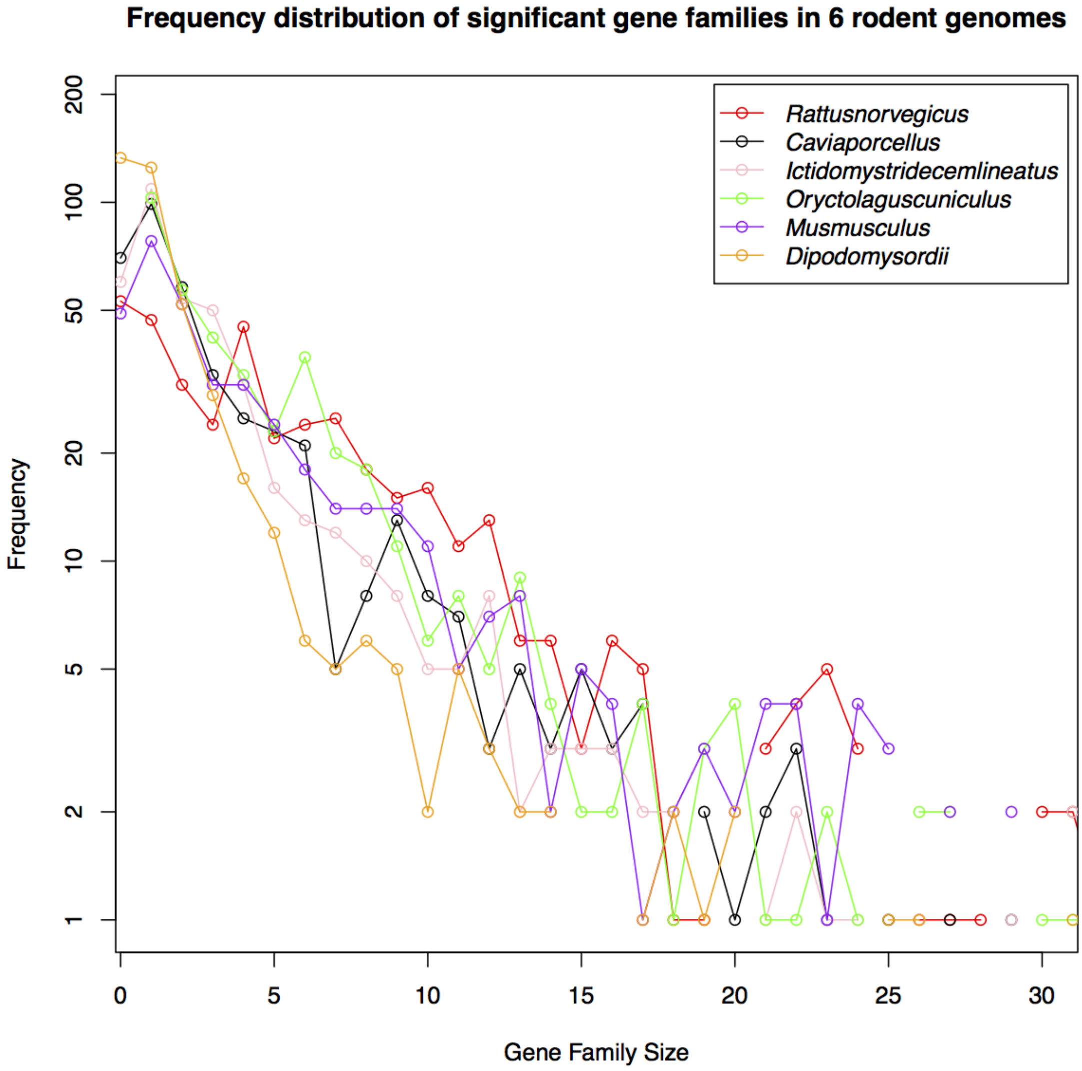


Figure 4.7 - Frequency distribution of significantly expanded or contracted gene family sizes in each rodent for release 66 of the rodents gene family data. A cut-off of 30 used as maximum on the x-axis as this represents the majority of the data. A 0-size gene family means complete loss in that species.

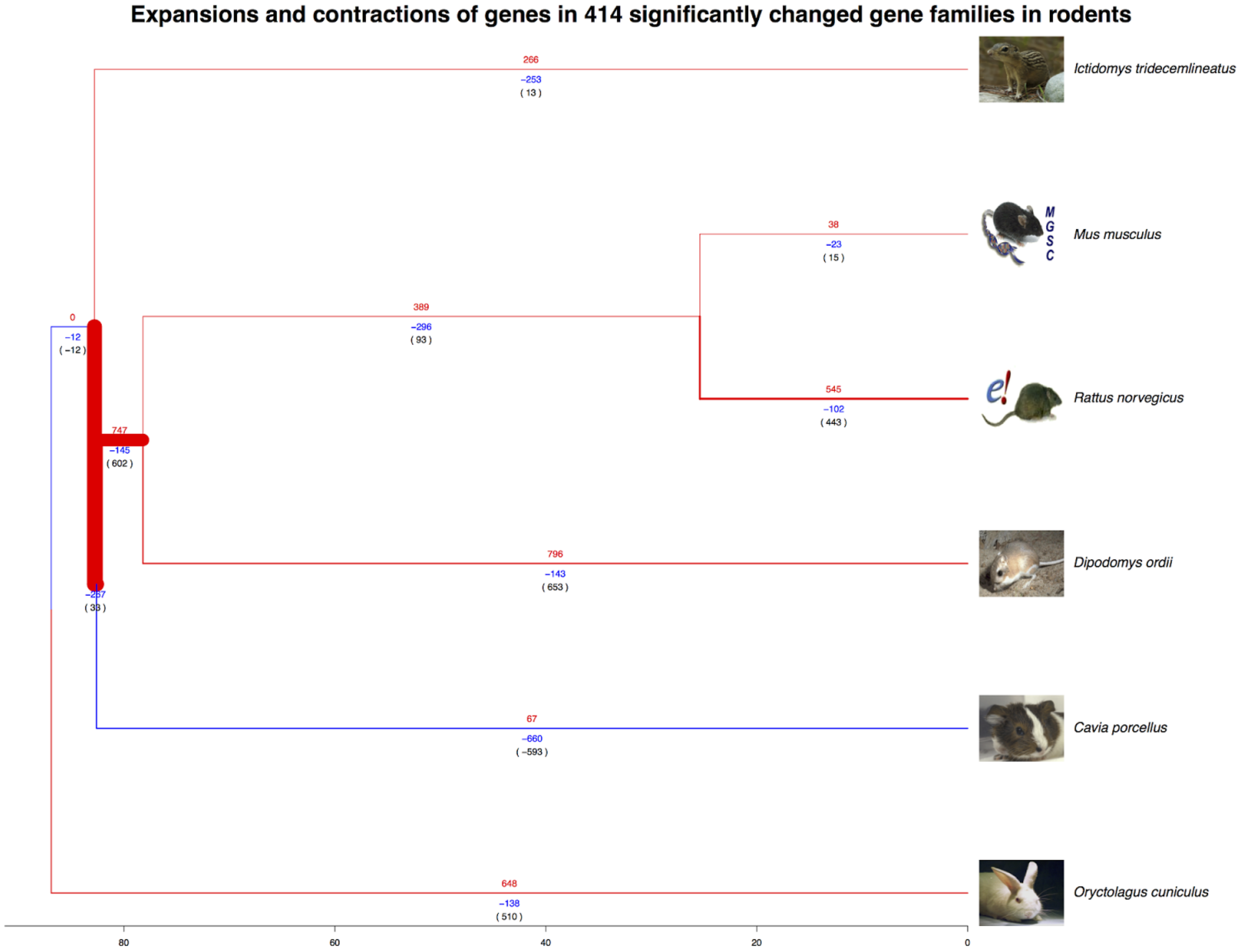


Figure 4.8 - Expansions and contractions of genes along the branches of the rodent phylogenetic tree. Blue coloured branches depict overall contraction, while red coloured branches depict overall expansion. Black branches would represent equal or no change. Branch thickness represents the number of gene copy number changes weighted by the time to the ancestral node for each branch as a proportion of the time to the root node.

Assembly and annotation information

Comparison of coverage and assembly

Manual comparison of sequence coverage and determination of assembly quality based on the chemistry used, original assembly release date, first gene build and date of most recent patch (see Table 4.5 and Table 4.6) were undertaken using resources available from Ensembl, GenBank and the respective genome consortium website (see Appendices 4.1 and 4.2).

Release 67 data

Data retrieval

To further determine the impact of annotation quality on the outcome of CAFE analyses an updated release 67 dataset was used to rerun the analyses. The “all families” method returned a total of 628,351 identified gene families, containing a total of 41,159 primate gene families with 226,693 gene members, corresponding to 208,105 unique gene IDs (see Table 4.7).

The data was trimmed for 0-size gene families in the outgroup, as per previous protocol, resulting in 11,024 gene families for further downstream analyses.

The maximum sizes of the gene families differed in the release 67 data with a new maximum of 271 in *Homo sapiens*. The mean, median and mode sizes remained low although also differed from the release 67 data (see Table 4.8).

The frequency distribution of pooled gene family sizes (see Figure 4.9) and species independent gene family sizes (see Figure 4.10) followed a similar pattern.

Determining expansions and contractions

A total of 11,024 gene families from the release 67 data were given as input to the CAFE software and separate runs were undertaken with a fixed lambda value across the tree, with a variable lambda value across the tree, and with a variable lambda .

CAFE results with a fixed lambda across the tree

CAFE identified 626 gene families as being significantly expanded or contracted. The minimum and maximum gene family sizes weren’t changed. The frequency distribution of these data followed a similar pattern (see Figure 4.11) and a large relative expansion was still observed in the branch leading to modern humans (see Figure 4.12).

Table 4.5 - Genome assembly information for primates species used in this study taken from release 66 of the Ensembl genome browser.

Species	Assembly	Date	Coverage	Genebuild Released	Genebuild Patched
Callithrix jacchus	Callithrix jacchus-3.2.1	January 2010	6X	May 2010	March 2011
Gorilla gorilla	gorGor3.1	December 2009	2.1X + 35X	March 2010	July 2011
Homo sapiens	GRCh37.p6	February 2009	5.11X + 7.5X	April 2011	February 2012
Macaca mulatta	MMUL 1.0	February 2006	5.1X	August 2006	May 2010
Microcebus murinus	micMur1	June 2007	1.93X	March 2008	May 2010
Nomascus leucogenys	Nleu1.0	January 2010	5.6X	April 2011	April 2011
Otolemur garnettii	OtoGar3	March 2011	137X	December 2011	December 2011
Pan troglodytes	CHIMP2.1.4	February 2011	6X	December 2011	December 2011
Pongo abelii	PPYG2	September 2007	6X	March 2008	May 2010
Tarsius syrichta	tarSyr1	July 2008	1.82X	February 2009	May 2010
Tupaia belangeri	tupBel1	June 2006	2X	February 2007	May 2010

Table 4.6 - Genome assembly information for rodents species used in this study taken from release 66 of the Ensembl genome browser.

Species	Assembly	Date	Coverage	Genebuild Released	Genebuild Patched
Cavia porcellus	cavPor3	March 2008	6.79X	September 2008	May 2010
Dipodomys ordii	dipOrd1	July 2008	1.85X	February 2009	May 2010
Mus musculus	NCBIM37	April 2007	7X	January 2011	March 2012
Oryctolagus cuniculus	oryCun2	November 2009	7X	March 2010	December 2011
Rattus norvegicus	RGSC 3.4	December 2004	7X	September 2009	May 2010
Ictidomys tridecemlineatus	speTri1	June 2006	1.90X	April 2007	May 2010

Table 4.7 - Absolute numbers of data retrieved for the primates from release 67 of the Ensembl Core and Compara database using the "all families" method.

	"all families" method
Raw data count	628,351
Raw gene family count	226,693
Primates gene family count	41,159
Gene family members	226,693
Number of genes	208,105
Number of species	11

Table 4.8 – Primate gene family (GF) sizes in the release 67 "all families" gene family data.

Species	Min GF Size	Max GF Size	Mean GF Size	Median GF Size	Mode GF Size	Skew
Callithrix jacchus	0	220	1.48	1	1	44.91
Gorilla gorilla	0	228	1.40	2	1	56.26
Homo sapiens	0	271	1.51	2	1	54.90
Macaca mulatta	0	262	1.48	2	1	57.60
Microcebus murinus	0	125	1.17	1	1	36.27
Nomascus leucogenys	0	214	1.28	2	1	66.04
Otolemur garnettii	0	212	1.43	2	1	42.25
Pan troglodytes	0	227	1.29	1	1	64.70
Pongo abelii	0	188	1.32	2	1	53.32
Tarsius syrichta	0	151	0.98	1	1	53.83
Tupaia belangeri	1	99	1.40	1	1	30.56

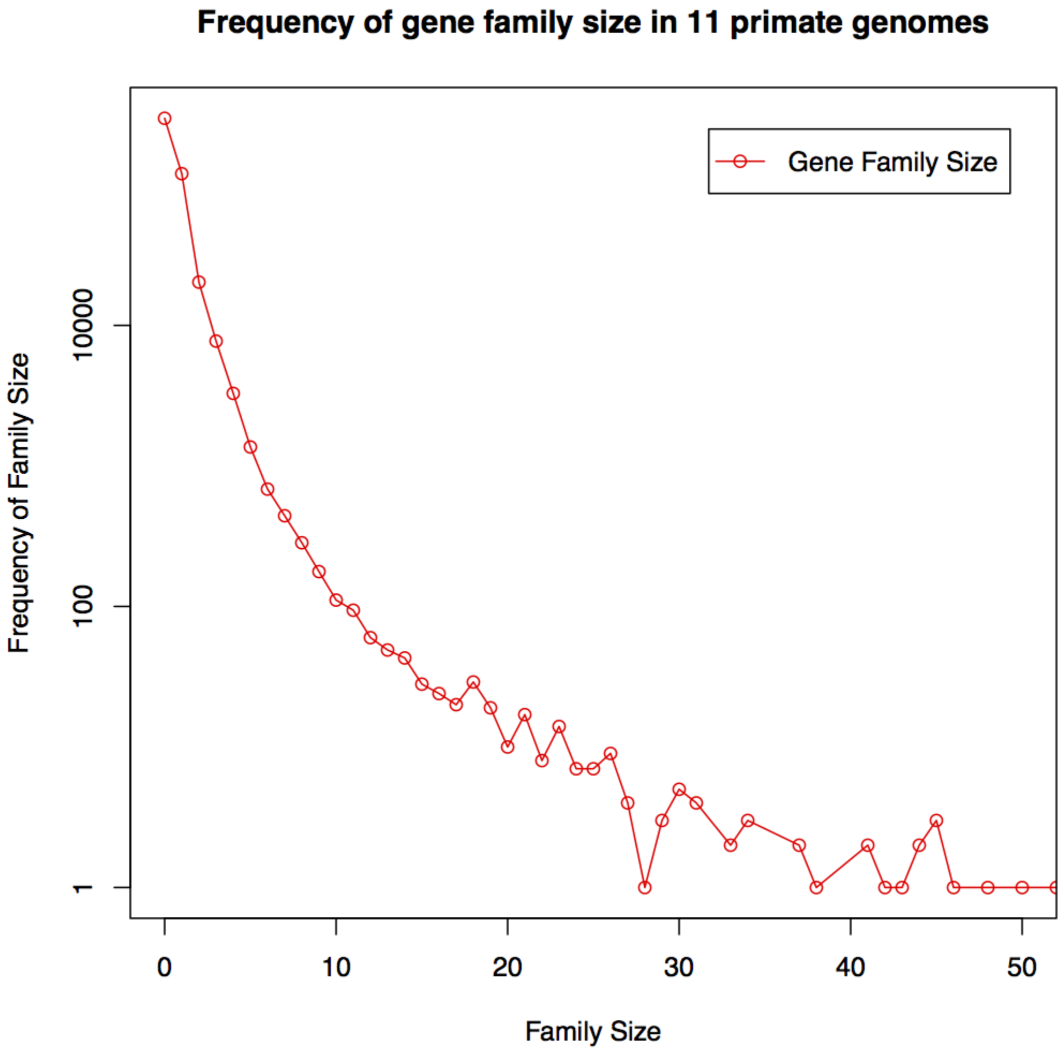


Figure 4.9 - Frequency distribution of pooled gene family sizes for release 67 of the primates gene family data. A cut-off of 50 is used as maximum on the x-axis as this represents the majority of the data. A 0-size gene family means complete loss in that species.

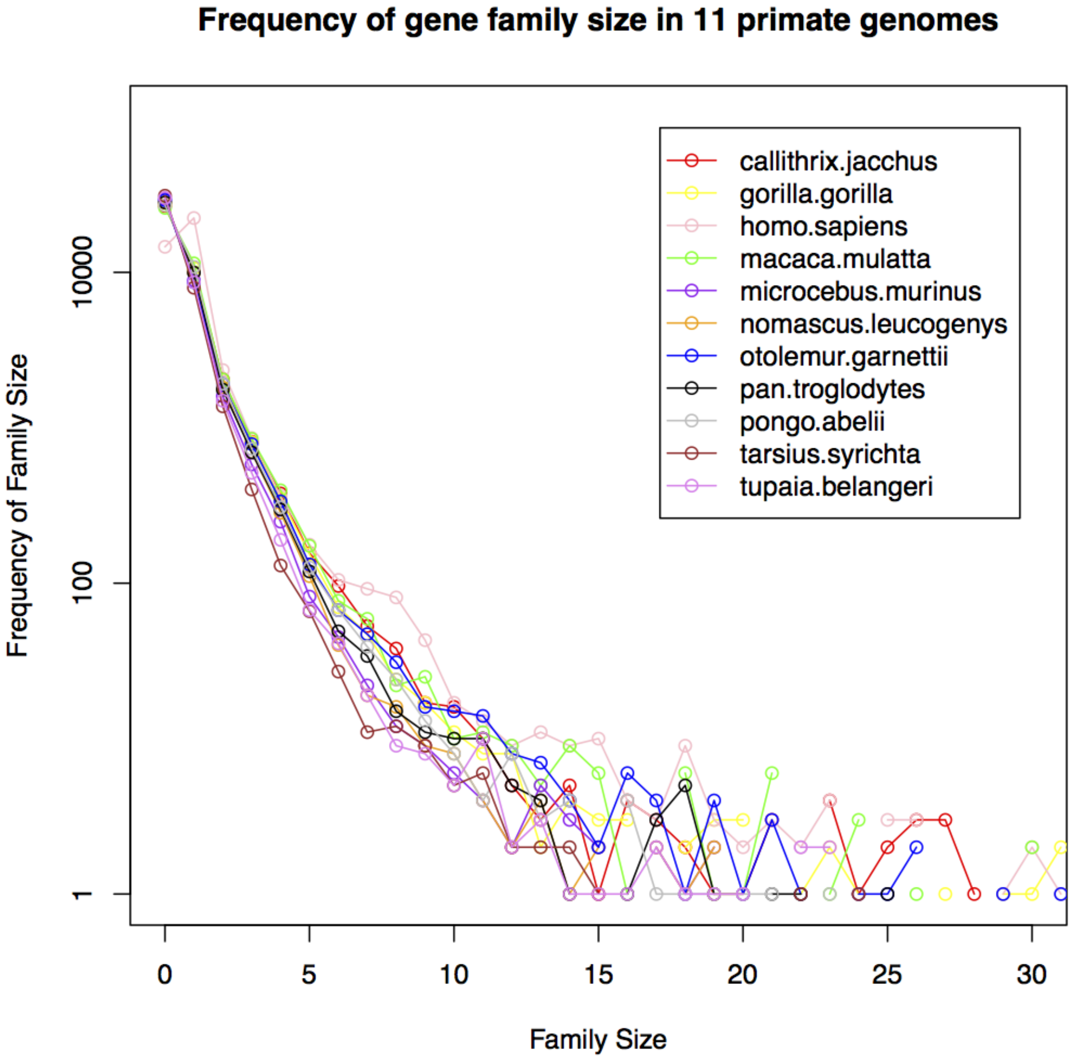


Figure 4.10 - Frequency distribution of gene family sizes in each primate for release 67 of the primates gene family data. A cut-off of 30 is used as maximum on the x-axis as this represents the majority of the data. A 0-size gene family means complete loss in that species.

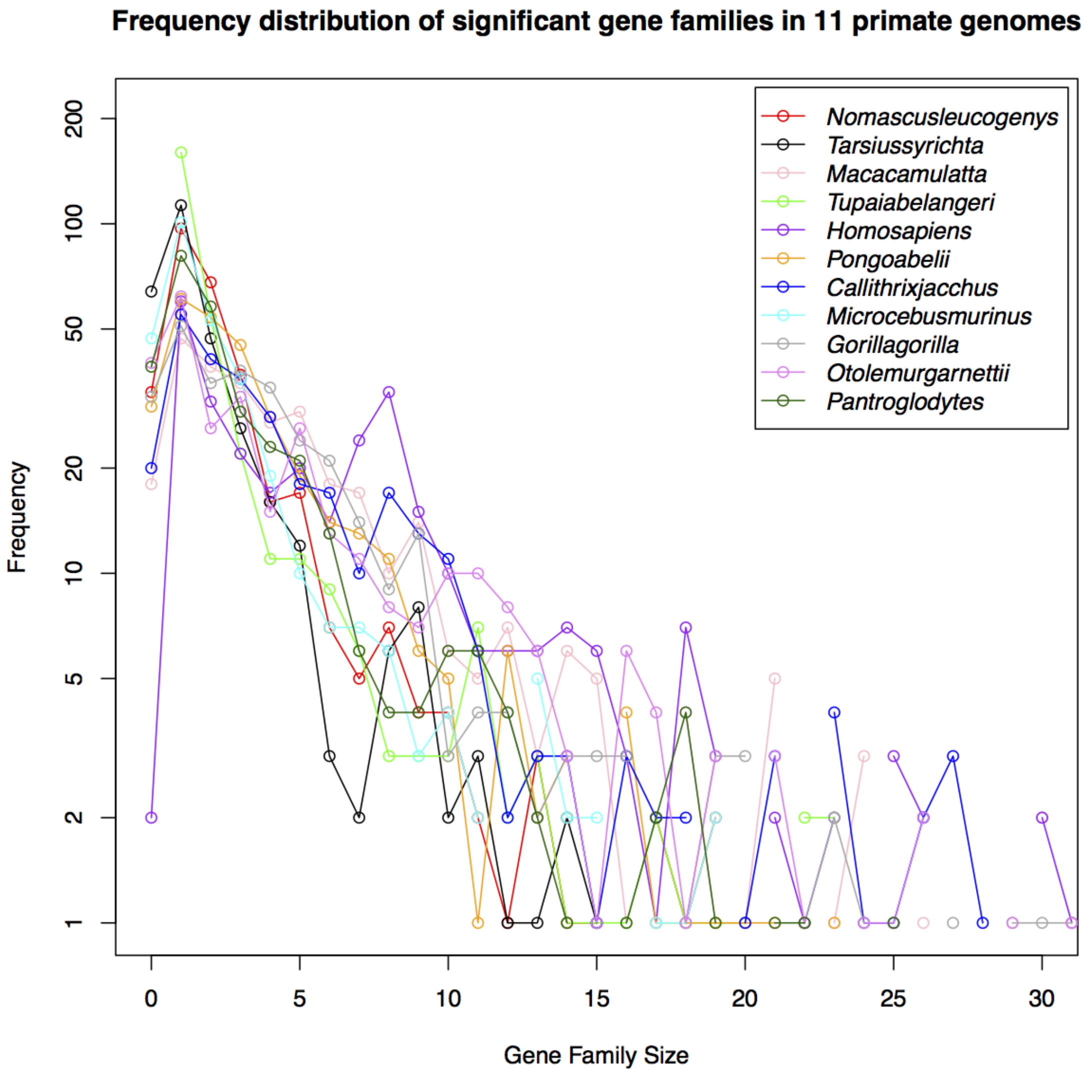


Figure 4.11 - Frequency distribution of significantly expanded or contracted gene family sizes in each primate for release 67 of the primates gene family data using a fixed lambda across the tree. A cut-off of 30 used as maximum on the x-axis as this represents the majority of the data. A 0-size gene family means complete loss in that species.

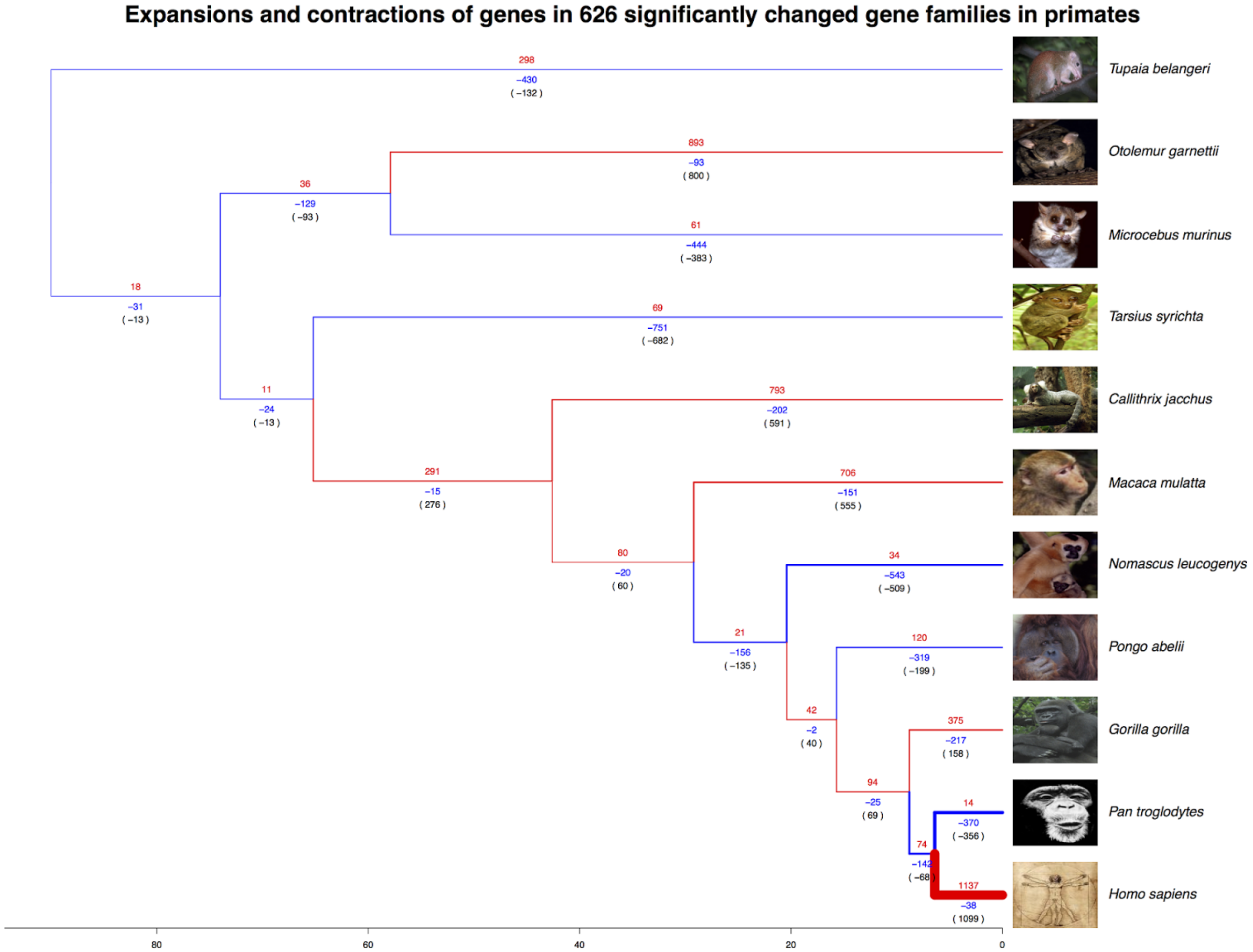


Figure 4.12 - Expansions and contractions of genes along the branches of the primate phylogenetic tree for release 67 of the primates gene family data using a fixed lambda across the tree. Blue coloured branches depict overall contraction, while red coloured branches depict overall expansion. Black branches would represent equal or no change. Branch thickness represents the number of gene copy number changes weighted by the time to the ancestral node for each branch as a proportion of the time to the root node.

CAFE results with a varied lambda across the tree

CAFE identified 309 gene families as being significantly expanded or contracted. The minimum and maximum gene family sizes weren't changed. The frequency distribution of these data followed a similar pattern (see Figure 4.13) and a large relative expansion was still observed in the branch leading to modern humans (see Figure 4.14).

CAFE results with a varied lambda for each gene family and across the tree

CAFE was run to estimate values for each family and with a variable lambda across the tree. When using CAFE to estimate individual lambdas for each gene family, it isn't possible to undertake ancestral state reconstruction or perform Monte Carlo simulations for all families combined. 1,000 iterations were performed per family to estimate the lambda that maximised the log likelihood value and a consensus of the lambda values were output for each family (see Appendix 4.3). It is then possible to analyse each family independently, requiring CAFE to be executed 11,024 times.

The 11,024 individual CAFE runs identified 163 gene families as being significantly expanded or contracted. The minimum and maximum gene family sizes weren't changed. The frequency distribution of these data followed a similar pattern (see Figure 4.15) and a large relative expansion was still observed in the branch leading to modern humans (see Figure 4.16).

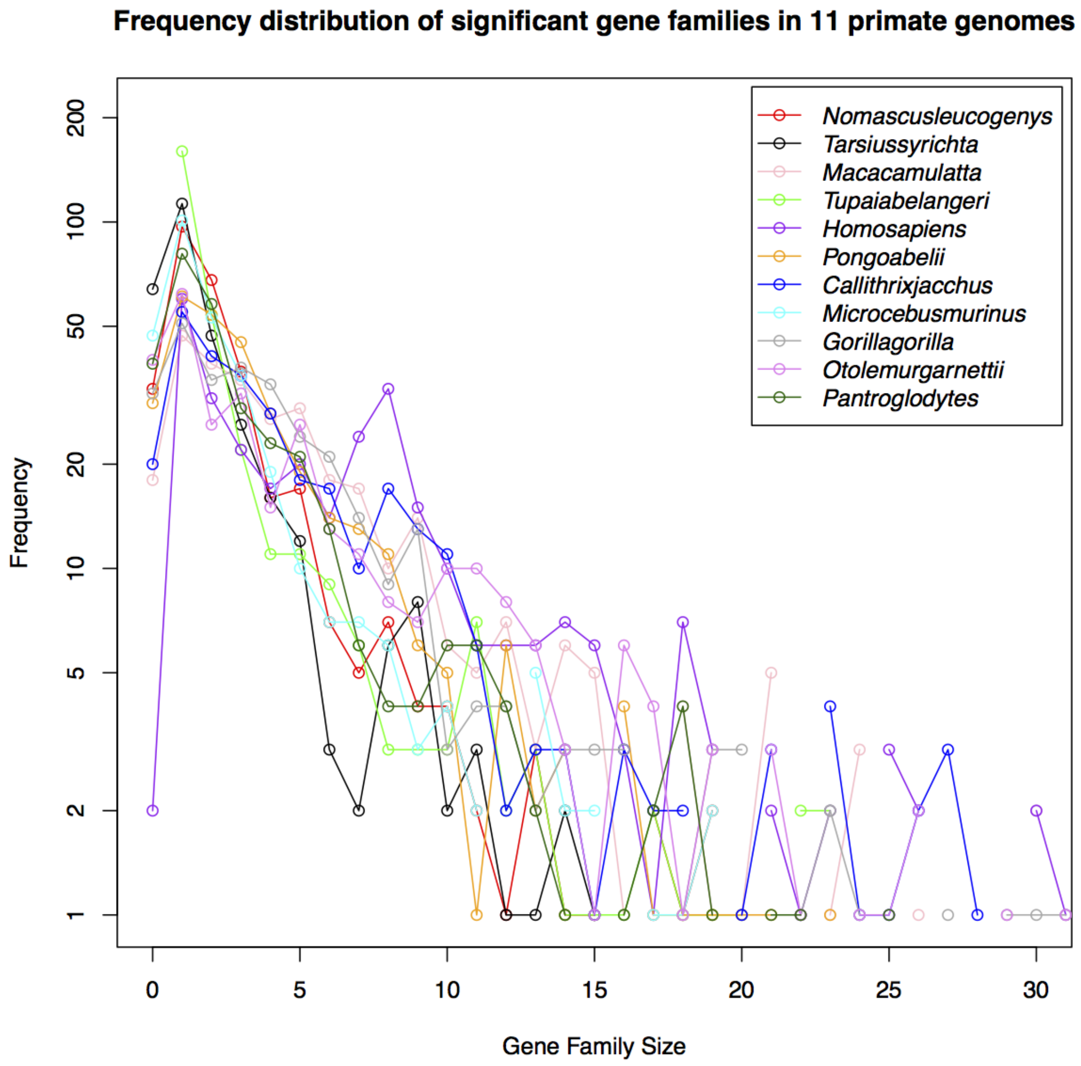


Figure 4.13 - Frequency distribution of significantly expanded or contracted gene family sizes in each primate for release 67 of the primates gene family data using a variable lambda across the tree. A cut-off of 30 used as maximum on the x-axis as this represents the majority of the data. A 0-size gene family means complete loss in that species.

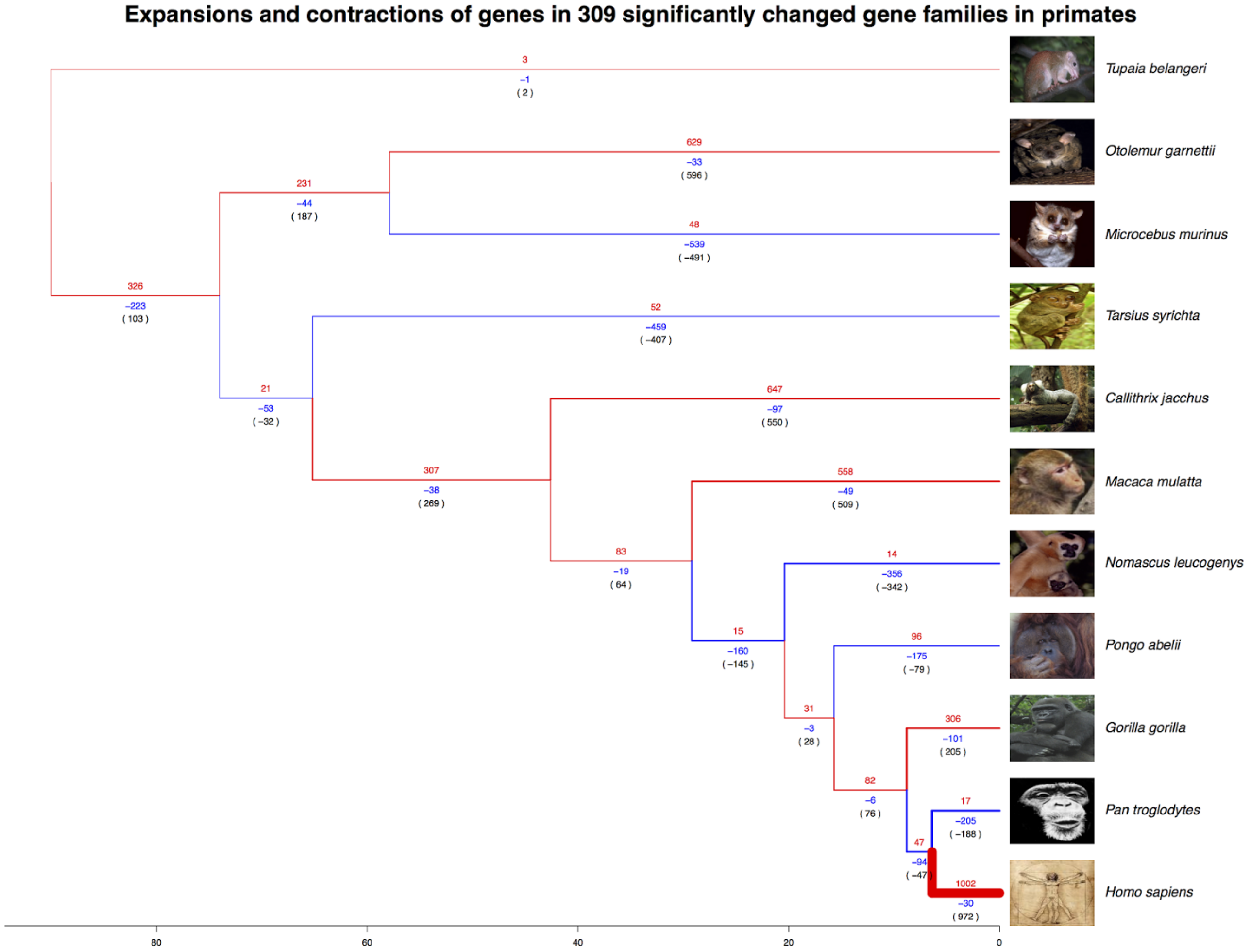


Figure 4.14 - Expansions and contractions of genes along the branches of the primate phylogenetic tree for release 67 of the primates gene family data using a variable lambda across the tree. Blue coloured branches depict overall contraction, while red coloured branches depict overall expansion. Black branches would represent equal or no change. Branch thickness represents the number of gene copy number changes weighted by the time to the ancestral node for each branch as a proportion of the time to the root node.

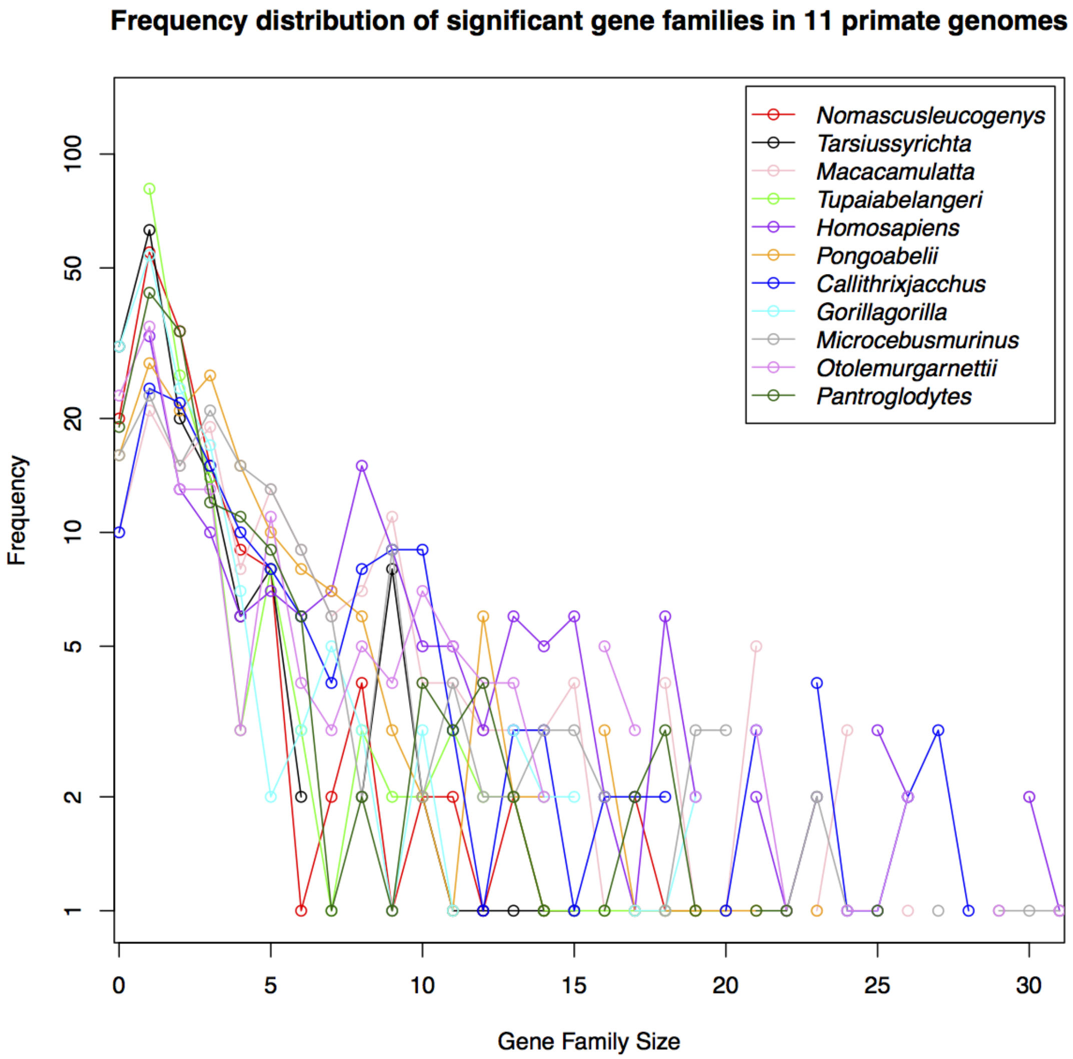


Figure 4.15 - Frequency distribution of significantly expanded or contracted gene family sizes in each primate for release 67 of the primates gene family data using a variable lambda for each gene family and across the tree. A cut-off of 30 used as maximum on the x-axis as this represents the majority of the data. A 0-size gene family means complete loss in that species.

Expansions and contractions of genes in 163 significantly changed gene families in primates

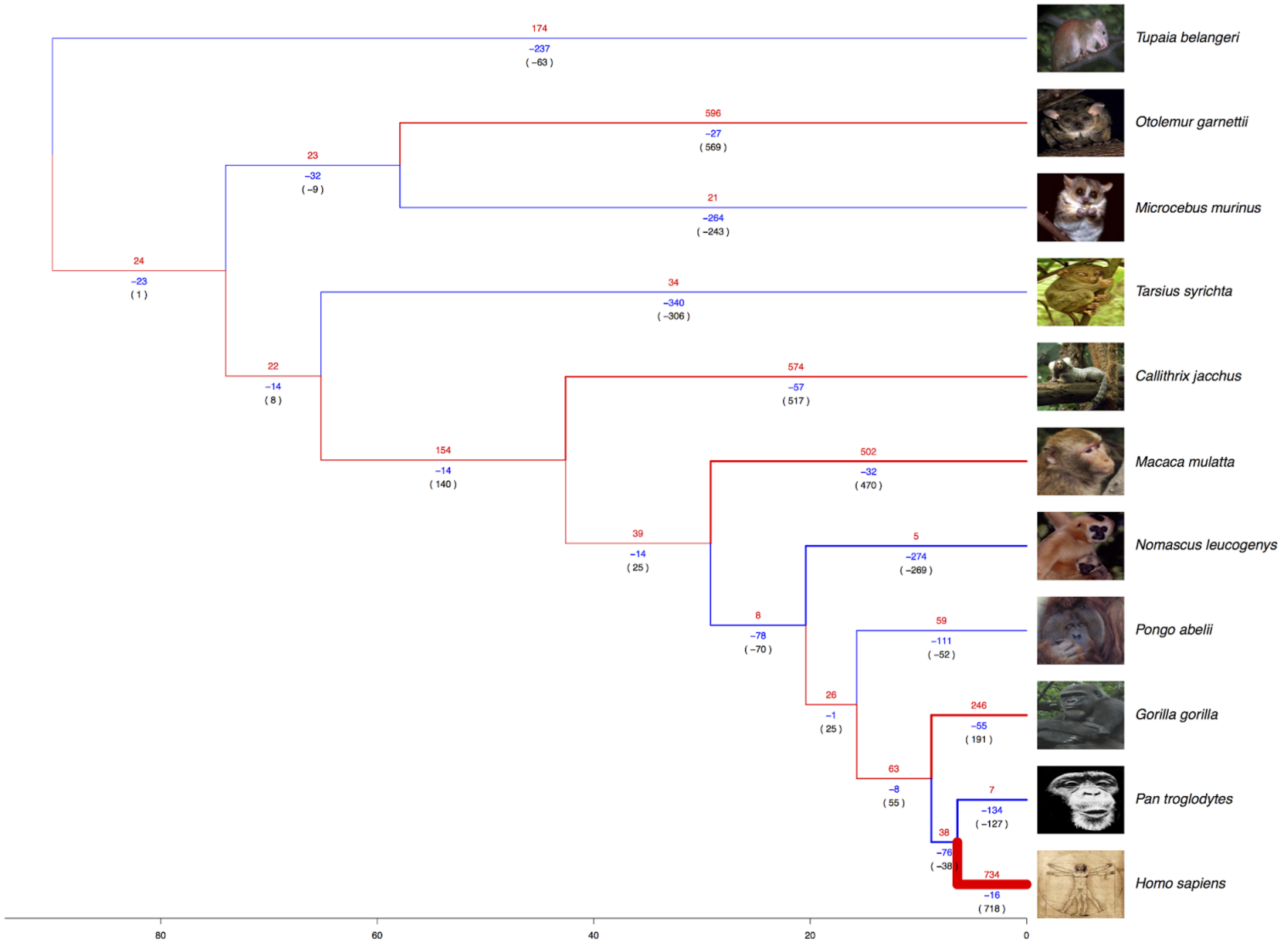


Figure 4.16 - Expansions and contractions of genes along the branches of the primate phylogenetic tree for release 67 of the primates gene family data using a variable lambda for each gene family and across the tree. Blue coloured branches show contractions, while red coloured branches depict expansions. Branch thickness is weighted by time since the ancestor node for each branch.

Functional classifications of significantly expanded or contracted primates gene families

Functional classifications of the significantly expanded or contracted gene families were retrieved via the Ensembl Compara API.

Release 66 data

The fixed lambda gene family CAFE data was used to retrieve information on the functional nature of the 538 gene families identified as being significantly expanded or contracted.

This analyses retrieved 150,440 entries from the Ensembl Compara database relating to 517 of the 538 gene families. Functional classifications were missing for 21 families. Out of the 150,440 raw entries, there were 392 unique descriptions annotated by the Ensembl Compara pipeline (see Appendix 4.4). A total of 15,979 unique gene IDs existed within these data. These genes fell into two biotypes; `protein_coding` and `IG_V_gene`, of which 150,427 and 13 entries belonged to those classifications respectively. The genes were located in only 4,039 different locations, corresponding to their top level location association assigned by Ensembl.

The metadata that Ensembl assigns via external reference to the GO Database returned 2,911 GO IDs and 2,891 GO definitions respectively, belonging to one of 4 different GO domains. The GO domains assigned were `molecular_function` (MF), `cellular_component` (CC), `biological_process` (BP) and NULL.

The family description assignment from Ensembl was used to plot the relative contribution of each of the 392 values for all species collectively (see Figure 4.17) and for each species individually (see Figure 4.18).

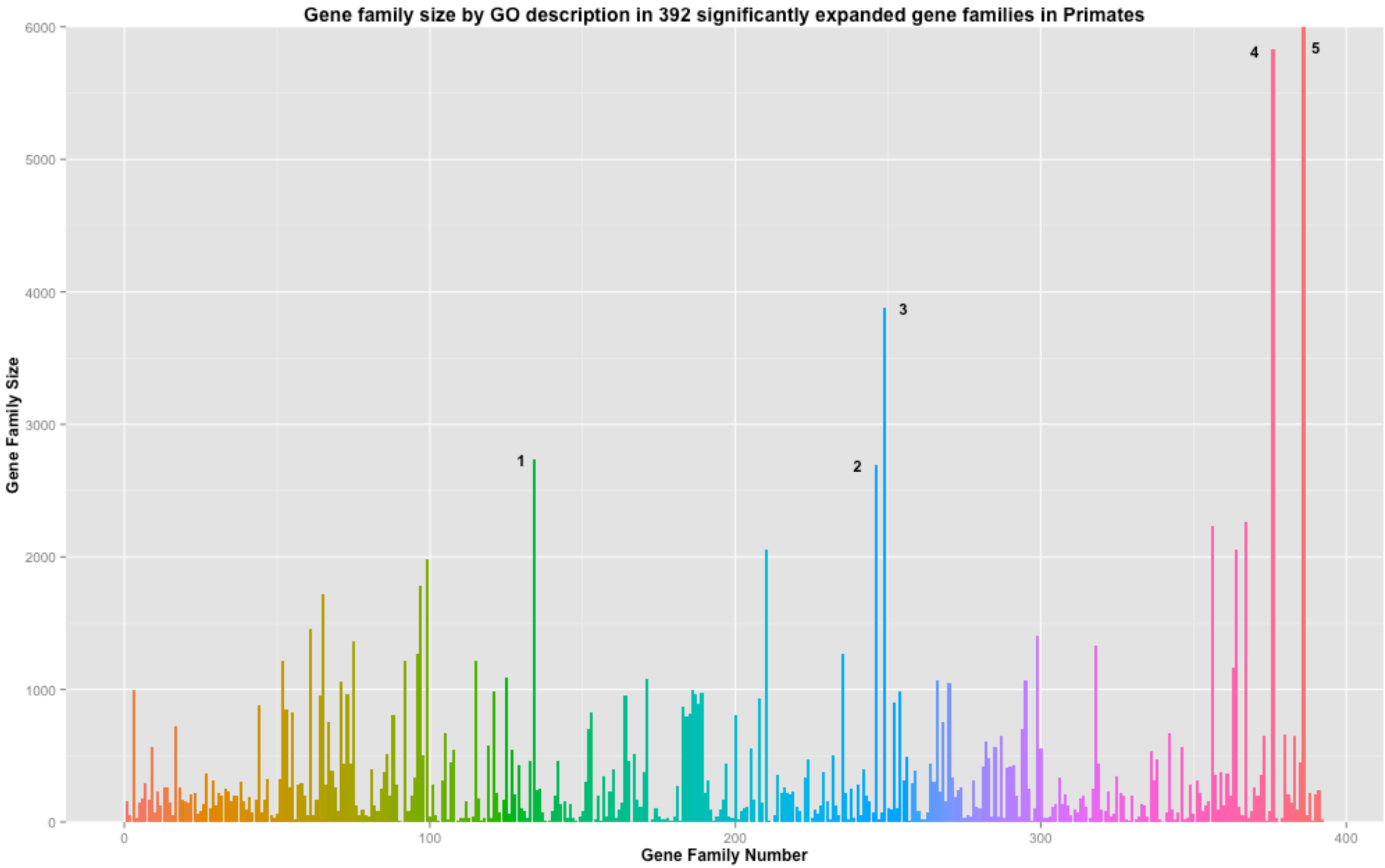


Figure 4.17 - Bar chart showing a breakdown of the functional classifications of significantly expanded or contracted gene families for the release 66 primates gene family data. All species data are pooled. Annotations correspond to values above an arbitrary cut-off of 2,500 members. Annotations are: 1) Epithelial Discoidin Domain Containing Receptor, 2) and 3) Olfactory Receptor, 4) Unknown, 5) Zinc finger (truncated at 6,000 – actual size 18,306). See Appendix 4.4 for list of gene family descriptions.

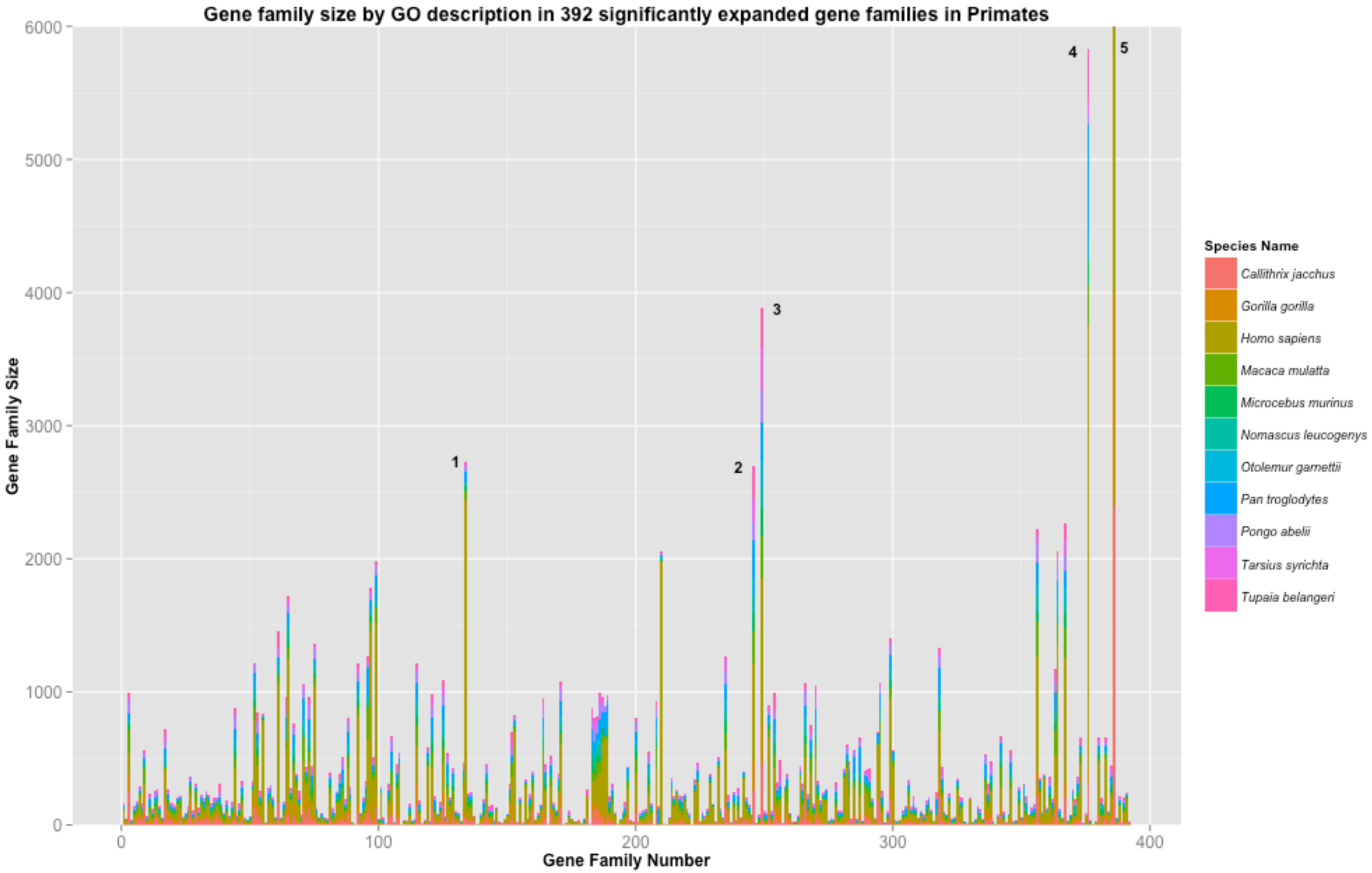


Figure 4.18 - Bar chart showing a breakdown of the functional classifications of significantly expanded or contracted gene families for the release 66 primates gene family data. All species data are represented individually to highlight per species contributions. Annotations correspond to values above an arbitrary cut-off of 2,500 members. Annotations are: 1) Epithelial Discoidin Domain Containing Receptor, 2) and 3) Olfactory Receptor, 4) Unknown, 5) Zinc finger (truncated at 6,000 – actual size 18,306). See Appendix 4.4 for list of gene family descriptions.

Release 67 data

The variable lamba individual gene family CAFE data was used to retrieve information on the functional nature of the 163 gene families identified as being significantly expanded or contracted.

This analyses retrieved 80,801 entries from the Ensembl Compara database relating to 162 of the 163 gene families. Functional classifications were missing for 1 family. Out of the 80,801 raw entries, there were 122 unique descriptions annotated by the Ensembl Compara pipeline (see Appendix 4.5). A total of 10,369 unique gene IDs existed within these data. These genes fell into four biotypes; `protein_coding`, `IG_V_gene`, `IG_C_gene`, and `TR_V_gene` of which 80,616, 141, 32 and 12 entries belonged to those classifications respectively. The genes were located in only 2,358 different locations, corresponding to their top level location association assigned by Ensembl.

The metadata that Ensembl assigns via external reference to the GO Database returned 1,389 GO IDs and 1,376 GO definitions respectively, belonging to one of 4 different GO domains. The GO domains assigned were `molecular_function` (MF), `cellular_component` (CC), `biological_process` (BP) and NULL.

The family description assignment from Ensembl was used to plot the relative contribution of each of the 122 values for all species collectively (see Figure 4.19) and for each species individually (see Figure 4.20).

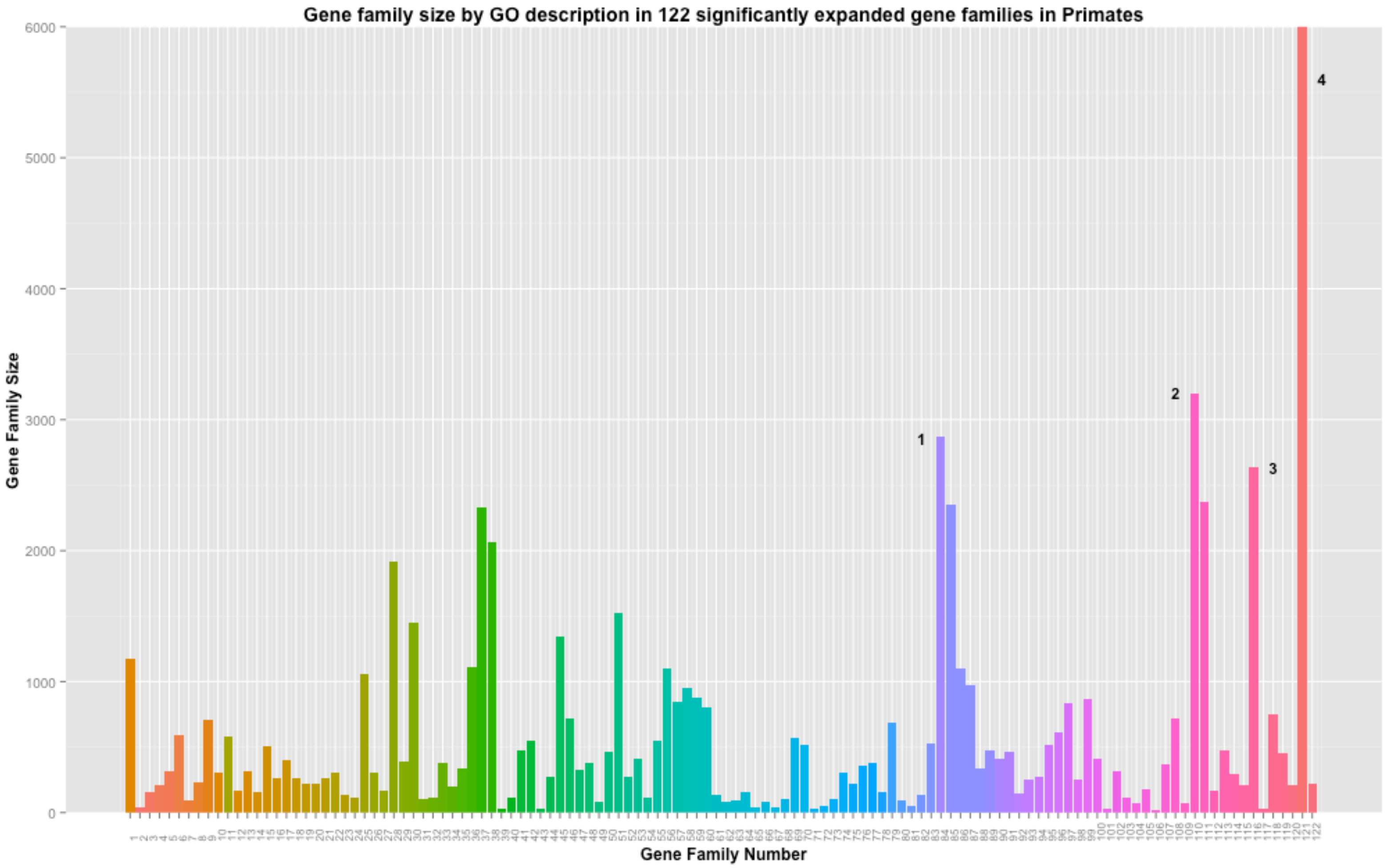


Figure 4.19 - Bar chart showing a breakdown of the functional classifications of significantly expanded or contracted gene families for the release 67 primates gene family data. All species data are pooled. Annotations correspond to values above an arbitrary cut-off of 2,500 members. Annotations are: 1) Olfactory Receptor, 2) Tripartite Motif Containing 3) Unknown, 5) Zinc finger (truncated at 6,000 – actual size 17,905). See Appendix 4.5 for list of gene family descriptions.

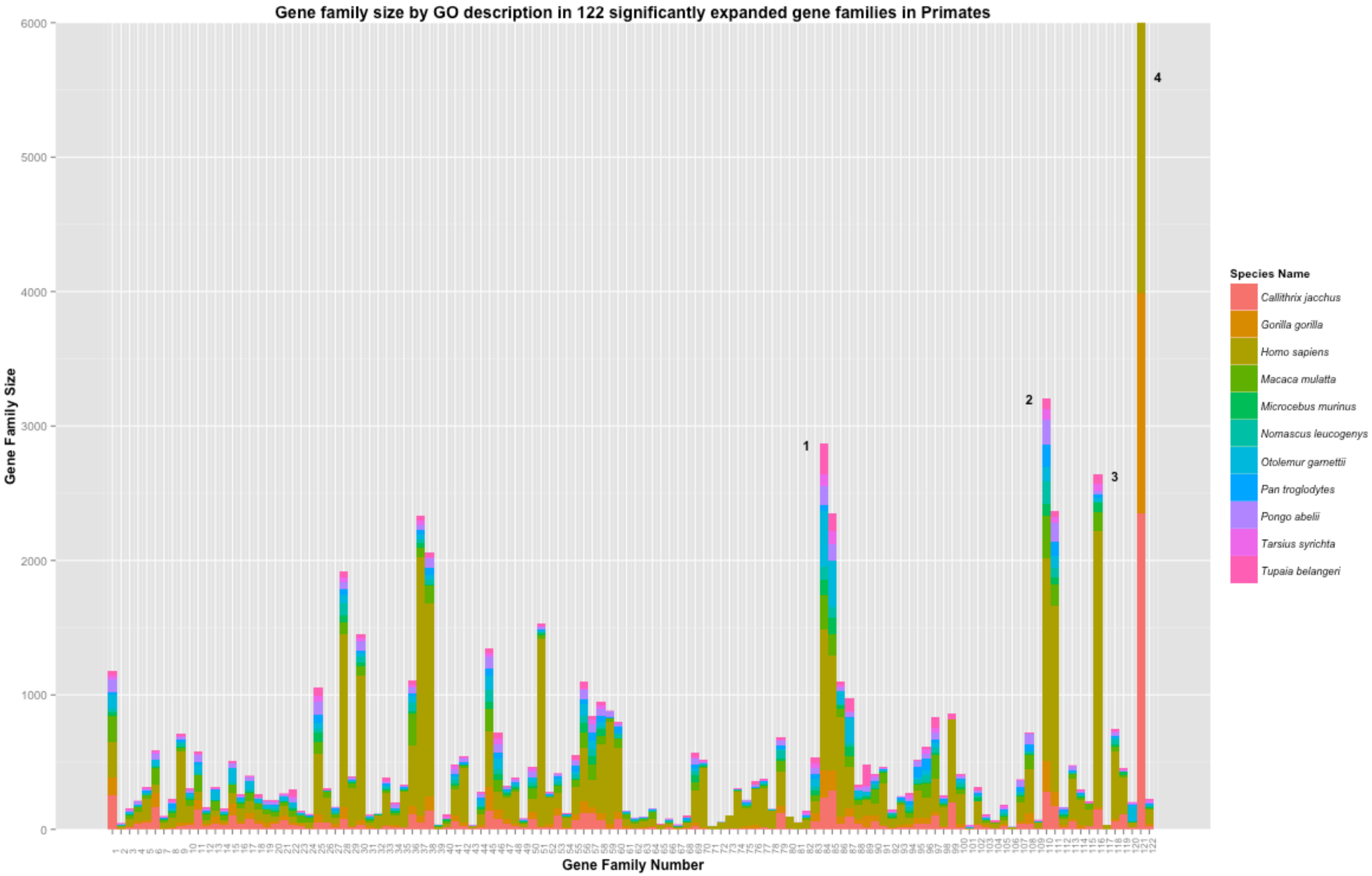


Figure 4.20 - Bar chart showing a breakdown of the functional classifications of significantly expanded or contracted gene families for the release 67 primates gene family data. All species data are represented individually to highlight per species contributions. Annotations correspond to values above an arbitrary cut-off of 2,500 members. Annotations are: 1) Olfactory Receptor, 2) Tripartite Motif Containing 3) Unknown, 5) Zinc finger (truncated at 6,000 – actual size 17,905). See Appendix 4.5 for list of gene family descriptions.

Comparison with previously identified functional classifications

A comparison was made to the findings of a previous study (Dumas *et al.*, 2007) that focused on gene copy number variation across 60-million years of primate evolution. Dumas *et al* identified a total of 186 gene names from clusters having copy number changes with LS changes in humans. Of those 186 gene names, 15 were classified as having noteworthy LS amplifications. These were queried against Ensembl and 183 Ensembl gene IDs retrieved. Of those 183 gene IDs, 107 Ensembl gene family IDs were retrieved for comparison. All 107 family IDs were found within the 41,159 raw gene family IDs retrieved from release 67 of the Ensembl database. Only 77 matched the trimmed dataset of 11,024 family IDs given as input to CAFE. A total of 25 family IDs were recovered as being significantly expanded or contracted by comparison with the fixed lambda CAFE run (see Appendix 4.6.

Of the 15 noteworthy LS amplifications, 12 Ensembl gene IDs were identified belonging to 10 family IDs. All 12 of these were found within the 41,159 raw gene family IDs dataset, with 11 in the trimmed dataset and only 2 in the fixed lambda CAFE run. These 2 changes were in the same gene family with the *ID* ENSFM00250000000661 having the *DESCRIPTION* GAMMA GLUTAMYLTRANSPEPTIDASE PRECURSOR GGT (see Appendix 4.7).

Discussion

Analyses of gene families

The analyses to identify significantly expanded or contracted gene families in primates produced some unexpected results, highlighting a large relative expansion in gene family sizes in the branch leading to modern humans (see Figure 4.4), though confirming the findings of previous studies (Huynen and van Nimwegen, 1998; Koonin *et al.*, 2002; Karev *et al.*, 2002; Barabási and Oltai, 2004), which state that gene family sizes approximate a power-law distribution (see Figure 4.2). Other interesting findings included a large loss in gene family sizes in the branch leading to the Philippine tarsier (*Tarsius syrichta*). It is unclear why these findings have not been previously identified. Many of the initial genome sequencing projects work with low coverage immature raw sequence data that has yet to be comprehensively scrutinised and validated by the scientific community, although that is certainly not the case here with humans. The tarsier was however sequenced as part of the Mammalian Genome Project (Lindblad-Toh *et al.*, 2011) at 1.82X coverage in order to provide intermediate scaffolds between existing mammalian species to help improve their annotation, and so is more likely to represent an incomplete assembly and annotation itself. In addition, the approach used here is novel, and the first time such analysis has been undertaken in this way.

Given the multiple possible sources of error that can exist in the process of sequencing, assembling and annotating a genome (Brenner, 1999; Devos and Valencia, 2001; Alkan *et al.*, 2010), the outcome of any analyses needs to be scrutinised in great detail. There could be multiple confounding factors that impact on the findings in this chapter. Many of the primates have their sequences aligned and mapped back to humans for example, and as the human genome is the most complete, a gene that exists in the human genome yet doesn't in another due to a low coverage assembly may mistakenly be identified as a loss in that species. Likewise, this could also result in the annotation of a gain in copy number in the human genome when it is really down to mapping against a poor assembly. A comparison was therefore undertaken with the available rodent gene family data to determine if these findings were due to the nature of the human data, an artefact of the approach used in identifying significantly changed gene families, or real biological change.

Understanding the data

There has been an increasing medical investment in genomics (The 1000 Genomes Project Consortium, 2011; Gibson and Visscher, 2013; Zhang *et al.*, 2013) and diversity panels (e.g. The 1000 Genomes Project Consortium, 2011) have made the human genome data rich in comparison to other species. The focus on population level re-sequencing is limited to a small group of model organisms as well as domesticated, horticultural, and agricultural species (Ellegren, 2014), though the medical interest in human genomics has seen projects undertaking sequencing at a much larger scale (The 1000 Genomes Project Consortium, 2011; Hall *et al.*, 2013). This likely makes the human genome the most well annotated and understood of all genomes. If the human lineage specific expansion of gene families observed in this study was an artefact of being such a relatively well annotated genome, we should expect to see a similar expansion on the branch leading to *Mus musculus* relative to the rodent phylogeny. In contrast to the 538 gene families identified as being significantly changed in primates (see Figure 4.4), only 414 gene families were significantly changed in rodents. The expectation of an expansion in the branch leading to *Mus musculus* was rejected following visualisation of the significant gene family data, where we instead see a large relative expansion towards the base of the tree (see Figure 4.8). Before making any inferences using these outcomes however, it is wise to try and understand the data in more detail.

The sizes of the trimmed data (see Materials and Methods) used as input for CAFE were 11,307 for primates and 11,947 for rodents, which are approximately the same, though they are taken from full data sets containing 49,737 and 25,610 gene families respectively (see Table 4.1 and Table 4.3). The total number of gene families is approximately double in primates compared to rodents. The reason behind this is unclear. It is possible that this could be an artefact of looking at only 6 primate genomes, compared to the 11 primate genomes. Downsampling the primate analyses may be helpful, though it is probable that population size has an effect here. Primates tend to have longer generation times and smaller population sizes. The shorter generation times and larger population sizes in rodents are likely therefore to impact the ability of duplicates to become fixed within any population. This is because the duplicate landscape will change more frequently in such a population reducing the likelihood that any one duplicate can drift to fixation. A higher duplicate death rate (as well as a lower birth rate) in *Mus musculus* in comparison to *Homo sapiens* adds weight to this argument (Lynch and Conery, 2003a). The lower number of significant changes in *Mus musculus* might therefore be expected, though conversely, the influence of these population variables could also be argued as resulting in higher rates of evolution (Martin and Palumbi, 1993; Bromham, 2009; Santos, 2012; Steipera and Seifferte, 2012), meaning we might have expected to see more significant changes in the rodents.

The differences between the two datasets seems negligible in terms of the input numbers given to CAFE, however the maximum gene family sizes for each primate species differs considerably from the maximum values seen in the rodents, with the exception of *Mus musculus* (see Table 4.2 and Table 4.4). When looking at the divergence times of these species (see Appendix Table 4.8) we can see that the distance between the available rodents represents a more regularly sampled time gradient in primates, with the five members of the *Hominoidea* superfamily diverging within the same period of time since the common ancestor of *Mus musculus* and *Rattus norvegicus*. When assembling and annotating the genomes of these species, sequence data is often necessary from relatively closely related species in order to build more comprehensive scaffolds, as a form of intermediate reference. If the divergence times are too great then elements that may actually be conserved between more disparate species are missed. This was an issue for early genome scientists when the available genome assemblies were hugely deviating (human, mouse, chicken), and thus intermediate species were sequenced to assist in building a more accurate picture of these genomes from a comparative evolutionary perspective (Lindblad-Toh *et al.*, 2011). As the rodents are more divergent, it is likely that the differences between the gene family data are influenced by factors surrounding their assembly and annotation (see Appendix 4 – Table 4.8 and Table 4.9), as well as their respective sequencing protocols.

The nature of the data

The variability in the quality of the underlying data used in analyses of this nature are the most likely cause of inconsistencies in the results. Reaching robust conclusions from the results of error prone computational experiments is extremely difficult and it is hard, if not impossible to account for all sources of error. It is possible to consider some areas that might introduce bias however, and in analyses of this type incorrectly collapsed duplicates (ICD) are likely to have a big influence (Brenner, 1999; Devos and Valencia, 2001; Khaja *et al.*, 2006). ICDs are highly variable and can be directly linked to sequencing coverage, sequencing quality, and assembly protocol. Indeed, the sequencing coverage, quality and protocols used in these species reflects this (see Table 4.5 and Table 4.6). Movement towards cheaper and more accurate sequencing, in addition to maximum-likelihood model based comparison of genome assemblies is liable to improve things moving forward (Medvedev and Brudno, 2009; Clark *et al.*, 2013; Hunt *et al.*, 2013; Ghodsi *et al.*, 2013).

Ensembl pulls in data from external sources in order to provide experimental validation of their *de novo* gene predictions on a continual basis. This occurs roughly every three months and utilises new assemblies as they become available. The data retrieved from external sources for validation of the human annotations however, are more extensive than any of the other genomes in Ensembl. The effect of the fast pace of experimentation involving human genome data is likely to compound this. How different is the mouse data from the human data, however? Both of these projects are extensively funded, have regularly updated assemblies and annotations, and benefit from additional manual annotation from the Wellcome Trust Sanger Institute's HAVANA group. The additional sources of data for the human genome are likely to improve specific areas of the genome's annotation, but conversely there is little change in raw counts over time (see Table 4.1 and Table 4.7), as quality of assembly and annotation moves closer to the optimum, however the number of significant gene family changes actually increased from release 66 to release 67 (see Figure 4.4 and Figure 4.12).

We might consider CNV as accounting for differences in these data, with the multiple sources of data aggregated into the composite human genome resulting in overrepresentation of the duplicate landscape. Although this is plausible, genome assemblies are constantly updated over time. Problematic areas are highlighted and issues such as over- or under-representation of duplicates at particular loci are addressed based on new experimental evidence. All these data are used to build a more accurate reference genome, with variation data being stored separately for additional analytical purposes (Chen *et al.*, 2010; Rios *et al.*, 2010). If CNV, due to increased population level sequencing, were the cause of the apparent gene family changes in humans we might expect to see a similar, albeit less pronounced, pattern in the mouse genome. This would be dependent on similar levels of population re-sequencing, however, and comparable levels of sequence divergence between individuals, which is unlikely due to the much greater focus on human population level genomics (The 1000 Genomes Project Consortium, 2011; Hall *et al.*, 2013). We instead see a relatively large expansion towards the base of the rodent phylogenetic tree (see Figure 4.8). It is useful to remember here that the figure weights the branch thickness by branch length from node to the LCA for each branch of the tree. The relative numbers of changes are clear however, with 843 expansions and 117 contractions in humans and only 38 expansions and 23 contractions in mouse (see Figure 4.4 and Figure 4.8). Divergence time is a median of 6.1 Mya between human and chimp in comparison to 22.0 Mya between mouse and rat (see Appendix 4.8 and Appendix 4.9), so more adequate sampling between mouse and rat may have an impact on these results. Having higher sampling within a clade and thus more closely related species to compare to, can assist in identifying less divergent changes and pinpoint elements that are conserved between those species. With more divergent species the ability to identify duplicates may thus be more difficult due to greater sequence divergence. This effect is unlikely to be large in placental mammals, but certainly more profound in the rodent genomes used here than in the primates.

It is possible to exclude species from the analyses to test for the introduction of bias from a particular branch, or to introduce an artificially greater divergence between certain species. If the nature of the human data were having an impact on the results across the tree, we would expect to see changes when removing the human data from the CAFE analyses. Doing so appears to have little effect on the results however (see Appendix 4.10). It is likely, therefore, that the nature of the data in relation to the underlying sampling and its impact on assembly and annotation quality has more repercussions for the results, though there is also an effect seen by differing the parameters used in the CAFE analyses.

Evaluation of the approach

CAFE allows a number of different inputs when estimating the probability of gene gain and loss over time (lambda). In particular, it allows the user to give explicit lambda values for each node of the tree or has the ability to search for the values that maximise the log-likelihood across the tree as a whole. When searching the parameter space for the most likely lambda, different values can be estimated for different nodes and likewise some nodes can be fixed with static values. Additionally, it is possible to estimate lambda values for each gene family individually. When using the individual gene family approach, it isn't possible to estimate values for all families combined, and requires manual aggregation of the data for comparison with results reached using different parameters.

Using different parameters for CAFE allows us to address the data with different assumptions about their evolutionary history. Using a fixed lambda assumes that the probability of gain and loss is equal across the tree, though this is unlikely to be the case; both in terms of the nature of the data and different the different evolutionary and life history characteristics of the species in question. It is therefore more likely that a variable lambda reflects the significant changes in the duplication landscape more accurately. A fixed lambda is separate from fixed birth (B) and death (D) parameters, which are considered to be equal in CAFE, thus B and D are equally as likely using this model regardless of the probability of gain and loss along the branch. Previous research has shown us that B and D rates differ between species (Lynch and Conery, 2003a), so this is also an unrealistic limitation of the model. Using an explicit or estimated lambda that varies at different nodes is likely to lead to a more realistic representation of the underlying expansions and contractions of gene families.

By utilising a fixed and varied lambda with the release 67 data there is a difference in the estimated significantly changed gene families of 626 for fixed (see Figure 4.12) and 309 for varied (see Figure 4.14) respectively. Although there is little change in the values seen for primates, there is a big change in the outgroup species *Tupaia belangeri*, which exhibits 298 expansions and 430 losses in the fixed tree, but only 3 gains and 1 loss in the varied tree. The assembly information shows that this genome is subject to poor coverage (see Table 4.5), and is thus likely of dubious quality. With this being a divergent outgroup with no close relatives in the Ensembl databases our confidence in the annotation of family sizes is reduced, but even still this is an unexpected and unlikely result. There are also expansions where there were previously contractions in the basal branch leading to the common ancestor of the primates, and the branch leading to common ancestor of *Otolemur garnettii* and *Microcebus murinus*. *Microcebus murinus* is, like *Tupaia belangeri*, a low coverage genome, however *Otolemur garnettii* is very high coverage, which may be the reason for this variability towards the root of the tree between the fixed and varied runs.

The most likely scenario, from an evolutionary perspective, is variation in the probability of gain and loss in both individual taxa and individual gene families. By emulating these assumptions via the CAFE software we reach an outcome of 163 significantly changed gene families (see Figure 4.16). The discrepancies in the outgroup between the fixed and varied runs are restored, though we instead see expansions along the branches leading to the common ancestor of the primates and the common ancestor of the *Haplorhini* (prosimian tarsiers and the anthropoids). The change in humans is more modest at 734 expansions and 16 contractions, given the > 1000 expansions and ~30 contractions in the fixed and varied runs. As this method reduces overall bias imposed by both individual gene family sizes and individual taxa, it is reasonable to suggest it represent the most likely state of significant gene family changes in these species. It is clear however that, as with many other bioinformatics algorithms, varying the parameters and assumptions made by these models can have big impacts on the conclusions that we draw from them.

There are a number of other programs that take varying approaches to the identification of gene family data, as well as discussing the limitations of the CAFE software in both terms of its underlying model and flexibility in parameters (Ames *et al.*, 2012; Liu *et al.*, 2011; Vieira and Rozas, 2011). These software primarily focus on parsimony and maximum likelihood models (Ames *et al.*, 2012; Vieira and Rozas, 2011), but show that there is no significant impact on the results, given species that diverged less than ~100 Mya (Ames *et al.*, 2012). The power of Bayesian statistics to account for more diversity in the underlying distributions makes it more powerful however in widely divergent species (Liu *et al.*, 2011; Rannala and Yang, 2003; Yang and Yoder, 2004; Mayrose *et al.*, 2004; Beerli, 2006). CAFE has certainly stood the test of time, being an integral tool in a numbers of studies (Hahn *et al.*, 2005; De Bie *et al.*, 2006; Hahn *et al.*, 2007; Han *et al.*, 2013). Its adoption by Ensembl (Flicek *et al.*, 2012) to include information on significant changes across the entire Ensembl species tree, emphasises how it is considered as the most robust approach to this problem. Many of its limitations, as highlighted by alternate software (Ames *et al.*, 2012), such as an inability to account for biases in the underlying sequence quality or coverage that can often impact on the accuracy of the genome assembly and annotations for individual species, have now been addressed with the release of CAFE 3 (Han *et al.*, 2013). Error models have now been incorporated into the CAFE algorithm that allow weights to be applied to each gene family for each of the species. CAFE can then infer models of gene family gain and loss, and change in gene family size independently for individual gene families and species, in the presence of potential errors. This should greatly improve the inference of significant gene family changes across the phylogenetic tree.

Real biological change?

The analyses here have considered and accounted for the effects of branch length, individual species bias and variance in gene family size across the tree as a whole. Previous research shows that these have little impact on the results (Hahn *et al.*, 2007), though their approach groups often quite divergent species into the same lambda categories, assuming equal birth and death rates along those branches, and thus reducing the ability to highlight such effects. The data presented here show that when estimating different lambdas for each branch of the tree, the impact of individual gene family size is more prominent (see Figure 4.14 and 4.16). This approach is likely to be the most accurate as it removes the assumption of equal change across all branches and in individual gene families, therefore reducing bias in the results.

To support this we can consider the birth and death rates of duplicate genes in humans (*B*=0.0049, *D*=0.081), which have been shown to be asymmetric (Lynch and Conery, 2003a; Goodstadt and Ponting, 2006). CAFE assumes an equal birth and death rate, which Hahn *et al.* (2007) determine as being no more likely to result in rejection of the null hypothesis when compared with results given a much higher birth rate. Their conclusions assume that the human and chimp lineage have an equal rate of birth and death (B,D=0.0039), which although similar to the previous birth estimates in humans, varies considerably from death rates (Hahn *et al.*, 2007; Lynch and Conery, 2003a). The difference in sequence divergence and copy number variation between these species doesn't seem to support the decision to apply an equal rate for B and D to these species (Demuth *et al.*, 2006; Marques-Bonet *et al.*, 2009). These birth and death rates are of course averages and so applying an average for the entirety of the duplications across the tree or within a clade seems counter-intuitive. By estimating lambda for each branch and gene family individually, we reduce this effect, and can achieve more accurate results (B,D=0.0068 in humans, B,D=0.0056 in chimps; see Appendix 4.3). Changes in the CAFE software should improve the accuracy even further and indeed by accounting for error in annotation and assembly a higher rate for humans is seen (B,D=0.0062) that is closer to our results, though this still assumes equality in the probability of birth and death of duplicates for humans and chimps, as well as an equal birth and death rate (Han *et al.*, 2013).

Given the variability in the quality of data from the raw reads, the reference assembly and associated annotation, it is difficult to reach a solid conclusion that the expansion in the branch leading to modern humans is the result of real biological change in gene family size over the last 6.1 million years. In addition, there are various sources of methodological bias, however I attempt to account for these limitations in this study and believe the results represent a more accurate picture of the landscape of significant expansions and contractions across the available primate genomes.

Functional classifications of significantly expanded or contracted primates gene families

In order to understand whether adaptation plays a role in the significant gene family changes highlighted here we can examine the functional classification of these data. The evolutionary fate of duplicates can be influenced by environmental pressures, and thus those genes that have an adaptive benefit are more likely to be fixed within a population (Lynch and Conery, 2000; Lynch and Conery, 2003a). By critiquing these annotations in relation to previous studies, we can provide additional weight to support or reject our finding of large significant changes along the branch leading to modern humans. Adaptation to a changing environment is commonplace in nature (Sandve *et al.*, 2008; Chen *et al.*, 2008; Turner *et al.*, 2010; Fischer *et al.*, 2011; Sheik *et al.*, 2011; Oh *et al.*, 2012; Wu *et al.*, 2012; Chao *et al.*, 2013; Jiang *et al.*, 2013; Norman, 2014) and there is a great deal of evidence to suggest its occurrence during human evolution (Pronk *et al.*, 1982; Perry *et al.*, 2007; Zozulya *et al.*, 2001; Young *et al.*, 2008). In particular previous studies have highlighted expansions in gene families involved in reproductive, immunological and development roles (Hahn *et al.*, 2005; Demuth *et al.*, 2006), with various expansions in the great ape lineage (Mikkelsen *et al.*, 2005; Marques-Bonet *et al.*, 2009).

Making accurate inferences on the adaptive landscape of the significantly changed duplicates requires accurate data across all stages of the process. As previously discussed, errors in the underlying raw sequence data can propagate through higher level stages, resulting in flawed data for analyses. The different assumptions made by analytical programs also impacts the results in varying degrees. The release 66 fixed (see Figure 4.17) and release 67 varied individual (see Figure 4.19) data sets agree on the functional classifications of a number of significantly changed families, however conflict in raw number of unique descriptions (r66=392, r67=122), which emphasises how differing levels of annotation and the parameters and assumptions made in the analysis can impact on the outcome (Ames *et al.*, 2012; Han *et al.*, 2013).

Values over a threshold frequency of 2,500 are taken for further discussion as these represent the most abundant. There are 5 peaks highlighted above this cut-off (Epithelial Discoidin Domain Containing Receptor, 2 Olfactory Receptors, Unknown, and Zinc Finger). These families all fall within the broad classifications of having reproductive, immunological or developmental functions. Class II Histocompatibility Antigen falls within the broader Major Histocompatibility Complex, which are involved in presenting peptides to CD4+ lymphocytes as part of the innate immune system. This family has a UniProt (Apweiler *et al.*, 2004) BP annotation of “Immunity”, however GO lists a wider range of BP annotations including “T cell receptor signalling pathway” and “antigen processing and presentation of exogenous peptide antigen via MHC class II”. The Epithelial Discoidin Domain Containing Receptor is an epithelial cell membrane protein involved in developmental processes such as cell migration, differentiation, survival and cell proliferation. It has UniProtKB BP annotations of “Lactation” and “Pregnancy” however has a broad range of GO BP annotations including “ear development”, “embryo implantation” and “regulation of cell-matrix adhesion”. The Olfactory Receptor is a cell membrane protein responsible for odour reception. It has UniProtKB BP annotations of “Olfaction” and “Sensory transduction” and a GO BP annotation of “detection of chemical stimulus involved in sensory perception of smell”. Zinc Finger is a broad classification, likely encompassing a wider set of proteins that contain this structural motif. Previous studies have excluded this family to assess its impact on the results of their analyses (Hahn *et al.*, 2007), however it was show to have little effect on the outcome and additionally the approach taken with the r67 variable lambda individual gene family dataset here doesn't bias the results of other families with its inclusion. The UniProtKB BP annotates these proteins as being involved in "Transcription" and "Transcription regulation" and GO BP annotations include "regulation of transcription" and "transcription, DNA-dependent". These families alone confirm the findings in previous studies and support our hypothesis that these changes are likely adaptive. There are also other interesting findings (see Appendix 4.4 and Appendix 4.5) that are supported by the literature, including a significant expansion in humans of the Alpha Amylase Precursor EC_3.2.1.1 family. This protein family has the HGNC (HUGO (Human Genome Organisation) Gene Nomenclature Committee) prefix AMY1A, which is a salivary amylase and is thought to be expanded due to adaptation to an increased amount of starch in the diet following the advent of agriculture (Pronk *et al.*, 1992; Perry *et al.*, 2007).

The functional analyses highlight a large number of "UNKNOWN" and "AMBIGUOUS" annotations in these data, with the former class being above our 2,500 frequency cut-off. There could be a number of reasons behind this. When focusing on the "UNKNOWN" class, the most obvious reason is that these are *de novo* predictions that have no evidence to support them. However, along with the "AMBIGUOUS" classification these groupings could also be the result of error in the underlying raw sequence data and assembly. If short low quality or highly repetitive reads that form overlapping segments are incorrectly collapsed as duplicates (Brenner, 1999; Devos and Valencia, 2001), they may subsequently have no hits against known proteins ("UNKNOWN" class) or have low similarity hits against one or more known proteins ("AMBIGUOUS" class). As sequence quality, assembly and annotations improve, along with the development of more robust models that reduce or improve on the assumptions made (Han *et al.*, 2013), there should be a reduction in these numbers, which these data support (r66 frequencies; "UNKNOWN"=5833, "AMBIGUOUS"=1456, r67 frequencies; "UNKNOWN"=2637, "AMBIGUOUS"=305).

By plotting the proportion of the functional data that is influenced by individual species (see Figure 4.18 and Figure 4.20) it is possible to highlight where there may be a bias in particular species due to overrepresentation in the underlying data. It is clear that there is a large influence from humans in both the release 66 and release 67 data. This inequality in the data may be representative of the underlying patterns in the change in gene family sizes to a degree. Due to the difference in quality of these genome annotations however, with humans being by far being the most well annotated (due to being the most well studied), there is likely to be an unfair representation of the true values of the non-human genomes. Those species that are more divergent or within less well sampled clades are likely to be impacted even more so, as it is more difficult to build an accurate assembly for these species due to increased sequence divergence. If only a single individual is used to represent the species, then CNV may decrease our confidence in the outcome of gene family annotations. Additionally, the lower coverage of many of these genomes compounds this issue as the resulting quality of the raw data and assemblies are likely to lead to increased errors, particularly when attempting to correctly collapse duplicates.

By comparing the functional classifications for humans determined here with a previous study (Dumas *et al.*, 2007) we can assess the effect of assembly, annotation and approach even further. The results reached herein vary considerably from those obtained by Dumas *et al.* The impact of varying the approach taken and the annotations used in this study alone emphasise these differences (release 66=25 matches, release 67=2 matches). We must also consider the data and approach taken by Dumas *et al.* Their data is based on the UCSC hg18 annotations (Kuhn *et al.*, 2007) built on top of the NCBI36 assembly from 2006. Our data utilises the Ensembl release 67 annotations from May 2012 (Flicek *et al.*, 2011), which is based on the GRCh37 assembly, in particular patch 7, which was updated in January 2012. The last release of Ensembl to use the NCBI36 human genome assembly was release 54 from May 2009 (Hubbard *et al.*, 2008). The methods used to identify gene copy numbers was vastly different too. Dumas *et al.* used array comparative genomic hybridization (aCGH) of novel biological samples, mapping the position of cDNA clones with high fluorescence ratios to the UCSC assembly and retrieving annotations accordingly. They additionally performed BLAT (Kent, 2002) searches using the GenBank accessions corresponding to their cDNA clones, to determine whether lineage specific (LS) changes corresponded to the values retrieved via analyses of their aCGH data.

The method used by Dumas *et al.* (2007) is considerably different from the predominantly automated computational identification of copy numbers used by Ensembl. Although Ensembl utilise sources of experimental data as evidence for their annotations, their computational annotation procedure is biased towards introducing errors (Brenner, 1999; Devos and Valencia, 2001; Alkan *et al.*, 2010) that are less likely when manually annotating the data (though see Iliopoulos *et al.*, 2003; Curwen *et al.*, 2004; Potter *et al.*, 2004). Although Dumas *et al.* take a different approach and reach widely different conclusions, their method is no more accurate or conclusive than my own. Studies based on their approach are known to suffer from the poor quality of cDNA libraries (Hubbard *et al.*, 2002; Paszkiewicz and Studholme, 2010). This is compounded by the known limitations of aCGH (Weiss *et al.*, 1999; Oostlander *et al.*, 2004). More recent studies emphasize these effects even further and discuss the importance of considering the nature of the data in addition to its biological context when making inferences based on the results (Gazave *et al.*, 2011; Sudmant *et al.*, 2013; Han *et al.*, 2013).

Sudmant *et al.* (2013) in particular use an approach that takes 97 sequenced genomes of multiple individuals within populations across the Great Ape phylogeny. These genomes are sequenced to high coverage, including 75 at 25X coverage on the Illumina HiSeq 2000 platform as part of the Great Ape Genome Diversity Project (Prado-Martinez *et al.*, 2013), and others from the Orang-utan Genome Project and Denisova Genome Project. The raw reads were aligned back to the UCSC hg18 NCBI36-based human reference genome making use of their high quality genome annotations (Kuhn *et al.*, 2007) using mrsFASTc aligner (Hach *et al.* 2010). Read depth profiles were constructed, and following a number of correction and validation steps, including high GC content masking and aCGH, were used to determine absolute copy number of a loci across the populations and the numbers of deletions and duplications of genes across the Great Ape phylogeny.

As with Dumas *et al.* (2007) the results varied considerably, and in particular the extent of human expansions and contractions identified in my analyses was not reflected. A total of 407 lineage specific duplications, and 340 deletions were identified. This is in contrast to the 186 changes in copy number identified by Dumas *et al.* and the 163 changes identified in my CAFE analyses using a variable lambda for each gene family and across the tree (varied_indiv data). Of the 747 duplications and deletions identified by Sudmant *et al.*, this resolved to 11,503 copy number changes corresponding to a total of 4,831 unique gene names. In comparing the HGNC symbol to the Ensembl genome database using Ensembl BioMarts (Kinsella *et al.*, 2011) via the Bioconductor package (Durinck *et al.*, 2005; Durinck *et al.*, 2009), a total of 5,000 Ensembl gene IDs were returned. This is likely due to the existence of multiple gene copies in the Ensembl database that corresponded to the associated HGNC symbols. When comparing the retrieve Ensembl gene IDs from the Sudmant study with the non-significant varied_indiv data from my study, a total of 3,586 gene IDs were matched. When comparing with the significantly expanded or contracted varied_indiv data, 88 gene family IDs were matched of which there were 483 copy number changes, corresponding to a total of 474 unique gene IDs. Although the majority of the specific duplications and deletions discussed by Sudmant *et al.* (e.g. PRDM7, C1QTNF, AMACR, and BOLA2) were identified within the larger non-significant dataset, only the CYP2C18 gene family identified as being deleted in Humans and Chimps was recovered from the significant varied_indiv dataset.

The approach taken to assembling and annotating these data differs considerably, even in relation to Dumas *et al.*; though where the latter study focuses on using aCGH as a primary means of copy number identification, the Sudmant *et al.* study uses aCGH as a validation step to confirm the correct mapping of raw reads to the UCSC hg18 (NCBI36-based) reference genome. This method seems to work well in that the gene families identified as being expanded or contracted largely conform to my findings, however, there are still a number of gene families that aren't identified using their techniques. Only approximately 50% of the gene families identified as being significantly expanded or contracted in the varied_indiv part of my study were identified by Sudmant et al's method. Of the total 11,503 copy number changes in total this equated to 6,458 genes in my non-significant dataset. As previously mentioned the approach used to identifying copy number changes is only as good as the underlying assembly, and in using the older NCBI36 assembly to map their reads to, Sudmant *et al.* have fallen short of identifying the true repertoire of expansions and contractions across the Great Ape phylogeny. The robustness of the underlying (GRCh37.p7) assembly, alongside Ensembl's automated annotations, computational validation against the results of molecular experiments, and manual gene curation means their data is more likely to report a comprehensive and correct detail of the deletion and duplication landscape across the entire primate phylogeny, albeit with some bias towards more well studied organisms. This situation is likely to improve as population level sequencing and experimentation provides additional data for Ensembl to validate against.

Conclusions

The ability of bioinformatics algorithms to correctly identify the duplication landscape of the genome are limited, and an ongoing area of research in computational biology. There is great bias in the choice of algorithm used to determine duplications, with those that make more or different assumptions, reaching variable conclusions. The complexity in the identification of duplicates are confounded by numerous sources of error that stem from the choice of individuals, the sequencing technologies used, the assembly software chosen, and the annotation software applied. It is dubious, whether we have the power to account for all possible sources of error. The algorithms available certainly make such a vast number of assumptions that it is difficult to consider their impact when making predictions based on those data. At best, our inference of the structure of the genome is a rough guess based on a static representation of something that constantly changes through space and time.

It is clear however that computational advances have greatly improved the assembly and annotation of genomes over the past 3 decades. Improvements have been made at all levels of genome sequencing projects; from sequencing, through assembly, to annotation. Robust computational protocols have been produced to allow for what is by its nature a very error prone process. Manual annotation is the gold-standard, however the increasing scale of the data involved means that manually annotating, especially vertebrate genomes, is bound to become a dying art. New models that compute likelihood scores of genome assemblies allowing for comparison and determination of the maximum likelihood model of the genome are growing in popularity and will likely become the *de facto* standard in years to come. Additionally, more comprehensive algorithms and models are being continuously developed that can utilise the advances in computing hardware to achieve more accurate results in more reasonable runtimes. The more we are able to account for hidden assumptions and consider the nature of the data within its biological context, the more powerful our conclusions will be.

[This page is left intentionally blank]

CHAPTER FIVE: ANALYSES OF THE INFLUENCE OF GENE FAMILY SIZE AND GENOMIC LOCATION ON THE EVOLUTION OF INTRONS

Introduction

Mechanisms of gene family and intron evolution

Gene families have a big impact on the coding portions of the genome through processes such as non-homologous recombination, retrotransposition, and relaxed selection due to redundancy (Zhang, 2003). This makes them effective drivers of adaptive evolution in particular (Force, 1999; Lynch and Force, 2000; Lynch, 2002; Zhang *et al.*, 2002; Swanson, 2003; Francino, 2005; Wong and Wolfe, 2005; Han *et al.*, 2009; Ames *et al.*, 2010; Nygaard *et al.*, 2011; Iskow *et al.*, 2012; Lynch, 2012), but it is currently unclear how their presence and particularly number of gene members, might impact on the evolution of non-coding regions, such as introns. Introns are very important as they contribute a large proportion of nucleotides towards the total size of vertebrate genomes, often contributing more to overall gene length than exons and UTRs (untranslated regions) combined (Hong *et al.*, 2006; Roy and Gilbert, 2006; Yandell *et al.*, 2006; Stajich *et al.*, 2007; Gazave *et al.*, 2007; Zhu *et al.*, 2009; Jiang and Goertzen, 2011; Rogozin, *et al.*, 2012; Zhang and Edwards, 2012). They are also very effective in highlighting the neutral forces of molecular evolution, as well as regulatory mechanisms underpinning splicing and gene expression (Yu *et al.*, 2002; Yeo *et al.*, 2005; Will and Lührmann, 2005; Hong *et al.*, 2006; Farlow *et al.*, 2012). In order to understand how gene family characteristics might impact on intron evolution, one needs to relate the underlying molecular mechanisms of gene family size change to the gain and loss of introns over time. Gene families can change size in a number of ways; including polyploidy, aneuploidy, replication slippage, non-homologous recombination and retrotransposition (Zhang, 2003). Introns can also propagate through a number of mechanisms; including intron transposition, transposon insertion, tandem genomic duplication, intron transfer, intron gain during double-stranded break repair (DSBR), insertion of a group II intron and intronization (Crick, 1979; Sharp, 1985; Rogers, 1989; Hankeln *et al.*, 1997; Irimia *et al.*, 2008; Roy, 2009; Li *et al.*, 2009; Yenerall and Zhou, 2012). Their loss has generally been limited to the mechanisms of gene conversion or genomic deletion, however (Derr and Strathern, 1993; Roy and Gilbert, 2006). By comparing and contrasting these mechanisms it is possible to highlight common processes involved in the evolution of both of these genomic features. Any correlations between the two can then be examined in further detail to understand how these processes might have an effect.

Polyploidy

At the most abstract level, whole genome duplication (polyploidy), which can occur due to nondisjunction of chromosomes during meiosis, could create duplicate copies of all regions of the genome, albeit generally followed by rapid gene loss (McLysaght et al., 2002; Schmutz et al., 2010; Jiao et al., 2011; Wolfe, 2015). This would result in a symmetric increase in both orthologous genes, and orthologous introns in the first instance however. The longer term impact on intron characteristics is less clear with the survival of polyploidy depending on population genetic factors and mating compatibility. Mating between individuals with even copied numbers of chromosomes generally produces viable offspring, in contrast to mating between even and odd ploidy individuals (Acquaah, 2007). This is perhaps why we see a great number of tetraploid individuals in nature (Comai, 2005). Species with longer generation times and larger effective population sizes also tend to see a longer term maintenance of polyploidy due to slower fixation of these changes. Polyploidy may not necessarily be fixed *per se* in these individuals, but merely reflect that they are a slower moving targets for effective selection towards a smaller genome size. The ability for any alleles to reach high enough frequencies in a population is drastically reduced however, and it is these changes in gametic and filial frequencies that will impact on genome content and structure over time (Comai, 2005). Sustained ploidy can be advantageous, as it provides the raw material for adaptation on a much larger scale than single gene duplications (Comai, 2005), which is thought to be due to an increased recombination rate due to a greater likelihood of homologous and non-homologous recombination between duplicated regions of the genome (Lynch, 2002; Wendel *et al.,* 2002; Gaut *et al.*, 2007; Tiley and Burleigh, 2015). This is particularly true in plants, which see some of the greatest ploidy numbers, largest genome sizes and most diverse genome content (Otto and Whitton, 2000; Bennett, 2004; Parfrey *et al.*, 2008; Kejnovsky *et al.*, 2009; Jaillon *et al.*, 2009; Levasseur and Pontarotti, 2011; Jia *et al.*, 2013; Nystedt *et al.* 2013), as well as many adaptations to environmental pressures due to neo- and subfunctionalization of duplicated genes (Lynch and Force, 2000; He, 2005; Rastogi and

Liberles, 2005; Chain and Evans, 2006; Shiu *et al.*, 2006; Proulx, 2012; Assis and Bachtrog, 2013). Though polyploidy events will create a great deal of variation for evolution to utilise, it is clear that smaller scale mechanisms will have more influence on intron content over time.

Aneuploidy

In contrast to polyploidy, aneuploidy, or single chromosome duplication, might result in the expansion of gene families on a single chromosome only. In most cases the offspring of aneuploidy are not viable, particularly in mammals, resulting in spontaneous abortion during pregnancy (Carp *et al.* 2001; Sullivan *et al.*, 2004; Bianco *et al.*, 2006). Those individuals that are viable generally suffer from debilitating genetic disorders related to gene dosage effects and therefore these changes are unlikely to spread throughout a population. As aneuploidy is a less common occurrence, the power it has in influencing the evolution of genome content and structure is limited, but it might influence gametic and filial frequencies in similar ways to polyploidy. As the fixation of any changes are dependent on population genetic factors, the likelihood of any significant or lasting effect on genome content is dubious. The immediate impact on intron characteristics is likely to be similar to polyploidy, though longer term, it is intuitive to suggest that smaller scale mechanisms will have greater control over intron evolution.

Replication slippage

Replication slippage occurs when there is a detachment of DNA polymerase from the DNA strand during the replication process (Canceill *et al.*, 1999; Michel, 2000; Viguera *et al.*, 2001). This happens more often in highly repetitive regions of the genome. DNA polymerase encounters a direct repeat and the polymerase complex halts replication and is released. The newly synthesised strand then detaches and binds with a direct repeat upstream, before the DNA polymerase binds back to the template strand to resume replication. Rebinding of the polymerase complex isn't always precise however, resulting in the duplication of di- or trinucleotides. This asymmetry in repeat content causes an inability of pairing between the template and daughter strands. The repair of such changes results in the gain or loss of nucleotides, with trinucleotide expansion occurring most frequently so as not to impact on transcription. Expansion of repeats at this scale one might think is unlikely to impact greatly on the evolution of genomes. If these repeat expansions occur in coding regions they can have profound impacts on the individual, as seen in Huntington's disease in humans (Brown, 2002; Yoon *et al.*, 2003; Chi and Lam, 2005). If the changes occur in non-coding regions however, such as within introns, it is plausible that they could impact on a change in intron size over time, without affecting the fitness of the individual. It is also plausible that the creation of trinucleotide repeats might impact on intron content via tandem genomic duplication (Rogers, 1989; Yenerall and Zhou, 2012) though intron transposition due to partial recombination is also a possibility (Sharp, 1985; Yenerall and Zhou, 2012).

Non-homologous recombination

Non-homologous recombination can be seen as similar to replication slippage in its impact on intron content, however the regions over which it can occur may be much greater in size. Non-homologous recombination occurs when highly similar, yet non-homologous regions of the genome align during meiosis (Roth *et al.*, 1985; Roth and Wilson, 1986; Weterings and van Gent, 2004; Gong *et al.*, 2011). This results in portions of the genome being either duplicated or removed depending on how chromosome pairs segregate in the gametes and survive in the progeny. As with replication slippage these changes occur most often in highly repetitive regions of the genome. This may result from smaller-scale segmental duplications or larger-scale tandem duplications, resulting in either parts of genes or whole genes being duplicated or lost. The impact on intron loss is clear, with non-homologous recombination being a recognised cause of intron loss due to gene conversion or genomic deletion (Derr and Strathern, 1993; Roy and Gilbert, 2006). The impact on intron gain is less clear, though through the duplication of parts or whole copies of genes containing introns one can see how tandem genomic duplication will likely play a role in intron gain (Rogers, 1989; Yenerall and Zhou, 2012). Additionally double-stranded break repair (DSBR) and nonhomologous end joining (NHEJ) have been highlighted as a mechanism of intron gain (Li *et al.*, 2009) as has intron transposition (Sharp, 1985; Yenerall and Zhou, 2012). DSBR also plays a role in gene conversion, which is a sort of intermediate between strand slippage and nonhomologous recombination, where regions of generally between 200 bp to 1,500 bp in length are duplicated and then one allele is converted to the other resulting in a homogenisation at that locus (though see Chen *et al.*, 2007). Depending on the content of the regions in question this may result in gain or loss of introns, or a change in the size of those introns. A bias towards higher GC content in gene conversion has been observed (Galtier *et al.*, 2001; Duret *et al.*, 2006; Duret and Galtier, 2009) along with a bias towards intron gain in more GC rich regions albeit at shorter lengths (Wang and Yu, 2011), so one might expect an overall gain in intron number in duplicate regions, yet a decrease in intron length.

Retrotransposition

How introns are gained or lost due to duplication is apparent via the aforementioned mechanisms, however the change in size of introns with respect to duplication is less clear. Retrotransposition is a mechanism by which we can make more robust predictions. As retrotransposition often results in the removal of introns from the duplicated copy of a gene, we can be confident that it will play more of a role in intron loss (Derr and Strathern, 1993; Yenerall and Zhou, 2012). Retrotransposition exhibits a higher instance of pseudogenization of a particular locus due to loss of promoter sequences, and loss of splice sites due to introns being removed. This can also impact on intron size and density depending on whether the pseudogenes can still be processed. If a transcript is inserted within an existing intron then it will result in an increase in the size of that intron as we see with transposon insertion (Crick, 1979; Yenerall and Zhou, 2012), however the intron number may also increase if an exon is able to be processed in what we might call 'exon transposition' (see intron transposition in Sharp, 1985), though this would require modification of terminal splice sites and is perhaps why processed pseudogenes more often result in intron loss (Zhu and Niu, 2013). Conversely, a pseudogene could be inserted within another gene in such a way that it doesn't impact on the fitness of the organism and the pseudogene is processed as part of the transcription of that gene resulting in an increase in exon size. This would result in a decrease in intron density by increasing the size of the gene, but not affecting the number of introns. This is unlikely however due to the likelihood of this affecting the viability of the product of the gene transcripts translation. The increased mutation rate in pseudogenes due to relaxed selection often causes them to diverge in similarity much more quickly than in coding regions however. This makes it more difficult to test their impact on intron loss. The mechanisms for identifying duplicate loci often use similarity based approaches and so the increased divergence of pseudogenes impacts much more on our ability to annotate them as belonging to gene families.

Relaxed selection due to redundancy

As with pseudogenes, functional duplicate copies of genes can see an increased mutation rate and thus greater divergence than singleton genes, sometimes even greater than the pairwise distance between species. This relaxed selection in duplicate genes allows for mutations that might usually impact on the fitness of an organism in singletons to be masked by their duplicate copy. Both copies of the gene, through genetic drift, can come to share sub-functions of the original gene. This leads to fixation of both copies, as deletion of a copy would reduce the organism's fitness. This is known as the duplication degeneration complementation model or DDC (Force *et al.*, 1999). The majority of mutations aren't beneficial however and any mutation in a single gene would likely be under strong purifying selection. Likewise, selection should be stronger in smaller gene families, where fewer copies of a gene would mean less ability to complement deleterious mutations in other copies. Larger gene families should see an increased ability to complement mutations. Sometimes a gene copy may accumulate so many deleterious mutations that it becomes a pseudogene, though in others we might see multiple beneficial mutations give rise to multigene families whose members have a diversity in functionality through subfunctionalization and neofunctionalization (Lynch and Force, 2000; He, 2005; Rastogi and Liberles, 2005; Chain and Evans, 2006; Shiu et al., 2006; Proulx, 2012; Assis and Bachtrog, 2013). One might predict that larger gene families would create an environment where we see a bias towards increase in intron number and size due to the cumulative effect of all the mechanisms discussed here, though again population genetic factors such as effective population size have a strong role in the fixation of any such mutations. With the outcome of many duplication events resulting in pseudogenization, the ability to identify a correlation between gene family size and intron characteristics may be limited in its power however.

Mechanisms of location based intron evolution

Understanding how duplication impacts on the evolution of regions of the genome is very important, not least in determining how organisms adapt to changes in their environment or in describing the evolution of novel gene function. The amount of the genome that is contained within gene families is relatively small however, with the portion of the genome attributed to protein coding widely acknowledge to be approximately 1.5% of the total genome size in humans for example (Lander *et al.*, 2001; Venter *et al.*, 2001; International Human Genome Sequencing Consortium, 2004; Rands *et al.*, 2014). It is important therefore to gain a more thorough understanding of the forces shaping the genome at a larger scale. There have been several studies (Galtier *et al.*, 2001; Montoya-Burgos *et al.*, 2003; Meunier and Duret, 2004; Duret and Galtier, 2009; Marsolier-Kergoat and Yeramian, 2009; Weber *et al.*, 2014) that focus on chromosome location effects on the evolution of lower level components of genome architecture such as gene density, in particular highlighting the impact of GC-rich isochores, recombination, and gene conversion on these regions. There are none that take a comprehensive comparative genomics approach across multiple species, however. Likewise, there have been some studies (Fullerton *et al.*, 2001; Prachumwat *et al.* 2004; Haddrill *et al.*, 2005; Zhe *et al.*, 2009; Li *et al.*, 2009; Maeso *et al.*, 2012; Zhang and Edwards, 2012) that look at the evolution of introns within this context, but their power is limited as they fail to synthesise all relevant information (i.e. the subtle influences of GC-bias, recombination, and gene conversion) from other studies and thus don't account for the bigger picture. In understanding how genome structure and content evolves over time one needs to consider both the lower level forces of evolution such as point mutations and the higher level constraints imposed at the chromosome level.

GC-rich isochores

A number of studies have focused on the existence of GC rich regions of the genome, particularly in vertebrates, known as CpG islands over shorter distances of between 300 to 3,000 bp (Gardiner-Garden and Frommer, 1987; Saxonov *et al.*, 2006; Deaton and Bird, 2011) and isochores over regions of greater than 300,000 bp (Lander *et al.*, 2001; Oliver *et al.*, 2001; Gao and Zhang, 2006). Gene density has been shown to be much higher in these GC rich regions of the genome (Galtier *et al.*, 2001; Montoya-Burgos *et al.*, 2003; Bernardi, 2012). Conversely introns have been shown to be shorter (Galtier *et al.*, 2001) and under greater selective pressure in GC rich regions (Zhu *et al.*, 2009; Wang and Yu, 2011), though this depends largely on their position in the gene (Haddrill *et al.*, 2005; Gazave *et al.*, 2007) with first introns seeing a greater length and increased GC content overall (Kalari *et al.*, 2006; Wang and Yu, 2011). In order to understand how GC isochores in particular influence the evolution on intron characteristics we must first understand how these regions of the genome arise.

GC-rich isochores are a common occurrence in mammals (Bernardi, 2000; Ellegren *et al.*, 2003; Romiguier *et al.*, 2010; Lartillot, 2012; Nabiyounia *et al.*, 2013), though there is evidence to suggest changing patterns and a great deal of variation in content (Duret *et al.*, 2002; Belle *et al.*, 2004; Fujita *et al.*, 2011). Many of these studies have largely focused on the impact of GC bias on the evolution of components of the genome without thoroughly exploring how these regions arise however. Several studies in bacteria document a replication related organization of their genomes (Lobry, 1996; Eyre-Walker and Hurst, 2001; Rocha, 2004; Flynn *et al.*, 2010), which can be linked to the increased mutation rate of single-stranded DNA during replication and a bias in the DNA repair mechanisms (Wu *et al.*, 2005). Indeed, using GC skew plots is a common bioinformatics method of identifying origins of replication (Grigoriev, 1998; Eng *et al.*, 2009; Pevzner and Shamir, 2011). Likewise a relationship between replication and isochores in eukaryotes has been highlighted with early replicating regions of the genome being GC rich and short and late replicating regions being GC poor and long (Oliver *et al.*, 2001; Pac̀es *et al.*, 2004; Schmegner *et al.*, 2007; Costantini and Bernardi, 2008; Watanabe *et al.*, 2009; Costantini *et al.*, 2013). Although little is known about the precise mechanisms behind this there is evidence highlighting a correlation between replication timing and chromosome structure (Hiratani *et al.*, 2009; Hiratani and Gilbert, 2009; Schwaiger *et al.*, 2009; Pope *et al.*, 2010; Ryba *et al.*, 2010; Farkash-Amar and Simon, 2010), which indicates a possible epigenetic effect. Chromatin modifications have been examined in reference to replication, showing a strong correlation between epigenetic marks and replication timing (Ryba *et al.*, 2010; Ryba *et al.*, 2011; Appasani, 2012).

Recombination

A positive correlation between replication timing and epigenetic modifications, and a number of genomic features such as GC content, mutation rate, gene density and transcriptional activity have been emphasised above (Woodfine *et al.*, 2004; Stamatoyannopoulos *et al.*, 2009; Appasani, 2012). Epigenetic modifications can’t be directly associated with changes in intron characteristics as examined in this study however. There is a positive correlation between recombination rate and GC content however (Fullerton *et al.*, 2001; McVean *et al.*, 2004; Meunier and Duret, 2004), which is more likely to have a direct impact on the underlying sequence evolution. Biased mutation rates have been shown not to be solely responsible for changes in GC content however (Lercher and Hurst, 2002), with evidence suggesting that mutation biases depend on the existing GC content of the region (Fryxell and Moon, 2005), pointing towards effects due to recombination or biased gene conversion. The impact of recombination rate on the number of genes has been explored and although mutation rates are higher in regions of higher recombination, there is no significant correlation with gene density (Hey and Kliman, 2002). Non-homologous recombination is a recognised cause of intron loss due to gene conversion or genomic deletion (Derr and Strathern, 1993; Roy and Gilbert, 2006) and conversely intron gain (Li *et al.*, 2009) so it seems likely that non-homologous recombination and gene conversion play are larger role in the evolution of introns at least.

Gene conversion

Mismatch of bases following DNA strand transfer can initiate DNA repair mechanisms. DNA repair during homologous recombination can result in the conversion of one allele to another resulting in homozygosity at that the location involved. The replacement of a gene with its alternate allele, may impact on the number of introns present in the gene if the number of introns are polymorphic. It isn't clear whether this would result in a bias towards increase or decrease in intron number within the gene, though GC content has been shown to impact this; with GC-rich regions containing many genes with short introns, and GC-poor regions having virtually no genes at all (Galtier *et al.*, 2001). As GC-bias can impact on the likelihood of gene conversion occurring during homologous recombination (Galtier *et al.*, 2001; Duret *et al.*, 2006; Duret and Galtier, 2009) and also on the number of genes and size of introns, one might conclude that areas of higher recombination, as we find in isochores, should see an increased gene density and intron density, especially given the role of double stranded break repair in intron gain (Li *et al.*, 2009; Wang and Yu, 2011). This doesn't always appear to be the case however (Hey and Kliman, 2002; Derr and Strathern, 1993; Roy and Gilbert, 2006) though it is wise to consider the distances over which gene conversion can occur. Gene conversion typically occurs over shorter distances of between 200 bp to 1,500 bp and requires a high level of similarity between copies (Gaultier, 2003). Orthologous gene conversion between homologous regions (thus with high similarity) of the chromosomes during meiosis would explain the correlation with recombination rate, but given the short distances may only result in exchange of material between genes therefore not affecting gene density, though these would be under strong purifying selection (Kjeldbjerg *et al.*, 2008; Petronella and Drouin, 2013; Zid and Drouin, 2013). In contrast gene conversion can also occur ectopically, known as non-allelic gene conversion, between paralogous copies of genes. Gene conversion here can result in the creation of novel gene content when material is rearranged within genes (Willett, 2013) though it can also result in increased homogenisation of paralogous gene copies giving rise to concerted evolution (Teshima and Innan, 2004). Gene conversion may not play a dominant role in the evolution of paralogues however as there seems to be no GC bias between paralogous copies of duplicate genes (Assis and Kondrashov, 2012).

It is clear that there are many influences on the evolution of genes and introns at a number of different levels. At the lowest levels point mutations and indels create minor variations that can be acted upon by selection or allowed to drift within a population. Intermediate mechanisms such as duplication and gene conversion play a role in rearranging content at the gene level. Numerous mechanisms of repair with biases towards particular outcomes compound matters. In the context of this study it seems that larger scale mechanisms such as replication timing and epigenetic modifications may play an overarching role however. Their impact on gene conversion results in an increase in gene density due to greater homogeneity within regions of the genome and therefore increased recombination rates. The effect of strand slippage, gene conversion and larger-scale recombination on gene density and intron characteristics are also widely discussed. The bias towards increased GC content in these mechanisms correlates with greater gene and intron density, and reduced intron length. Increase in intron length is likely predominantly due to other mechanisms such as transposon insertion. The constraints imposed by selection, developmental process or population specific factors make for a landscape that is ripe with contradiction and noise, making it difficult to pinpoint the precise causes behind genome evolution. High quality sequence data and annotations utilised within multi-species comparative studies will allow an assessment of the diversity of influences at a broader scale however.

Chapter goals

In this chapter I will attempt to tease apart the mechanisms described here by answering the following questions; 1) Is intron evolution independent in different gene copies? 2) Does size of a gene family influence intron evolution? 3) Is there heterogeneity in intron evolution across chromosomes, and between sex chromosomes and autosomes? The first two questions will look at the impact of duplications, and in particular attempt to understand whether change in gene family size impacts on the evolution of introns. The first question will address whether introns evolve independently within gene families, where as the second question will address whether the overall size of the gene family has an impact on the gain or loss of introns through time. The third questions will examine non-family effects on intron evolution. I will attempt to understand whether physical location in the genome influences intron properties, and whether sex chromosomes and autosomes differ in their impact on intron characteristics. These questions will be addressed in a large-scale comparative genomics study allowing for the consideration of multiple life history traits and population genetic influences. By approaching these questions at such a large-scale and by utilising exploratory data analysis techniques, descriptive and inferential statistics, and data visualisation it will provide a great deal of power in understanding how genome content and structure changes over time across vastly divergent distances, therefore highlighting common mechanisms in the evolution of genomes in all species.

Materials and Methods

Species used in this study

All 61 species available in the Ensembl release 70 (January 2013) databases were used. These species were *Ailuropoda melanoleuca* (giant panda), *Anolis carolinensis* (Carolina anole lizard), *Bos taurus* (European cow), *Caenorhabditis elegans* (roundworm), *Callithrix jacchus* (common marmoset), *Canis familiaris* (domestic dog), *Cavia porcellus* (guinea pig), *Choloepus hoffmanni* (Hoffmann's two-toed sloth), *Ciona intestinalis* (sea squirt), *Ciona savignyi* (Pacific transparent sea squirt), *Danio rerio* (zebrafish), *Dasypus novemcinctus* (nine-banded armadillo), *Dipodomys ordii* (Ord's kangaroo rat), *Drosophila melanogaster* (common fruit fly), *Echinops telfairi* (lesser hedgehog tenrec), *Equus caballus* (horse), *Erinaceus europaeus* (European hedgehog), *Felis catus* (domestic cat), *Gadus morhua* (Atlantic cod), *Gallus* (chicken), *Gasterosteus aculeatus* (three-spined stickleback), *Gorilla* (western lowland gorilla), *Homo sapiens* (human), *Ictidomys tridecemlineatus* (thirteen-lined ground squirrel), *Latimeria chalumnae* (West Indian Ocean coelacanth), *Loxodonta africana* (African bush elephant), *Macaca mulatta* (rhesus macaque), *Macropus eugenii* (tammar wallaby), *Meleagris gallopavo* (Wild Turkey), *Microcebus murinus* (gray mouse lemur), *Monodelphis domestica* (gray short-tailed opossum), *Mus musculus* (house mouse), *Mustela putorius furo* (ferret), *Myotis lucifugus* (little brown bat), *Nomascus leucogenys* (northern white-cheeked gibbon), *Ochotona princeps* (American pika), *Oreochromis niloticus* (Nile tilapia), *Ornithorhynchus anatinus* (platypus), *Oryctolagus cuniculus* (European rabbit), *Oryzias latipes* (medaka), *Otolemur garnettii* (Northern greater galago), *Pan troglodytes* (common chimpanzee), Pelodiscus sinensis (Chinese softshell turtle), *Petromyzon marinus* (sea lamprey), *Pongo abelii* (Sumatran orangutan), *Procavia capensis* (rock hyrax), *Pteropus vampyrus* (large flying fox), *Rattus norvegicus* (Norwegian brown rat), *Saccharomyces cerevisiae* (yeast), *Sarcophilus harrisii* (Tasmanian devil), *Sorex araneus* (common shrew), *Sus scrofa* (pig), *Taeniopygia guttata* (Zebra Finch), *Takifugu rubripes* (pufferfish), *Tarsius syrichta* (Philippine tarsier), *Tetraodon nigroviridis* (green spotted puffer), *Tupaia belangeri* (northern treeshrew), *Tursiops truncatus* (Atlantic bottlenose dolphin), *Vicugna pacos* (alpaca), *Xenopus tropicalis* (Western clawed frog), and *Xiphophorus maculatus* (southern platyfish).

Gene family data

The GCAT API and scripts developed as part of Chapter Two and Chapter Four (e.g. `get_protfam_dist.pl`) were used to retrieve data on all of the gene families for all 61 species available in Ensembl's release 70 databases (Flicek *et al.*,2012). The "all families" method was used to retrieve these data (see Materials and Methods in Chapter Four). GCAT helper scripts were developed using R (`gene_family_distribution.R`) to provide descriptive statistics and visualizations of these data.

Intron data

The GCAT API and scripts developed as part of Chapter Three (e.g. `get_intron_counts.pl` and `get_introns.pl`) were used to retrieve data on the intron counts for each of the genes in all of the 61 species available in Ensembl's January 2013 release 70 databases (Flicek *et al.*, 2012). All intron numbers were retrieved for all genes via the canonical transcript for those genes, but only those contained within genes that were members of gene families were considered for further analyses. GCAT helper scripts were developed using R (`gene_family_introns.R`) to provide descriptive statistics and visualizations of these data. Intron sizes were retrieved for all species (using the GCAT plugin script `get_intron_lengths.pl`) to use when testing for relationships between intron characteristics and spatial location of those introns on their respective chromosome.

Gene family and intron data correlations

GCAT helper scripts were developed using R (`gene_family_introns.R`) to undertake munging (transformation, formatting and sub-setting) of the data as necessary. Spearman tests were performed between intron number, size and density, and gene family size to test for correlations. Intron density was calculated as the number of introns per bond i.e. intron count in the gene divided by the length of the transcript minus one (see Chapter Three). Only genes contained within gene families were considered. The Kruskal-Wallis test (the nonparametric equivalent of a one-way ANOVA) was used to test whether intron density, and gene family size originated from the same distribution, and thus whether there was a significant difference between the data.

Chromosome data retrieval

A novel GCAT plugin script (`get_chromosomes.pl`) was developed to retrieve data on the chromosomes for *Homo sapiens* only. Only coordinate system names annotated as `chromosome` were considered for further analyses. This disregards any `toplevel` slices such as contigs or supercontigs that stem from incompletely assembled sequence data. Where the name and length of a slice annotated as a `chromosome` was inconsistent with the expected chromosome names, then those data were discarded from downstream analyses, as they would likely represent incompletely assembled raw sequence data. GCAT helper scripts were developed using R (`gene_family_introns.R`) to mung these data, to perform descriptive and inferential statistics, and to produce visualizations of these data.

Intron characteristics by location on the chromosomes

The GCAT API and scripts developed as part of Chapter Three (e.g. `get_intron_counts.pl` and `get_introns.pl`) were used to retrieve data on the intron counts for each of the genes in the *Homo sapiens* genome. A GCAT helper script was developed using R (`gene_family_introns.R`) to mung the data as necessary. Only introns contained within protein coding genes were considered to examine whether genomic location had a greater effect on the evolution of introns than gene family membership. Intron density was calculated on a sliding window basis with the density given per 1Mb length of the chromosome. Sex chromosomes and autosomes were also compared to highlight the effects of recombination on intron evolution.

Results

Gene family data

A total of 1,102,993 records were retrieved corresponding to the total number of genes within gene families for all 61 species in the Ensembl release 70 databases. These genes were contained within a total of 130,090 gene families. Gene family count (GF) and gene count (G) vary from a minimum of 5,683 and 6,692 respectively both in *Saccharomyces cerevisiae* to a maximum of 14,637 and 26,157 respectively in *Caenorhabditis elegans* and *Danio rerio*. The gene/gene family ratio (G/GF) ranges from 0.489 in *Danio rerio* to 0.854 in *Ciona intestinalis*. Gene family size ranges from a size of 0, meaning complete loss in that species, to a maximum (Max GF) of 656 members in *Equus caballus* (see Table 5.1). The mean, median and mode sizes for each group were 11,649.48, 11,730, and 12,611 (GF); 18,081.85, 18,788, and 19,343 (G); 0.656, 0.645, and 0.705 (G/GF); 167.75, 121, and 121 (Max GF).

Gene family sizes were visualized and followed a power law distribution in all species as previously reported (see Figure 5.1 and Chapter Four).

Intron data

Intron counts and densities

A total of 1,426,601 records were retrieved corresponding to the total number of genes containing introns for all 61 species available in the Ensembl release 70 databases. This corresponded with a total of 10,367,439 individual introns. These intron data were trimmed to match the gene IDs identified as belonging to gene families, resulting in 1,102,993 records containing a total of 10,139,168 introns across all 61 species. The number of genes containing at least 1 intron was 1,007,633. The genes had a minimum length of 8 bp (in *Homo sapiens*) and a maximum length of 4,434,882 bp (in *Mus musculus*). Intron counts per gene ranged from 0 to 378 (in *Choloepus hoffmanni*) with a mean intron count of 9.192 (see Supplementary Table 5.1 and Supplementary Figure 5.1). The minimum transcript length was 8 bp (in *Homo sapiens* – likely not real as not a multiple of 3) and a maximum of 106,731 bp (in *Equus caballus*). The intron density (introns per bond – see Lunt, 2013) per gene ranged from 0 to a maximum of 0.132 (in *Choloepus hoffmanni*) with a mean intron density of 0.005 (see Table 5.2 and Figure 5.2).

Table 5.1 - Breakdown of gene family information for all 61 species available in release 70 of the Ensembl databases.

Species Name	Gene Family (GF) Count	Gene (G) Count	G/GF Ratio	Max GF Size
ailuropoda_melanoleuca	12,611	19,343	0.652	113
anolis_carolinensis	11,006	17,792	0.619	424
bos_taurus	12,344	12,344	0.617	154
caenorhabditis_elegans	14,637	20,517	0.713	55
callithrix_jacchus	13,052	20,993	0.622	224
canis_familiaris	12,854	19,856	0.647	150
cavia_porcellus	12,034	18,673	0.644	73
choloepus_hoffmanni	9,208	12,393	0.743	121
ciona_intestinalis	14,222	16,658	0.854	38
ciona_savignyi	9,655	11,604	0.832	43
danio_rerio	12,786	26,157	0.489	509
dasypus_novemcinctus	10,573	14,803	0.714	118
dipodomys_ordii	11,144	15,798	0.705	57
drosophila_melanogaster	11,295	13,937	0.810	34
echinops_telfairi	11,569	16,562	0.699	117
equus_caballus	12,161	20,449	0.595	656
erinaceus_europaeus	10,478	14,588	0.718	54
felis_catus	12,497	19,493	0.641	106
gadus_morhua	11,811	20,095	0.588	129
gallus_gallus	11,878	16,736	0.710	204
gasterosteus_aculeatus	11,730	20,787	0.564	197
gorilla_gorilla	13,825	20,962	0.660	201
homo_sapiens	13,881	23,260	0.597	271
ictidomys_tridecemlineatus	11,899	18,826	0.632	70
latimeria_chalumnae	11,563	19,569	0.591	128
loxodonta_africana	12,052	20,033	0.602	216
macaca_mulatta	13,842	21,905	0.632	231
macropus_eugenii	10,528	15,290	0.689	54
meleagris_gallopavo	9,785	14,125	0.693	20
microcebus_murinus	11,553	16,319	0.708	116
monodelphis_domestica	12,429	21,327	0.583	492
mus_musculus	13,548	22,783	0.595	283
mustela_putorius_furo	13,324	19,910	0.669	130
myotis_lucifugus	11,735	19,728	0.595	83
nomascus_leucogenys	12,769	18,575	0.687	211

Table 5.1 continued.

Species Name	Gene Family (GF) Count	Gene (G) Count	G/GF Ratio	Max GF Size
ochotona_princeps	11,282	15,993	0.705	70
oreochromis_niloticus	11,081	21,437	0.517	209
ornithorhynchus_anatinus	13,722	21,698	0.632	469
oryctolagus_cuniculus	11,711	19,018	0.616	92
oryzias_latipes	11,417	19,686	0.580	61
otolemur_garnettii	12,058	19,506	0.618	206
pan_troglodytes	12,831	18,759	0.684	212
pelodiscus_sinensis	11,224	18,188	0.617	541
petromyzon_marinus	7,517	10,402	0.723	42
pongo_abelii	13,907	20,424	0.681	212
procavia_capensis	11,392	16,044	0.710	83
pteropus_vampyrus	11,972	16,990	0.705	98
rattus_norvegicus	13,076	22,941	0.570	121
saccharomyces_cerevisiae	5,683	6,692	0.849	82
sarcophilus_harrisii	12,110	18,788	0.645	275
sorex_araneus	9,516	13,187	0.722	80
sus_scrofa	12,334	21,630	0.570	131
taeniopygia_guttata	10,823	17,488	0.619	228
takifugu_rubripes	10,037	18,523	0.542	75
tarsius_syrichta	10,046	13,615	0.738	146
tetraodon_nigroviridis	10,744	19,602	0.548	29
tupaia_belangeri	11,050	15,458	0.715	87
tursiops_truncatus	11,720	16,537	0.709	119
vicugna_pacos	8,978	11,752	0.764	85
xenopus_tropicalis	10,514	18,429	0.571	352
xiphophorus_maculatus	11,595	20,366	0.569	46

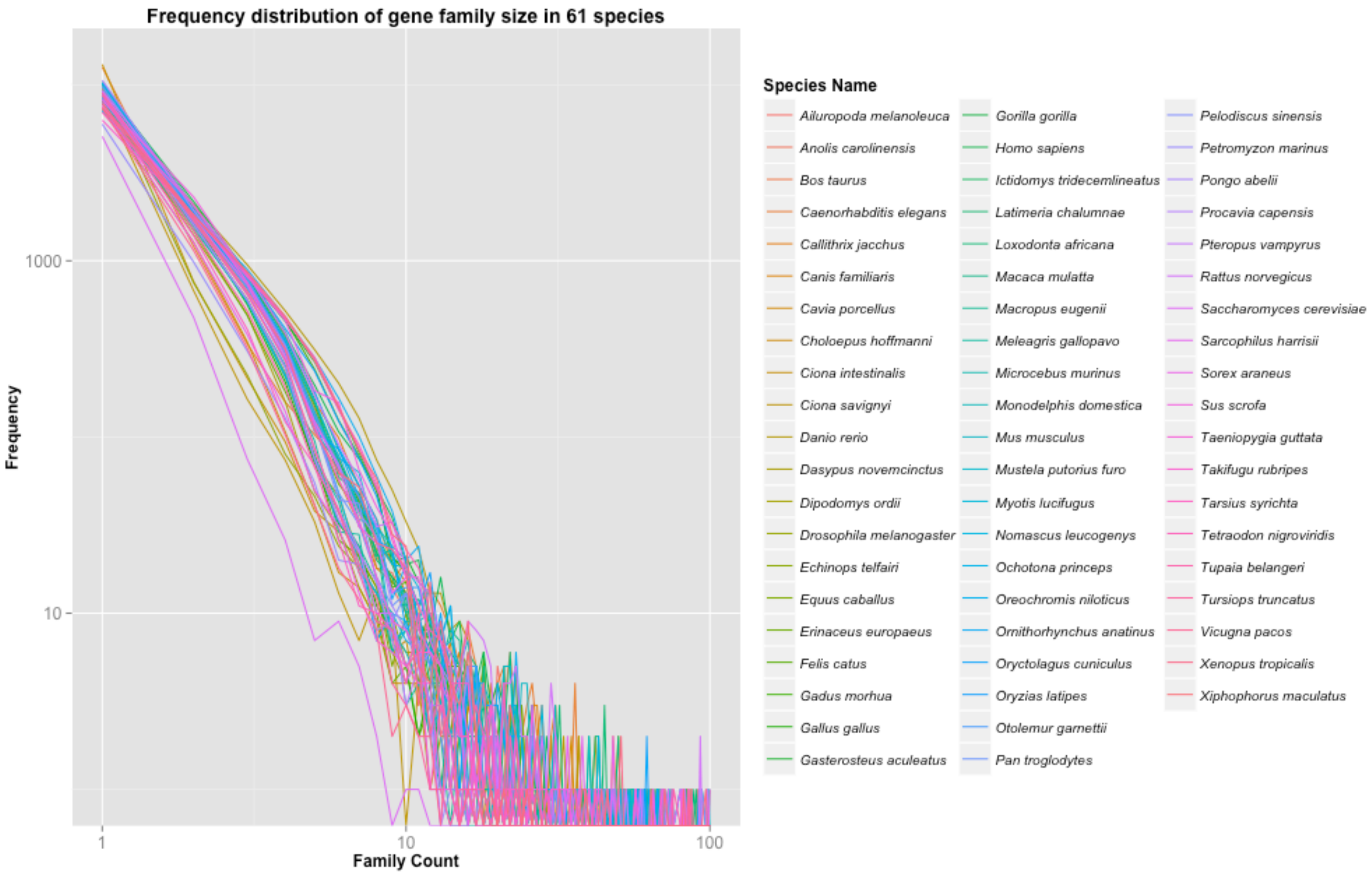


Figure 5.1 - Frequency distribution of gene family size in all 61 species available as of the January 2013 release (70) of Ensembl. A cut-off of 100 for gene family size is used as this represents the majority of the data. The maximum gene family size in these species is 656.

Table 5.2 - Breakdown of intron density (introns/bond) information for all 61 species available in release 70 of the Ensembl databases.

Species Name	Min	Max	Mean	Median	Mode
ailuropoda_melanoleuca	0.00018	0.02483	0.00593	0.00602	0.00763
anolis_carolinensis	0.00019	0.03106	0.00591	0.00608	0.00477
bos_taurus	0.00015	0.03636	0.00464	0.00451	0.00578
caenorhabditis_elegans	0.00023	0.03390	0.00409	0.00390	0.00403
callithrix_jacchus	0.00013	0.02857	0.00463	0.00427	0.00472
canis_familiaris	0.00014	0.03279	0.00448	0.00418	0.00704
cavia_porcellus	0.00021	0.03191	0.00606	0.00617	0.00787
choloepus_hoffmanni	0.00027	0.13208	0.00812	0.00748	0.00676
ciona_intestinalis	0.00020	0.02062	0.00462	0.00467	0.00448
ciona_savignyi	0.00010	0.01859	0.00500	0.00518	0.00610
danio_rerio	0.00005	0.03488	0.00424	0.00395	0.00768
dasypus_novemcinctus	0.00032	0.12500	0.00851	0.00766	0.01079
dipodomys_ordii	0.00022	0.13043	0.00813	0.00734	0.00763
drosophila_melanogaster	0.00009	0.01505	0.00192	0.00167	0.00192
echinops_telfairi	0.00026	0.10526	0.01014	0.00845	0.01000
equus_caballus	0.00018	0.03846	0.00511	0.00515	0.00135
erinaceus_europaeus	0.00027	0.11168	0.00894	0.00780	0.00730
felis_catus	0.00012	0.02296	0.00481	0.00462	0.00493
gadus_morhua	0.00032	0.06132	0.00819	0.00792	0.00917
gallus_gallus	0.00018	0.02857	0.00501	0.00496	0.00541
gasterosteus_aculeatus	0.00017	0.02326	0.00548	0.00559	0.00295
gorilla_gorilla	0.00006	0.09375	0.00480	0.00405	0.00270
homo_sapiens	0.00006	0.05000	0.00334	0.00279	0.00389
ictidomys_tridecemlineatus	0.00012	0.02410	0.00535	0.00543	0.00735
latimeria_chalumnae	0.00013	0.02273	0.00465	0.00405	0.00515
loxodonta_africana	0.00021	0.02363	0.00594	0.00611	0.00107
macaca_mulatta	0.00015	0.04348	0.00479	0.00448	0.00435
macropus_eugenii	0.00031	0.08911	0.00815	0.00768	0.00787
meleagris_gallopavo	0.00025	0.02247	0.00591	0.00591	0.00637
microcebus_murinus	0.00027	0.11429	0.00864	0.00755	0.00847
monodelphis_domestica	0.00013	0.02941	0.00375	0.00321	0.00106
mus_musculus	0.00002	0.04348	0.00324	0.00282	0.00286
mustela_putorius_furo	0.00005	0.02664	0.00399	0.00350	0.00214
myotis_lucifugus	0.00014	0.03448	0.00571	0.00579	0.00283
nomascus_leucogenys	0.00010	0.02146	0.00381	0.00338	0.00478

Table 5.2 continued.

Species Name	**Min**	**Max**	**Mean**	**Median**	**Mode**
ochotona_princeps	0.00033	0.13043	0.00897	0.00772	0.01317
oreochromis_niloticus	0.00012	0.02740	0.00446	0.00411	0.00633
ornithorhynchus_anatinus	0.00013	0.03349	0.00477	0.00433	0.00676
oryctolagus_cuniculus	0.00025	0.03165	0.00569	0.00578	0.00581
oryzias_latipes	0.00030	0.01930	0.00575	0.00588	0.00685
otolemur_garnettii	0.00016	0.02400	0.00552	0.00557	0.00752
pan_troglodytes	0.00011	0.05714	0.00421	0.00377	0.00541
pelodiscus_sinensis	0.00014	0.02206	0.00413	0.00367	0.00062
petromyzon_marinus	0.00021	0.03185	0.00632	0.00638	0.00730
pongo_abelii	0.00010	0.07241	0.00473	0.00388	0.00293
procavia_capensis	0.00036	0.10000	0.00834	0.00755	0.00763
pteropus_vampyrus	0.00021	0.11538	0.00809	0.00727	0.00662
rattus_norvegicus	0.00010	0.04255	0.00435	0.00409	0.00296
saccharomyces_cerevisiae	0.00018	0.01408	0.00354	0.00187	0.00019
sarcophilus_harrisii	0.00014	0.03141	0.00484	0.00461	0.00662
sorex_araneus	0.00021	0.07531	0.00872	0.00769	0.00503
sus_scrofa	0.00013	0.04255	0.00439	0.00403	0.00485
taeniopygia_guttata	0.00029	0.02243	0.00573	0.00576	0.00588
takifugu_rubripes	0.00015	0.02085	0.00596	0.00620	0.00649
tarsius_syrichta	0.00029	0.07368	0.00804	0.00737	0.00862
tetraodon_nigroviridis	0.00032	0.02353	0.00605	0.00614	0.00769
tupaia_belangeri	0.00026	0.08750	0.00925	0.00799	0.00459
tursiops_truncatus	0.00027	0.10526	0.00801	0.00723	0.00093
vicugna_pacos	0.00027	0.07547	0.00806	0.00748	0.00259
xenopus_tropicalis	0.00013	0.02101	0.00522	0.00507	0.00380
xiphophorus_maculatus	0.00011	0.02395	0.00444	0.00411	0.00637

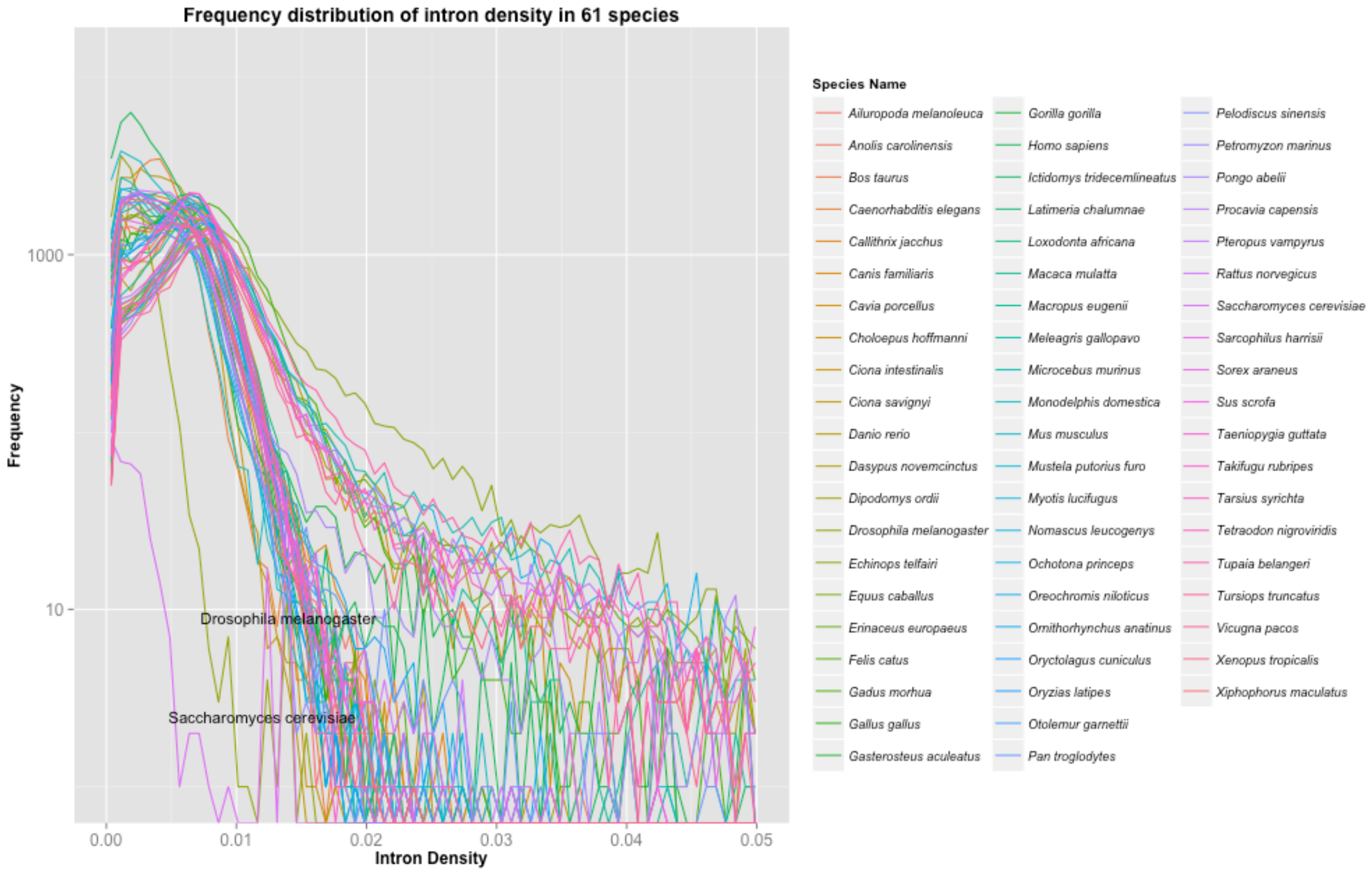


Figure 5.2- Frequency distribution of intron density in all 61 species in Ensembl release 70. Intron density is trimmed to 0.05 on the x-axis, which represents the majority of the data. This figure represents the right skew in the data, with the mean, median and mode all being approximately <= 0.005. The maximum intron density in these species is 0.1320755. See Appendix 5.2 for inset of 0 to 0.001 intron density.

Intron sizes

Intron sizes were retrieved for the 1,102,993 gene IDs identified as belonging to gene families, for all species available in the release 70 version of Ensembl's databases. A total of 1,007,633 gene IDs were recovered, which is less than the number of input gene IDs, indicating that 95,360 genes had no introns annotated. A total of 10,139,168 records were retrieved, which is the total number of introns annotated across all species. The total number of introns ranged from 358 in *Saccharomyces cerevisiae* to 235,593 in *Gadus morhua*. The total length of introns for each species was calculated, ranging from 111,916 bp (0.92% of the 12,157,105 bp genome) in *Saccharomyces cerevisiae* to 1,142,973,003 bp (31.70% of the 3,605,631,728 bp genome) in *Monodelphis domestica*. The minimum intron size annotated was 1 bp (which is certainly not a real intron) with a maximum size of 4,384,418 bp in *Mus musculus*. Whilst there are some extreme values at the minimum and maximum of annotated introns – which aren't explicitly annotate by Ensembl and instead are classified by the space between exons, regardless of whether they are correctly called – the majority of the data are the same across all species, covering a wide range of plausible intron sizes). The mean of all intron sizes ranged from 307 bp in *Caenorhabditis elegans* to 6,672 bp in *Monodelphis domestica*, though the mode of all species was 86 bp, which follows the pattern of intron size distribution highlighted in Chapter Two and Chapter Three (Moss *et al.*, 2011) (see Table 5.3 and Figure 5.3).

Table 5.3 - Breakdown of intron size information for all 61 species available in release 70 of the Ensembl databases.

Species Name	Number	Length	Max	Mean	Median	Mode
ailuropoda_melanoleuca	170,381	646,117,676	199,282	3,792.193	1086	88
anolis_carolinensis	151,837	390,119,547	223,783	2,569.331	1213	84
bos_taurus	178,229	734,909,925	1,184,028	4,123.403	1163	87
caenorhabditis_elegans	106,673	32,694,553	100,913	306.493	64	47
callithrix_jacchus	188,974	927,097,104	1,014,144	4,905.951	1242	84
canis_familiaris	171,762	742,116,143	265,212	4,320.607	1213	89
cavia_porcellus	167,913	585,117,916	294,846	3,484.649	991	83
choloepus_hoffmanni	141,059	318,464,028	120,973	2,257.665	828	100
ciona_intestinalis	98,146	46,968,103	43,597	478.553	323	57
ciona_savignyi	75,145	50,111,385	19,782	666.863	451	56
danio_rerio	222,283	624,323,314	378,145	2,808.687	991	86
dasypus_novemcinctus	173,790	824,800,226	412,727	4,745.959	697	100
dipodomys_ordii	187,156	512,470,546	1,433,833	2,738.200	727	100
drosophila_melanogaster	46,020	54,471,673	141,627	1,183.652	72	58
echinops_telfairi	209,317	791,550,321	340,784	3,781.586	354	100
equus_caballus	173,591	706,712,927	247,742	4,071.138	1129	88
erinaceus_europaeus	177,889	616,223,038	431,914	3,464.087	652	100
felis_catus	173,049	661,763,228	199,529	3,824.138	1144	88
gadus_morhua	235,593	277,037,249	759,882	1,175.915	282	100
gallus_gallus	146,217	376,944,883	488,850	2,577.983	801	86
gasterosteus_aculeatus	199,624	151,619,269	175,269	759.524	219	85
gorilla_gorilla	180,441	881,519,672	749,886	4,885.362	1304	85
homo_sapiens	202,870	1,132,066,118	4,250,947	5,580.254	1437	85
ictidomys_tridecemlineatus	163,915	635,665,385	321,457	3,878.018	1135	86
latimeria_chalumnae	178,098	733,614,509	199,449	4,119.162	1759	12
loxodonta_africana	167,993	705,725,509	334,004	4,200.922	1148	87
macaca_mulatta	174,625	896,959,214	988,812	5,136.488	1306	88
macropus_eugenii	193,533	431,626,711	135,339	2,230.249	514	100
meleagris_gallopavo	139,721	292,548,549	284,326	2,093.805	713	85
microcebus_murinus	188,440	700,848,350	697,784	3,719.212	717	100
monodelphis_domestica	171,320	1,142,973,003	312,330	6,671.568	1679	88
mus_musculus	184,137	873,871,082	4,384,418	4,745.766	1320	88
mustela_putorius_furo	174,466	802,941,219	279,537	4,602.279	1249	86
myotis_lucifugus	166,148	473,340,990	226,880	2,848.912	1041	85
nomascus_leucogenys	171,353	916,946,861	416,468	5,351.216	1461	85

Table 5.3 continued.

Species Name	Number	Length	Max	Mean	Median	Mode
ochotona_princeps	197,187	718,307,265	919,067	3,642.772	540	100
oreochromis_niloticus	214,987	307,960,522	195,966	1,432.461	328	85
ornithorhynchus_anatinus	146,825	440,063,833	390,999	2,997.200	1056	91
oryctolagus_cuniculus	161,876	676,971,002	440,526	4,182.034	1189	86
oryzias_latipes	185,494	219,591,667	295,125	1,183.821	252	77
otolemur_garnettii	170,593	684,625,122	540,755	4,013.208	1269	84
pan_troglodytes	170,795	943,847,045	1,086,464	5,526.198	1494	85
pelodiscus_sinensis	151,921	703,527,136	220,299	4,630.875	1587	90
petromyzon_marinus	81,132	119,353,863	172,894	1,471.107	764	102
pongo_abelii	179,779	947,900,833	3,022,163	5,272.589	1308	85
procavia_capensis	196,557	568,713,224	388,070	2,893.376	703	100
pteropus_vampyrus	198,495	587,449,131	464,755	2,959.516	734	100
rattus_norvegicus	181,135	756,185,018	898,102	4,174.704	1173	86
saccharomyces_cerevisiae	358	111,916	2,483	312.615	123	99
sarcophilus_harrisii	160,402	643,446,829	220,904	4,011.464	1279	88
sorex_araneus	153,980	553,862,343	458,340	3,596.976	625	100
sus_scrofa	167,069	631,980,866	512,204	3,782.754	1158	102
taeniopygia_guttata	141,480	364,683,662	233,394	2,577.634	775	86
takifugu_rubripes	187,962	108,524,412	93,537	577.374	143	78
tarsius_syrichta	154,157	441,679,163	238,254	2,865.126	997	100
tetraodon_nigroviridis	187,875	90,447,562	631,227	481.424	118	76
tupaia_belangeri	189,131	834,229,024	809,294	4,410.853	632	100
tursiops_truncatus	194,085	719,988,521	547,205	3,709.656	833	100
vicugna_pacos	135,032	626,204,310	778,219	4,637.451	1001	100
xenopus_tropicalis	176,758	375,692,909	218,827	2,125.465	830	84
xiphophorus_maculatus	202,391	263,410,565	196,977	1,301.493	379	83

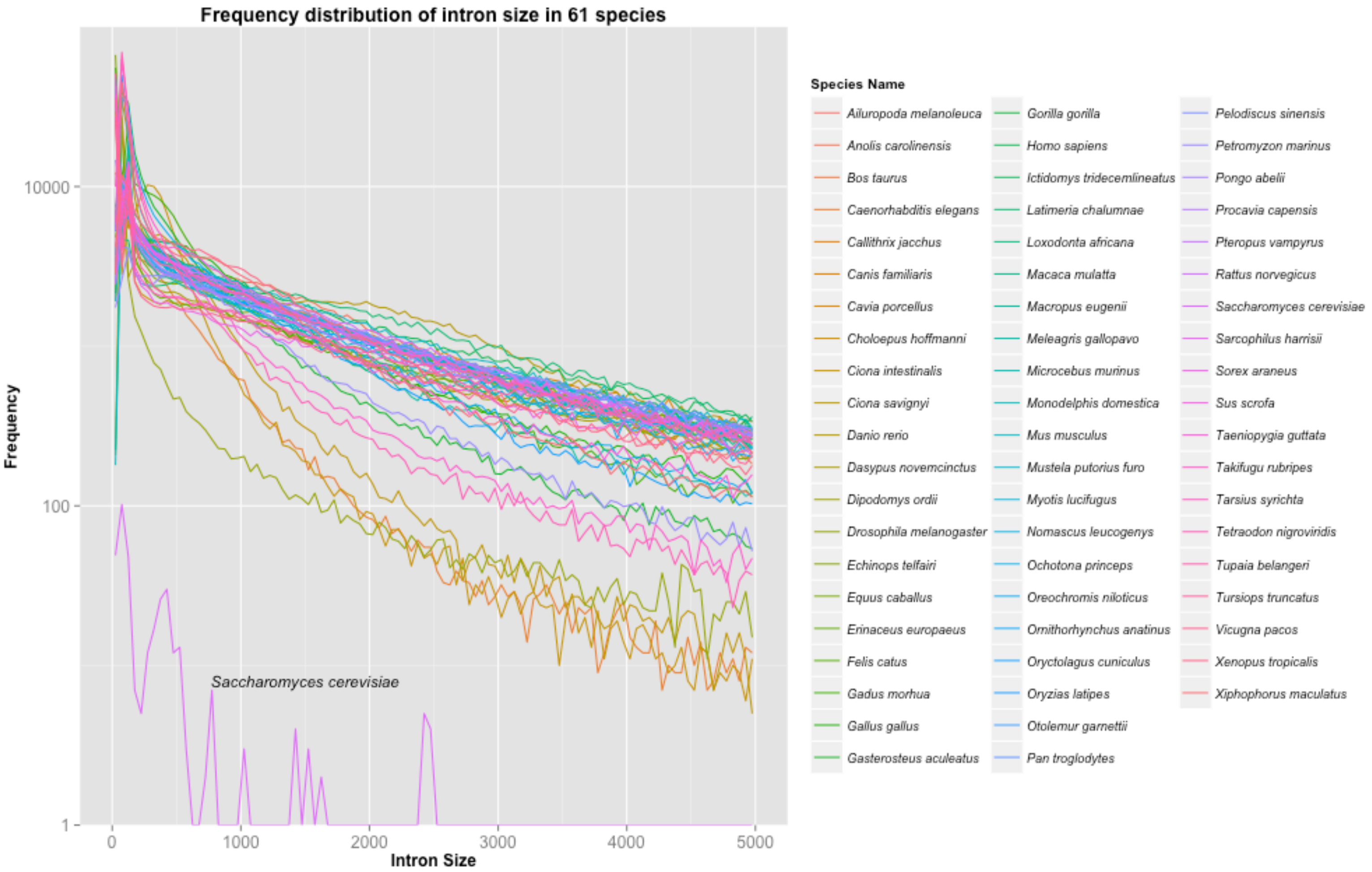


Figure 5.3 - Frequency distribution of intron size in all 61 species in Ensembl release 70. A cut-off of 5,000 bp intron size was used on the x-axis, though the data progresses up to a maximum of 4,384,418 bp at a frequency of approximately <= 1. Again the majority of the data are comparable as seen by the right skew in the distribution. See Appendix 5.3 for inset of 0 to 750 intron size.

Gene family and intron data correlations

Visualisations were plotted depicting the relationship between gene family size, and intron count, density and size using a pooled dataset of all the intron and gene family data from all 61 species (see Supplementary Figure 5.2, Figure 5.4 and Figure 5.5), showing a clear downward trend.

Correlation tests were performed between all data on intron count, density and size against gene family size using a Spearman's rank correlation test (see Table 5.4), showing a subtle positive relationship between intron count and gene family size, and a slight negative relationship between intron density and size against gene family size.

Table 5.4 - Table showing Spearman's rho for correlations between intron variable and gene family size pairs in 61 species.

Variable Pairs	rho
Intron Count ~ GF Size	0.06139003
Intron Density ~ GF Size	-0.04813767
Intron Size ~ GF Size	-0.005180116

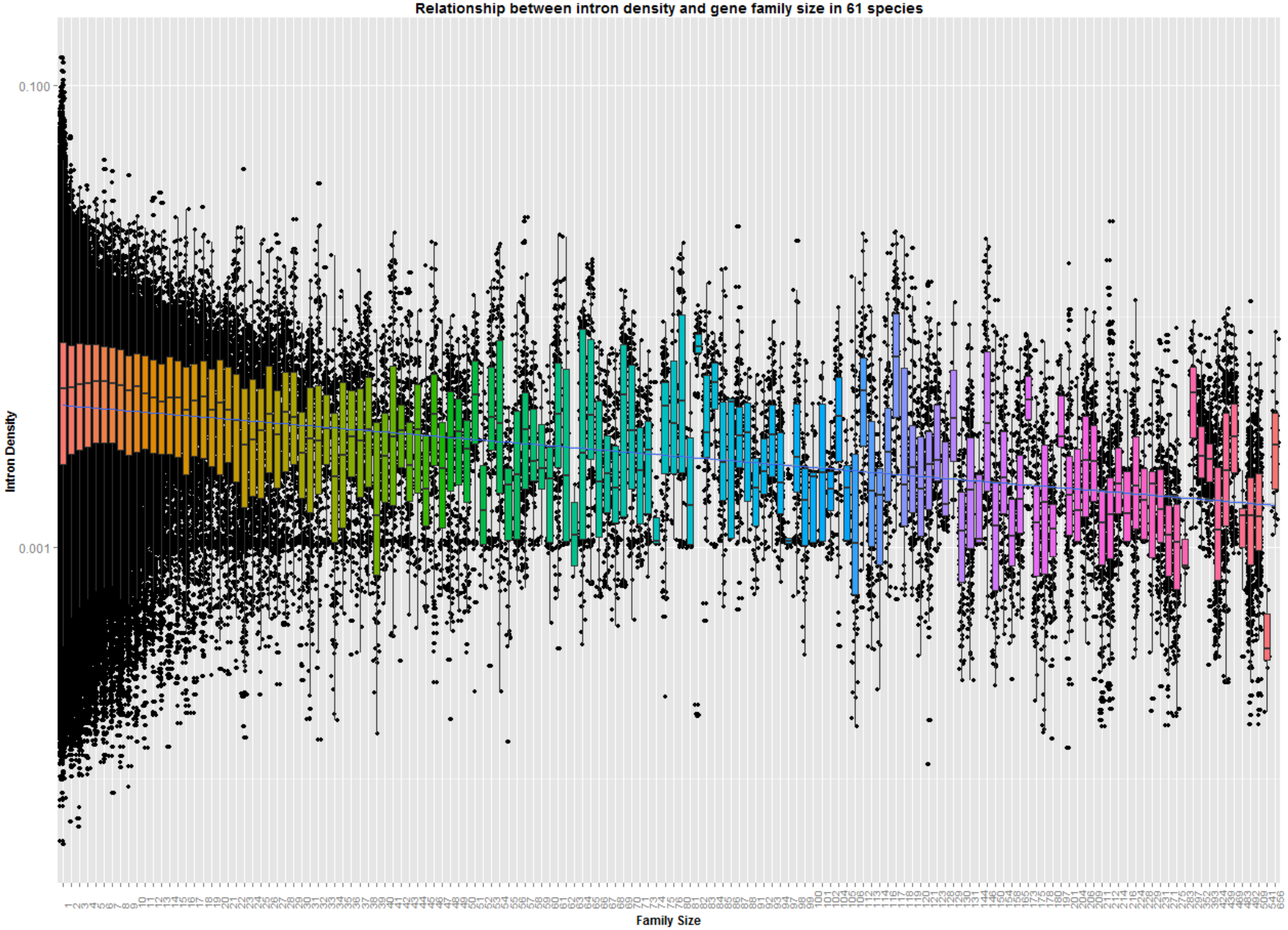


Figure 5.4 - A boxplot displaying the relationship between gene family size and intron density for the pooled intron and gene family data of all 61 species used in this study.

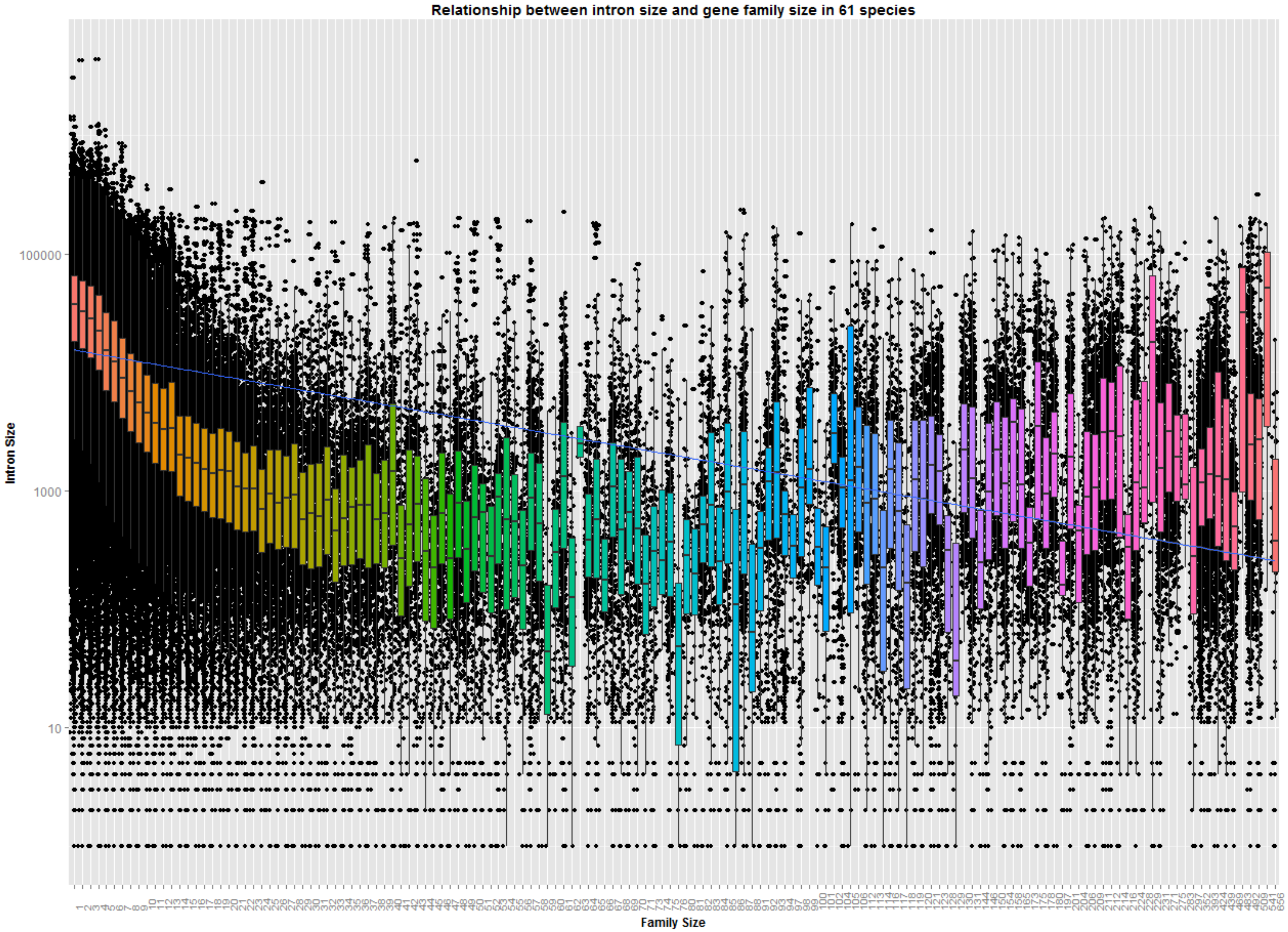


Figure 5.5 - A boxplot displaying the relationship between gene family size and intron size for the pooled intron and gene family data of all 61 species used in this study.

Intron characteristics by location on the chromosome

A total of 21,983 protein-coding genes were retrieved, of which 19,850 were identified as belonging to assembled chromosome sequences and thus considered for further analysis. The remaining 2,133 were part of unassembled contigs or non-standard chromosome pairs and thus removed. An additional 1,351 genes were discarded for having no introns, leaving 18,499 genes for downstream analyses. Of these 18,499 genes, a total of 186,650 introns (from 202,227 introns matching the full 21,983 protein coding gene dataset) were identified for analyses. A further 40 of these (leaving 186,610 introns) were trimmed for being less than 5 bp in length (93 out of the 202,227 intron dataset) as the absolute minimum requirement for a functional intron (though this is likely to be greater than 30 bp in reality – see Moss *et al.*, 2011). Intron count ranges from a minimum of 1 intron to a maximum of 362 introns. Intron density (introns per bond) ranges from a minimum of 0.00002969 to a maximum of 0.03571429. Intron length ranges from a minimum of 5 bp to a maximum of 1,240,120 bp. The mean intron density was calculated for windows of 250 Kb in size across chromosome 1 and plotted alongside an ideogram for *Homo sapiens* chromosome 1 (see Figure 5.6).

Sex chromosomes versus autosomes

Mean intron density was calculated per 250Kb window across the X and Y chromosomes (see Figure 5.7 and 5.8 respectively).

The Kruskal-Wallis Rank Sum Test was used to determine whether the samples for each intron characteristic came from the same distribution across groups of chromosomes. Intron count, density and size were tested against autosomes and sex chromosomes (defined by "Chromosome Name") to see whether a significant difference existed (see Table 5.6). A significant difference is highlighted for all but intron count and size in the sex chromosomes.

Table 5.6 - Table showing Kruskal-Wallis Rank Sum Test statistics for test of difference between intron variables and chromosome groups.

Measures and Groups	All Chromosomes	Autosomes	Sex Chromosomes

Intron Count ~ Chromosome Name	χ2 = 225.6736
	p < 2.2e-16
		χ2 = 200.4487
		p < 2.2e-16
			χ2 = 0.3927
			p = 0.5309

Intron Density ~ Chromosome Name	χ2 = 142.7524
	p < 2.2e-16
		χ2 = 85.3713
		p = 1.002e-9
			χ2 = 50.819
			p = 1.013e-12

Intron Size ~ Chromosome Name	χ2 = 6506.65
	p < 2.2e-16
		χ2 = 6509.875
		p < 2.2e-16
			χ2 = 0.4285
			p = 0.5127

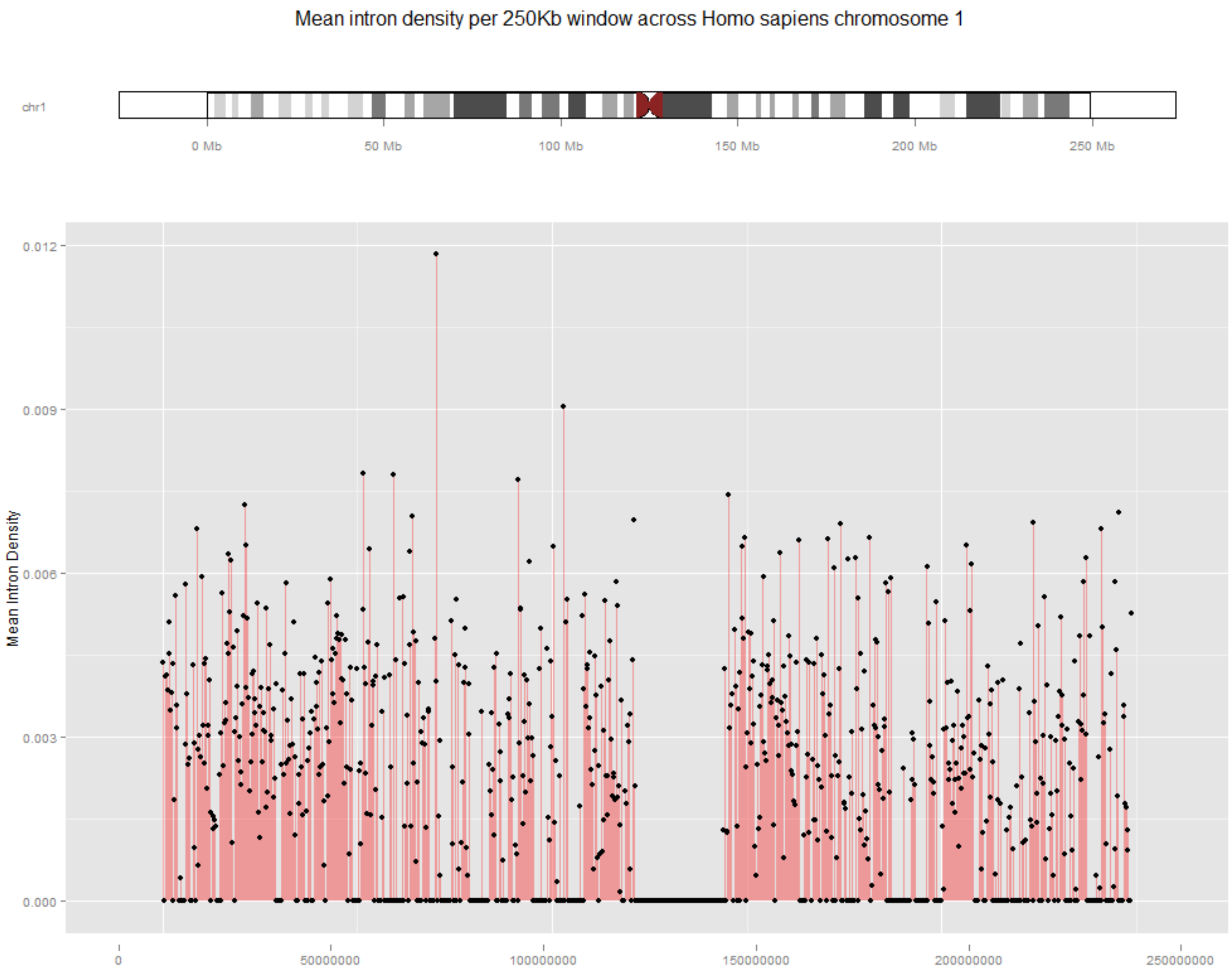


Figure 5.6 – Mean intron density calculated per 250Kb window across chromosome 1.

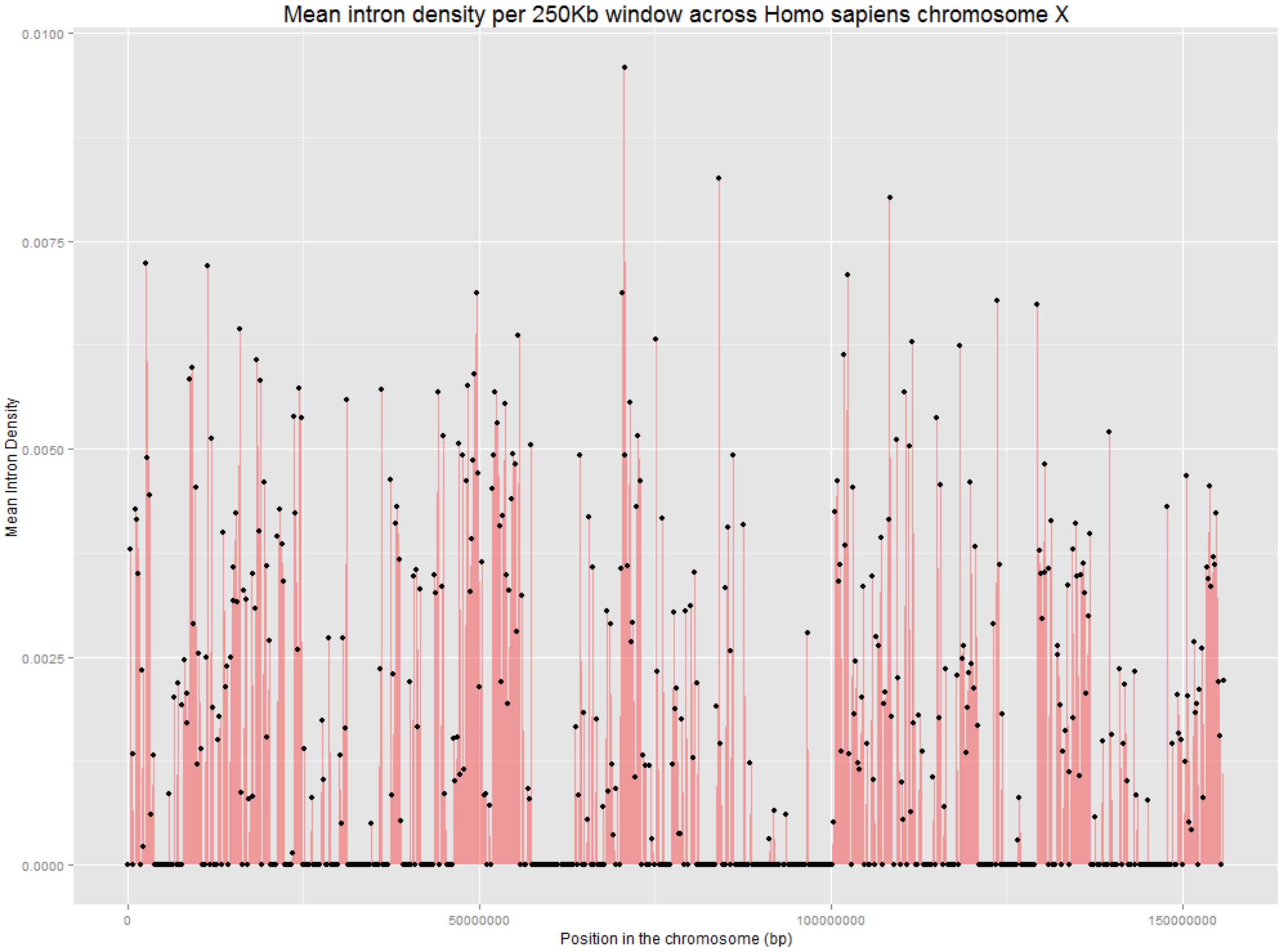


Figure 5.7 – Mean intron density calculated per 250Kb window across chromosome X.

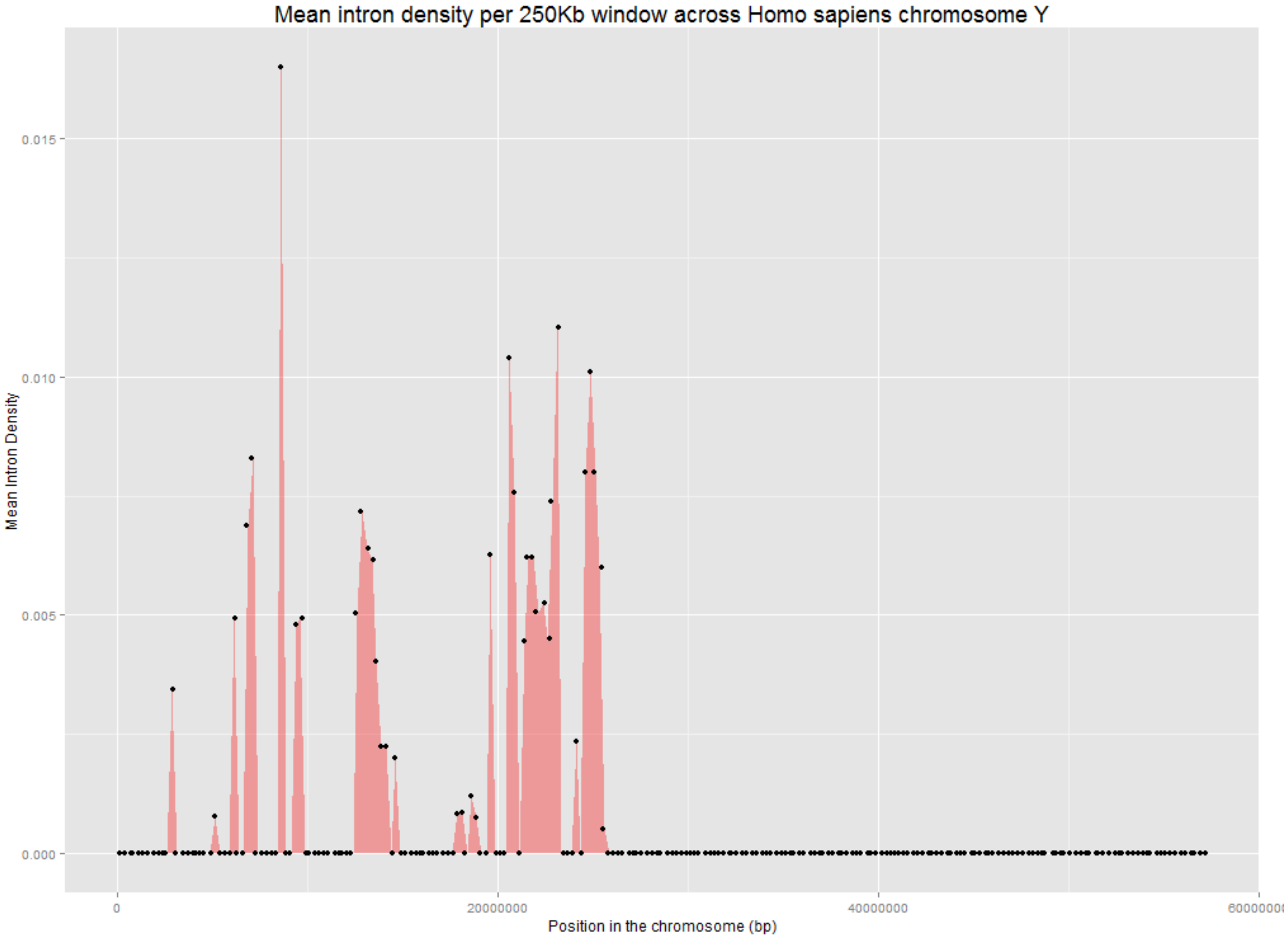


Figure 5.8 – Mean intron density calculated per 250Kb window across chromosome Y.

Discussion

This chapter attempts to understand whether gene duplication impacts on the evolution of introns, or whether there are other non-family effects that are more involved. There have been some attempts to understand the main drivers of intron evolution in relation to gene duplication previously (Deutsch and Long, 1999; Vinogradov, 1999; McLysaght *et al.*, 2000; Castillo-Davis *et al.*, 2004; Lin *et al.*, 2006; Chatterji and Pachter, 2007; Jabbari, 2013), albeit at a much smaller scale. This study takes a large-scale comparative genomics approach to the analyses of gene family size on intron evolution, and in doing so demonstrates that these data can be collated from a large number of species, allowing for comparisons across a diverse range of organisms with a variety of population and life history parameters. Previous studies that have suggested an increased rate of intron gain or loss have been limited to single organisms, or members of different subspecies (Castillo-Davis *et al.*, 2004; Lin *et al.*, 2006) and therefore the analyses in this chapter have much greater power. One might be concerned that that variation in assembly and annotation quality across multiple genomes might impact on the reliability of the results, as described by Ames *et al.* and Han *et al.* (2012; 2013). However, although the data here aren't controlled for these factors, the distributions are shown to be sensible and comparable throughout (see Figure 5.1 through Figure 5.5). This gives us increased confidence that we are likely working with the same classes of data, though being able to control for error in the models would be preferable. The volume and extent of the data being compared here provides further proof of the power of comparative genomics in determining the forces shaping genome structure and content.

Introns are very important as they contribute a large proportion of nucleotides towards the total size of higher eukaryote genomes, with vertebrate genomes being the primary focus of this study. Introns often contribute more to overall gene length than exons and UTRs (untranslated regions) combined (Deutsch and Long, 1999; Hong *et al.*, 2006), and there is often a correlation between intron size and genome size (Deutsch and Long, 1999; Vinogradov, 1999; McLysaght *et al.*, 2000). Introns are also very effective at highlighting neutral forces of molecular evolution, as well as some of the regulatory mechanisms involved in gene expression and splicing (Patel *et al.*, 2002; Patel and Steitz, 2003; Basu *et al.*, 2008; Rogozin *et al.*, 2012). Mechanisms such as replication slippage, and non-homologous recombination that are common in the gain and loss of introns (Roy and Gilbert, 2006; Yenerall and Zhou, 2012) are also involved in the duplication of genes (Zhang, 2003). One might expect, therefore, that larger gene families would have a greater density of introns due to an increased intron gain over intron loss, and there seems to be some evidence to suggest this in the literature (Babenko *et al.*, 2004), though change in size of introns, at least reduction in size, seems to be more linked to gene expression (Castillo-Davis *et al.*, 2002), with selection playing a role in maintaining smaller gene sizes (Zhang and Edwards, 2012). Chromosomal translocations are postulated to impact more on increased intron size (Jabbari, 2013), though recent evidence has highlighted a correlation between conservation at the protein level and increased intron burden (Gorlova *et al.*, 2014) potentially due to the benefits of conserving the gene and its splice variants outweighing the metabolic burden of maintaining the introns.

This chapter looks at a number of factors in relation to location of genes in the chromosome and intron evolution. Recombination rate is postulated to be an influential factor on the intron distribution (Comeron and Kreitman, 2000; Duret, 2001; Prachumwat *et al.*, 2004; Roy and Gilbert, 2006; Li *et al.*, 2009; Nam and Ellegren, 2012). I show a strong difference between the sex chromosomes and autosomes, and it is likely that recombination is a major contributing factor to this difference, which is discussed in greater detail below. The literature also describes a relationship between GC content, in particular larger scale GC rich isochores and the evolution of intron content in the genome (Zhu *et al.*, 2009; Fujita *et al.*, 2011; Chaurasia *et al.*, 2014; Sun *et al.*, 2015). The analyses in this chapter do not identify a clear relationship between GC content and intron characteristics and this agrees with the literature that shows there is likely to be a much more complicated relationship as the discussion below highlights (Fullerton *et al.*, 2001; Prachumwat *et al.*, 2004; Li *et al.*, 2009; Marsolier-Kergoat and Yeramian, 2009; Weber *et al.*, 2014). This chapter also investigates other chromosome location specific factors. The position of the introns at the chromosome level is explored in relation to whether higher-level forces such as replication timing and epigenetic modifications might play more of a role in the evolution of introns. We would expect to see a stronger relationship between chromosome location and intron density than there is with gene duplication in this case. Several studies highlight such non-family effects with larger introns being correlated with chromosomal translocations (Jabbari, 2013), genes with specific intron arrangements being conserved in given clusters of chromosomes (Salier, 2000; Sánchez *et al.*, 2003) and significant differences in autosomal vs X-linked introns, specifically in relation to GC isochores (Haddrill *et al.*, 2005). The spatial impact on intron characteristics was addressed here using data from *Homo sapiens* only as a pooled comparative analysis wouldn't be possible due to non-homologous chromosomes pairs across species. The use of the high quality annotated human genome, however, allows us to tease apart the contrasting signals from the different evolutionary forces at work across the genome in a more focused manner. Autosomes and sex chromosomes were also compared, allowing us a proxy for determining the effects of recombination on intron evolution.

Is intron evolution independent in different gene copies?

In understanding how gene duplication impacts on lower level molecular evolution, the gene family data for a diverse range of 61 species were retrieved and analysed. A total of 1,102,993 genes were fetched contained within 130,090 gene families. The pattern of gene family sizes followed the same approximation of a power-law distribution for all species, as previously highlighted in Chapter 4. Likewise intron data were retrieved for all 61 species accounting for a total of 10,139,168 introns. Intron count, density and size varied with species, which likely indicates the effect of population genetic forces and life history traits such as effective population size, generation time, and number of offspring on their underlying evolution. Species with larger effective population sizes are less susceptible to genetic drift, and more likely to see an effect of natural selection. For example, *Monodelphis domestica*, which can raise a relatively large number of offspring throughout their lifetime (Macrini, 2004), and therefore may be more susceptible to selection against larger introns through natural selection, had the largest intron sizes and a relatively large maximum gene family size, though the gene to gene family (G/GF) ratio was approximately 50%. This perhaps indicates a greater role of retrotransposition in increasing intron size in this species, which is feasible given the increased burden of transposable elements and other interspersed repeats within its genome (Mikkelsen *et al.*, 2007).

The data shows that 9.46% (95,360) of genes contain no introns. Although this is a cumulative value across all 61 species, it indicates that retrotransposition or gene conversion is a relatively common occurrence because these methods can result in the loss of introns, particularly in the case of retrotransposition where a processed transcript can be reinserted into the genome. Most retrotransposition events will result in pseudogenization on insertion (Lynch and Richardson, 2002) though some do get processed if they are inserted sufficiently close to downstream promoter sequences (Zhu and Niu, 2013). Most genes, at least in more complex eukaryotic species (this thesis focuses primarily on vertebrates), have introns (Gilbert, 1978; Lynch, 2005; Lynch, 2007). It is certainly possible that novel genes could arise without introns, though this is likely to be due to a recent retrotransposition event that places a transcript close to a downstream promoter sequence, which doesn't have a high probability of occurrence. It is more likely that these events will result in pseudoginisation on reinsertion, however. Pseudogenes aren't explicitly considered in this study however, so it is more likely that gene conversion plays the dominant role here. Gene conversion results in the homogenization of regions of 200 to 1,500 bp in length between orthologous (allelic) and paralogous (non-allelic or ectopic) genes. This has been shown to result in intron loss if the donor sequence contains no introns and gain if it does (Roy and Gilbert, 2005). The bias towards higher GC content in gene conversion predicts increased gene density and smaller introns (Galtier *et al.*, 2001), though intron loss has been shown to result from ectopic gene conversion (Morris and Drouin, 2011) particularly at the 3'-end, which perhaps provides some additional considerations for why first introns are shown to be longer (Bradnam and Korf, 2008).

Increased retrotransposition, non-homologous recombination, and gene conversion are just some of the mechanisms by which gene family expansion can impact on the evolution of introns (Crick, 1979; Sharp, 1985; Rogers, 1989; Derr and Strathern, 1993; Hankeln *et al.*, 1997; Roy and Gilbert, 2006; Irimia *et al.*, 2008; Roy, 2009; Li *et al.*, 2009; Yenerall and Zhou, 2012), however the pattern of intron count, density and size seems to follow a similar distribution regardless of whether we consider only introns within gene families or introns across the entire genome (see Chapter Two and Chapter Three; Moss *et al.*, 2011). One would expect that if gene duplication did impact on intron evolution that there would be a difference in intron characteristics that correlated with change in gene family characteristics. This is not the case here, and instead we see uniform patterns of intron evolution across different gene copies, highlighting common forces involved in intron evolution. Singleton genes weren't assessed in relation to intron content here, though would be a useful future direction to understand how non-family effects impact on intron characteristics regardless of position in the chromosome. Previous studies (Knowles and McLysaght, 2009; Bornberg-Bauer *et al.*, 2010) have highlighted that recent de novo singleton genes tend to be void of intronic sequence (however see Levine *et al.*, 2006 and Yang and Huang, 2011). De novo singleton genes arise from non-coding DNA and gain the capability of transcription and translation through mutation. This has been widely studied in yeast (Cai *et al.*, 2008) but applies to de novo genes without introns in their coding sequence. Introns can be gained by intronization of exonic sequences (Irimia *et al.*, 2008), but more often are linked to mechanisms such as intron transposition, transposon insertion, tandem genomic duplication, intron transfer between paralogs, and self-splicing type II intron insertion (Irimia *et al.*, 2008; Roy and Irimia, 2009; Zhu and Niu, 2013). In conclusions the analyses here point to a uniform process of intron distribution across gene copies, and implies the forces driving this are not specific to genes and more general across the genome.

Does size of gene family influence intron evolution?

Intron count, density and size follow a consistent pattern in all genomes with the observation of a unimodal, right-skewed distribution for all characteristics (see Supplementary Figure 5.1, Figure 5.2 and Figure 5.3), as previously described in the literature (Hong *et al.*, 2006; Moss *et al.*, 2011). This distribution represents the entirety of all individual intron data points pooled together, however; so all introns regardless of their location in the genome or the size of the gene family they belong to, are represented by their contribution to particular size classes. In order to understand how gene family size impacts on intron evolution however, it is necessary to examine the contribution of intron sizes within each gene family size class. We can do this visually using a simple boxplot. By examining the relationship between gene family size and intron characteristics both visually and statistically it is possible to build a more robust picture of the impact of gene family size on intron evolution. Comparing the intron characteristics visually allows us to 1) determine whether there is a loss of variation due to the decrease in number of observations with increase in gene family size, and 2) highlight any outliers in the distributions of intron sizes for each gene family size that might point towards a correlation between gene family size change and the evolution of intron characteristics.

The results show (see Supplementary Figure 5.2, Figure 5.4 and Figure 5.5) that there are a lower number of intron data points seen with increase in gene family size. This might be expected, as the total number of genes contained within each gene family class follows an approximate power-law distribution (see Figure 5.1) as previously described in the literature (Huynen and Nimwegen, 1998; Luscombe *et al.*, 2002). This downward trend is confirmed by the Spearman's rank correlation test (see Table 5.4), which returns a negative relationship between intron density and intron size, against gene family size. Intron count, however, returns a positive relationship, although there is clearly a downward trend in the plot (Supplementary Figure 5.2). This visual determination of a downward trend is likely due to the exceptionally large number of introns contained within gene family size classes of less than 30 members, as the trend above 30 shows a positive correlation. In general, there seems to be a distinct difference between gene families with less than 30 members and gene families with more than 30 members. A positive correlation between recombination rate and average gene family size has previously been observed (Tiley and Burleigh, 2015), which may partly explain the differences here, though why there would be a distinction between families with fewer members and families with greater members is still unclear. It is possible this could be due to annotation errors in larger gene family size classes. We would expect to see more introns overall within those size classes that exhibit a greater cumulative total number of genes. This is dependent on the average number of introns within the genes being the same however, which seems not to be the case in all species (Koonin *et al.*, 2012). Indeed, if the number of introns were the same in all genes we should expect to see a pattern of intron density that follows the same power-law distribution as gene family size. Instead we see a unimodal right-skewed distribution (see Figure 5.4 and Figure 5.5). This intron pattern is consistent across all genes, not just those in gene families (see Chapter Two and Chapter Three; Moss *et al.*, 2011). It is likely that the variation in age, selective pressure, and recombination frequency within individual gene families contributes to a deviation from the neutral expectation (Roy *et al.*, 2002; Rogozin *et al.*, 2003; Babenko *et al.*, 2004; Basu *et al.*, 2008; Kordiš and Kokošar, 2012).

We can plot gene family size against intron count, density and size along with a linear model of the data to understand the relationship between these variables in greater detail. The overlap in confidence intervals seen in Figures 5.5 through 5.7

demonstrates that intron count, density and size come from the same distribution regardless of their member gene family size. By performing a Spearman's rank correlation test the relationship between the data can be described using a single statistic. The relationship seen for intron density and intron size are closer to 0 (see Table 5.5), though exhibit a slight negative relationship, meaning that as gene family size increases intron density and size displays a very slight decrease. Intron count in contrast shows a slight positive relationship (see Table 5.5), meaning that as gene family size increases, intron count also increases slightly. These visualisations and tests tell us that gene family size has a weak influence at best on the underlying evolution of intron characteristics.

There is some evidence to suggest that intron density is greater within larger gene families (Babenko *et al.*, 2004), though we see the opposite of that here. Intron gain appears to be limited to specific types of gene such as domesticated genes (Kordiš and Kokošar, 2012) and plastid-derived genes (Basu *et al.*, 2008) as well as being more prevalent in vertebrates and plants (Rogozin *et al.*, 2003; Babenko *et al.*, 2004; Koonin *et al.*, 2012). The weak positive relationship between intron count and gene family size in our data may be a reflection of this, with the larger gene families (such as zinc-finger genes) tending to have adaptive relevance (Kordiš and Kokošar, 2012; Brunner *et al.*, 2014). A previous study in rice (Lin *et al.*, 2006) identified a bias towards intron loss in duplicate genes, with cases of intron gain being due to transposon insertion or retrotransposition of pseudogenes. Conversely, an analyses of introns in duplicate genes in human and mouse malarial parasites (Castillo-Davis *et al.*, 2004) showed an increased acceleration of both intron gain and loss due to relaxed selection and/or positive selection in duplicate copies, along with a weak correlation between protein divergence and intron gain/loss in orthologs, but not paralogs. This points to processes that work independently of duplication, and that are more intrinsically linked with the function and structure of genes.

Is there heterogeneity in intron evolution across chromosomes, and between sex chromosomes and autosomes?

If the size of the gene family that introns are contained within doesn't impact significantly on the evolution of their count, density or size, then there must be other forces at play. One possibility is that there are localised effects that exert a bias towards increase or decrease in intron characteristics, with mechanisms such as GC bias within isochores across the genome (Zhu *et al.*, 2009; Fujita *et al.*, 2011; Chaurasia *et al.*, 2014; Sun *et al.*, 2015) and recombination rate (Comeron and Kreitman, 2000; Duret, 2001; Prachumwat *et al.*, 2004; Roy and Gilbert, 2006; Li *et al.*, 2009; Nam and Ellegren, 2012) being strong contenders. These mechanisms seem to have predominance towards smaller intron size and fewer introns per gene however, which may be due to the positive correlation seen between GC bias and recombination rate (Fullerton *et al.*, 2001; Marsolier-Kergoat and Yeramian, 2009; Weber *et al.*, 2014). In contrast *Caenorhabditis elegans* demonstrates a positive correlation with intron number and recombination rate and also a positive correlation between introns of 100-1,000 bp (Prachumwat *et al.*, 2004; Li *et al.*, 2009), with the exception of the X chromosome that interestingly sees introns of > 1,000 bp arranged according to the recombination rate across the chromosome (Li *et al.*, 2009). Generally the impact of selection and recombination rate are much weaker and non-significant in introns of < 100 bp and > 1,000 bp, however.

The pooled intron data for *Homo sapiens* was analysed using a Kruskal-Wallis test, the non-parametric equivalent of the one-way analysis of variance (Kruskal and Wallis, 1952), to test whether intron counts, intron densities and intron sizes came from the same distribution across different chromosome groups. The Kruskal-Wallace test showed that when analysing all chromosomes together, and just autosomes that these samples originated from the same distributions with a high significance ($p=2.2x10^{-16}$, $p=2.2x10^{-16}$, $p=2.2x10^{-16}$ and $p=2.2x10^{-16}$, $p=1.002x10^{-9}$, $p=2.2x10^{-16}$ respectively – see Table 5.6). When comparing sex chromosomes, however, this isn't that case and there is no significant result for intron counts and intron sizes (p=0.5309 and p=0.5127), with the exception of intron densities ($p=1.013x10^{-12}$) (Table 5.6). The difference seen with the sex chromosomes is likely due to the composition of the Y chromosome skewing the results. The Y chromosome has experienced massive gene decay and degeneration, due mostly to reduction or loss of recombination and the accumulation of repetitive elements (Charlesworth and Charlesworth, 2000; Bachtrog, 2013). The Y chromosome is an extreme case, however, demonstrating how increased mutation rate, and reduced effective population size and recombination rate can impact on underlying sequence evolution (Charlesworth and Charlesworth, 2000) not just in humans (Cortez *et al.*, 2014). The increased mutation rate can lead to increased divergence between sequences, which is particularly the case in introns that are more subject to drift than selection, with the exception of some MSY-specific gene families that see > 98% nucleotide identity among family members in both introns and exons (Skaletsky *et al.*, 2003). Interestingly there was still a significant relationship between the distribution of intron densities across the X and the Y-chromosomes. This is likely due to the number of introns per bond, the definition of intron density used here, being similar in both groups due to the decreased size of genes on the Y chromosome, though the overall count and size of those introns varies considerably.

To gain greater insight and understanding of the distribution of intron characteristics across the chromosomes, mean intron density was calculated per 250Kb window across each chromosome, with a focus on chromosome 1, the X chromosome and the Y chromosome (see Figures 5.8, 5.9 and 5.10). The mean intron density showed a spatial pattern that seemed to fit with the distribution of cytogenetic bands, with 0 intron density being seen in heterochromatic regions, particularly around the centromeres and telomeres (Figure 5.6). The non-existence of introns in these regions is perhaps expected given the levels of variable and suppressed recombination (Choo, 1998), and lack of genes reported in a number of species (Sun *et al.*, 1997; Hosouchi *et al.*, 2002) particularly humans (Schueler *et al.*, 2001). However, it is wrong to assume that centromeres never experience transfer of genetic material, as past studies have identified active genes and widespread gene conversion (Nagaki *et al.*, 2004; Shi *et al.*, 2010) including the existence of introns in these regions (Nagaki *et al.*, 2004). This is in contrast to telomeres, which do experience homologous recombination, but are devoid of genes likely due to their variability in length through incomplete DNA replication as part of their involvement in genomic stability (Royle *et al.*, 2009; Basenko *et al.*, 2011). It is interesting to note, however, that sub-telomeric regions have been shown to contain rapidly evolving gene families as a result of frequent sub-telomeric gene conversion (Riethman, *et al.*, 2001), though these regions have been shown to be intron poor (Hunt *et al.*, 2001). As with telomeres, centromeres typically consist of tandem repeats that are highly homogenous in nature, though despite their conserved functionality they have been observed to evolve rapidly (Malik and Henikoff, 2009). The propagation of mutations in these regions has been proposed to be the result of gene conversion (Talbert and Henikoff, 2010), and indeed centromeres have been shown to experience levels of gene conversion no different to non-centromeric regions (Symington and Petes, 1988). Gene conversion is typically linked with intron loss, and would account for the reduction or lack of introns in these regions (Derr and Strathern, 1993; Roy and Gilbert, 2006).

The variability in the existence of centromeric genes in different species may in part be down to the difficulties in sequencing these regions, with long-read sequencing highlighting the existence of genes in previous assembly gaps (Aldrup-MacDonald and Sullivan, 2014), though there is unlikely to be excessive increases in the number of identified genes. That being said, given the scarcity of introns within regions of lower recombination identified in this study, particularly on the Y chromosome, it seems likely that recombination plays a key role in their maintenance and proliferation. This is in contrast with previous evidence that describes more and longer introns in regions of lower recombination (Comeron and Kreitman, 2000; Duret, 2001; Lynch, 2002; Prachumwat *et al.*, 2004; Roy and Gilbert, 2006; Li *et al.*, 2009; Nam and Ellegren, 2012), with the exception of *Caenorhabditis elegans* (Prachumwat *et al.*, 2004; Li *et al.*, 2009). Increased recombination may see increased occurrence of gene conversion on the basis of GC content (Galtier *et al.*, 2001; Weber *et al.*, 2014), and it seems likely, therefore, that the variability in GC content plays a big role in the fate of introns, with greater GC-content being shown to result in biased gene conversion and intron loss (Derr and Strathern, 1993; Roy and Gilbert, 2006). The process is complicated, however, and there is an intricate involvement of effective population size, recombination rate, and length of the heteroduplex in determining how effective biased gene conversion can be (Galtier *et al.*, 2001). This seems to be supported by low evidence of gene conversion (Katju and Bergthorsson, 2010) alongside a low (36%) GC-content (The C. elegans Sequencing Consortium, 1998) in *Caenorhabditis elegans*, with variable effective population sizes often < 10^4-10^5 (Félix and Duveau, 2012) and suppressed recombination in wild populations (Rockman and Kruglyak, 2009). It is clear that there are numerous influences on intron evolution, with non-family effects playing a predominant role.

Conclusions

The objectives of this chapter were to address the following questions; 1) Is intron evolution independent in different gene copies? 2) Does size of gene family influence intron evolution? 3) Are there other non-family effects, which influence intron evolution? In undertaking the analyses to answer these questions I have firstly been able to show that the necessary data can be easily collated and is sensible and comparable across vertebrate species. The detailed analyses show that intron characteristic distributions are largely uniform across all species and follow similar patterns regardless of gene family membership, as described in previous studies (Fedorov *et al.*, 2002; Zhu *et al.*, 2009; Moss *et al.*, 2011). There is a weak correlation at best between gene family size and the intron characteristics studied, which is contrary to what one might think, as recombination rate and gene family size have been found to be positively correlated (Kong *et al.*, 2004). This indicates that other mechanisms are more central in the evolution of introns, though are not necessarily entirely independent in different gene copies, due to the impact of similar evolutionary processes. The big difference between the sex chromosomes highlights the most extreme case of these processes at work, which ultimately refer to variations in population genetic parameters, recombination rate and GC-content.

There seems to be a strong spatial element too with intron characteristics being impacted by chromosomal translocations, and location within specific clusters or GC-rich areas of chromosomes (Jabbari, 2013; Salier, 2000; Sánchez *et al.*, 2003; Haddrill *et al.*, 2005). Euchromatic sequences in particular display vastly different intron characteristics in comparison with heterchromatic regions (Sun *et al.*, 1997; Hosouchi *et al.*, 2002; Schueler *et al.*, 2001; Nagaki *et al.*, 2004). This highlights the potential involvement of epigenetic forces at a higher level, and there have been a number of studies highlighting the involvement of epigenetic processes in replication timing, and the formation of GC-rich isochores (Oliver *et al.*, 2001; Paċes *et al.*, 2004; Schmegner *et al.*, 2007; Costantini and Bernardi, 2008; Watanabe *et al.*, 2009; Costantini *et al.*, 2013). The positive correlation between GC-content, and recombination rate and gene conversion as well as their impact on intron characteristics (Fullerton *et al.*, 2001; Galtier *et al.*, 2001; McVean *et al.*, 2004; Meunier and Duret, 2004; Duret *et al.*, 2006; Duret and Galtier, 2009) demonstrates how epigenetics might impact at this level. However, more work needs to be done to understand the intricacies of these molecular processes in the evolution of introns. In particular an across and within species comparative analyses of the impact of GC-content, recombination rate, and effective population size on intron evolution would be a helpful step forward in teasing these signals apart.

[This page is left intentionally blank]

CHAPTER SIX: DISCUSSION

This thesis has discussed the development and application of computational methods to the analyses of genomic data using a large-scale comparative approach. The power of comparative genomics analyses is highlighted throughout, with comparisons being made between groups of species diverging by millions of years. The importance of analysing the data within a robust evolutionary framework is emphasised in order to highlight the underlying biological signal in the vast volumes of genomic data. The work here shows the strength of comparative genomic analyses in identifying changes in the structure and content of genomes over time; though it also demands consideration of the assumptions made by the methods and algorithms used in their analyses. In addition to considering methodological bias, the nature of the underlying data is brought into question. The need for validation of genome assemblies and an allowance for error correction in analyses is a welcome step forward in recent years (Medvedev and Brudno, 2009; Medvedev, 2011; Salzberg et al., 2012; Ilie and Molnar, 2013), though much is still to be done to insure we aren't building our conclusions on top of artefacts in the data.

The need for automated genome informatics pipelines

In chapter two I show the necessity for developing comprehensive software pipelines that provide automated and reproducible approaches to comparative genomics analyses. Generic toolkits for this type of analyses are currently lacking, which has resulted in numerous different in-house approaches to the analyses of genomic data. This complicates the ability to reproduce the analytical steps of previous studies (Tan et al., 2010; Hothorn and Leisch, 2011). In many instances the code and description of methods used aren't exhaustive, which necessitates a great deal of effort in understanding how the data has been manipulated from one stage of its analyses to the next. If this fails then contact with the authors of the study is necessary, which presents problems of its own. For example, if the data are old, then the individuals responsible for its analysis may be no longer available, or at least difficult to track down. There may also be reluctance to share approaches due to fear of being scooped or having flawed analyses identified. Additionally, requests for co-authorship in exchange for such assistance can sour collaborative efforts.

GCAT attempts to rectify these problems by providing a robust and dynamic framework for the retrieval, analysis and visualisation of genomic data. It builds on top of the wealth of genomic data available in the Ensembl genome databases (Hubbard, 2002; Kersey et al., 2010; Flicek et al., 2013), which are among the most comprehensive in the world. By utilising these data I am able to highlight the power of GCATs approach to broad-scale genome informatics analyses. It is important to utilise basic file formats, simple APIs, and automated workflows in order to improve both the ability to reproduce one's results and the ease of use of such tools (Parker et al., 2003; Goble et al., 2010; Prlic and Procter, 2012). The investment in time required for learning how to use GCAT is reduced by taking this approach. Many common functions (such as retrieving all the introns for a given species) that can be relatively complicated when using the Ensembl Perl API (Stabenau et al., 2004), are wrapped into single subroutine calls or compartmentalised scripts. The requirement for previous programming experience is acknowledged however.

The need for increased adoption and acceptance of scientific computing

Producing user friendly, robust and comprehensive genome analysis toolkits is imperative, however there is quite often a lack of time or inclination to learn new methods amongst many scientists (Ranganathan, 2005; Kumar, S. & Dudley, J., 2007; Schneider et al., 2010). The diversity of computational knowledge and experience provides differing barriers to the adoption of new tools. Learning to program in an additional language is generally simpler than learning a language from scratch for example. Tools are often aggregated into classes according to the programming language in which they are developed; based on the popularity of the language, how common the language is in the field, or personal preference. This sort of approach fails to ask what might be the best language for the job. More complex algorithms tend to require programming in lower-level languages such as C, though the learning curve here is extremely high. As many scientists are self-taught programmers with only basic computing skills, learning the intricacies of C may not be feasible. It isn't clear how one can improve the uptake of computer programming or use of more suitable platforms (i.e. UNIX-based systems) for analyses (though see Dudley and Butte, 2009; Harold et al., 2011; Wilson, G., 2013; Petre and Wilson, 2013; Wilson et al., 2014), but one way of alleviating issues with learning new languages is to increase the availability of web services (Stein, 2002; Stein, 2008; Stein, 2010; Lord et al., 2004; Ramirez et al., 2011).

The Semantic Web (Berners-Lee et al., 1999; Berners-Lee et al., 2001) is an initiative that seeks to promote common data formats and sharing of data, predominantly between computers. The World Wide Web does an excellent job of distributing knowledge and information (Frystyk et al., 1999; Berners-Lee et al., 2010), however the format for its dissemination is designed for reading by humans. By improving the integration of computer networks with the flow of information, in a machine readable as well as human readable format, it is possible to harness the power of the Internet to do much of the legwork when it comes to propagating biological data and metadata (Stein, 2002; Stein, 2008; Stein, 2010). In chapter two I discuss how web services such as those provided via the REST architectural style (Fielding and Taylor, 2000; Fielding, 2000a; Fielding, 2000b) are likely to be the future when it comes to sharing and analysing data. These services provide simple interfaces for data retrieval that, particularly in the case of REST, can be accessed using simple HTTP methods. This has the benefit of allowing any tool that can utilise HTTP to be used for the retrieval of data, whether that be a web browser, command line utility, or programming language. This moves towards overcoming the reliance on programming language dependent tools. I show the power and simplicity of web services in the retrieval of genomic data, though there is much work to be done. The wider adoption of cloud computing in bioinformatics has been requested for over a decade (Stein, 2002; Stein, 2008; Stein, 2010). Its slow uptake is perhaps due to lack of appropriate infrastructure for the transfer of petabytes of biological data. Networking hardware and bandwidth are improving greatly however and so one would hope we will see a change, though large investment is needed, which is questionable given current financial climates (though see Gross, 2011; Crosswell and Thornton, 2012). Most importantly there needs to be acceptance and backing by biologists in order to change the paradigm. There seems to be a very disjointed approach towards research in many cases and realisation of the interdisciplinary nature of data analyses is paramount.

Using computational approaches to understand genome evolution

In chapter three I identify how directed exploratory data analysis (EDA) followed by focused hypothesis testing allows for the identification of patterns in genome scale data that can then be subject to intense scientific rigour. This approach towards understanding the distribution and evolution of intron sizes in the genomes of five teleost fish highlights a difference in the distribution of intron sizes in the zebrafish, *Danio rerio* (Moss *et al*., 2011). Exploratory data analysis is a powerful approach towards understanding data, especially given the growing volume and complexity of biological data sets. It is important to ground any such analyses within the remit of the specific biological questions, however. Blindly searching for patterns in biological datasets will likely yield many results due to the probabilities of sequences being repeated, but these are likely to encompass an overwhelming amount of noise. By proposing specific questions and directing EDA towards understanding well defined subsets of the whole genome data, it is possible to examine any prominent similarities or differences highlighted in the data using more stringent hypothesis testing approaches (Lindenmayer et al., 2012; Michener and Jones, 2012).

Due to the volumes of genomic data now available, many projects focus on EDA in the first instance, followed up by rigorous hypothesis testing. This has the benefit of highlighting interesting patterns in the data that can then be the focus of more objective scientific scrutiny, thus removing a lot of the noise. This is a common approach in computer science; reducing the complexity of computational analyses by reducing the search space and thereby decreasing the computational burden (Hsiao et al., 2006; Miller et al., 2010; McKenna et al., 2010; Ekanayake et al., 2013). EDA within well-defined boundaries can be seen as a powerful means of directing genomic analyses. These methods have been met with conflicting responses from different groups of scientists however (Jones et al., 2006; Lindenmayer et al., 2012; Michener and Jones, 2012; Hampton et al., 2013). There are those that believe that the data first approach undermines the traditional hypothesis testing scientific method (Lindenmayer et al., 2012). This opinion is mostly taken by individuals that work with datasets of no greater than a few gigabytes in size and that have assiduously collected and refined them over several decades of their careers. Unfortunately this view and approach is unrealistic given the fast-paced nature of genomics and particularly so when considering large-scale inter- and intra-species comparative analyses. Novel and creative approaches need to be developed that encompass an underlying knowledge of molecular evolution (Hsieh, 2002; Petrov, 2002; Koonin, 2011; Lynch, 2011; Paten et al., 2013; Luo, 2014), with robust experience of the system's biology and ecology, in addition to a contemporary and forward thinking pursuit of the interpretation and analyses of these data.

Identifying the forces shaping genome architecture

In chapters four and five I show how comparative genomics is a powerful means of identifying the forces shaping the evolution of genome architecture. Changes in the structure and content of genomes can be highlighted by comparing similarities and differences within and between species. It is important that comparative analyses are also grounded by specific questions however. As previously discussed, blindly comparing and contrasting the genomic data will inevitably lead to the identification of patterns that just represent random noise due to the inherent structure of the data. Understanding that similarity doesn't equate to homology (Pearson, 2013) and approaching interpretation of results with a strong understanding of the molecular mechanisms driving any potential patterns in the data is paramount. When properly used comparative genomics can be utilised to examine how genomes have changed over time between species that have diverged by millions of years or, as is the growing trend, comparisons at the whole organism or single cell population level.

What approach one takes is dependent on the questions being asked. In understanding how duplications have shaped the evolution of the genome I take an interspecies comparative approach. If I wanted to determine variation in copy number within a species, I would need to examine whole-organism population level data and perhaps use that knowledge in my interpretation of the previous duplication data. If my question concerned an understanding of the causes or susceptibilities to a particular disease then I would need to look at whole-organism, or cell line or tissue specific population data as part of a controlled studied. It is important to consider the bigger picture and focus one's data analyses towards that. This is one of the biggest criticisms of genome wide association studies, which tend to search for variants such as SNPs across groups of individuals with and without disease and then determine a causative role for any changes in the diseased individuals based on correlations between particular variants. This is of course a simplistic summary of the approach, but emphasises the need to consider the data in its broader biological context to ensure robust conclusions are made. Genome wide association studies might be considered as a useful form of EDA however, identifying regions of interest that can then be subject to more objective scrutinisation using traditional hypothesis testing in well-defined and controlled experimentation. An interesting outcome of changes to the cost of genome sequencing that incorporates a creative approach to genomic data analyses is the advent of personal genomics. Companies such as 23andme can sequence an individual's genome (or rather differences from the reference) for as little as $100 and then use this data in addition to family and medical history to generate a profile of susceptibility to disease, identify particular character traits, or assist in determining ancestry. It can be seen as a means of crowdsourcing genome wide association studies. The greater volume of genomes sequenced, the more population level data is made available and the more powerful the approach becomes, as a consensus in allele frequency for particular traits are identified.

Comparative approaches that focus on identifying mutations such as SNPs have the benefit of highlighting how shorter term changes in the genomes of individual cells or hereditary mutations can impact an organism (Zhang et al., 2009; Zöllner and Teslovich, 2010; Teng et al., 2015). Of course, variation is the raw material for evolution and so mutations that have become fixed over longer time periods can be useful in highlighting how organisms have adapted to environmental pressures, undergone bottlenecks in population size or changed due to the random nature of evolutionary processes (Nielsen et al., 2005; Wright and Andolfatto, 2008; Harris, 2008; Iskow et al., 2012). These mutations have little effect on the overall structure of the genome however, at least in less divergent species. Duplicates, repeats and introns have been shown to comprise the greatest portions of the genome (Lynch, 2003; Lynch, 2007) and much of this thesis focuses on the role they play in changes in genome complexity over time. It is the forces that drive the propagation of these elements at the molecular level that must be considered when interpreting the data however. It seems obvious that natural selection has a limited ability to impact on the evolution of many organisms. This is particularly relevant in so called “higher eukaryotes” where the functional components of genome are relatively small in comparison to bacteria for example (Pushker et al., 2004; Parfrey et al., 2008; Kejnovsky et al., 2009; Delihas, 2011; Doolittle, 2013; Niu and Jiang, 2013; Graur et al., 2013). Population genetic forces coupled with larger scale changes at the molecular level such as recombination, duplication, retrotransposition, and the impact of chromosomal location through GC-rich isochores for example, have much greater influence on how the genomes change over time (Charlesworth and Charlesworth, 2000; Galtier et al., 2001; Lynch, 2002; Kejnovsky et al., 2009; Melé et al., 2012; Weber et al., 2014).

In summary, by using a broad-scale comparative genomics approach, I show in chapter three that the evolution and propagation of introns seems to be driven by nonhomologous recombination, and it is likely that they originate from the proliferation of repetitive and transposable elements. Duplications seem to have little impact on the evolution of introns however, as discussed in chapter five. The fixation of duplications seems at least in part to be a result of selective pressures and adaptation to environmental triggers over time as shown in chapter four (and see Force et al., 1999; Lynch and Conery, 2000; Lynch et al., 2001; Lynch and Conery, 2003; Lynch, 2011; Lynch, 2012). The shorter-term impact of duplication highlights a great deal of variation in copy number that is then subject to the influences of population size and other life history traits, though the nature of the data in reaching reliable conclusions is addressed. The forces driving duplications are similar to those experienced by introns, particularly non-homologous recombination (Castillo-Davis et al., 2004; Lin et al., 2006; Iskow et al., 2012; Xu et al., 2012). As most duplicates contain introns, it is unlikely that retrotransposition of mRNA is a factor. Retrotransposition can often result in the creation of pseudogenes (Esnault et al., 2000; Mighell et al., 2000; Sen and Ghosh, 2013) however, which aren't consider in these analyses. Finally in chapter five I show that it is chromosomal location and therefore biases exerted on particular regions of the genome, such as GC-rich isochores or epigenetic modification that drive changes in gene density and intron density across a divergent group of species. It is clear that the evolution of the genome is extremely complex with a great number of parameters to consider in determining how and why it has changed over evolutionary time scales. There is certainly an influence at lower levels from the standard molecular genetic forces of mutation, recombination, migration, selection, and random genetic drift; though these seem to be directed by larger-scale arrangements at the chromosome level but ultimately are limited by population level forces and life history traits (Soltis and Soltis, 1999; Charlesworth and Wright, 2001; Lynch and Conery, 2003; Lynch, 2007). The extremes can be seen when we observe the huge differences between bacteria, which have much shorter generation times and larger effective population sizes, in contrast to plants, which have vastly longer generation times and smaller population sizes overall.

Conclusions

The limited power of algorithmic and statistical methods

Throughout this thesis the power of current algorithms and statistical methods is proven to be limited given the volumes of error prone biological data (Brenner, 1999; Devos and Valencia, 2001; Hubisz et al., 2011; Han et al., 2013). Most methods take an approach that make many assumptions or that otherwise reduce the complexity of computational analysis by approximating the results (Katoh et al., 2002; Stamatakis et al., 2005; Hsiao et al., 2006; Price et al., 2009; Money and Whelan, 2011). This comes at the detriment of reliable conclusions, especially when considering the cumulative effect of all processing of the data. The current state of computational power requires that we approach the manipulation and analysis of large, complex datasets this way however. To apply an accurate algorithm to the analyses of most genome data would exceed the available CPU operations, memory capacity and other performance measures, making for very inefficient use of computer time, if such an algorithm is even feasible (Wang and Jiang, 1994; Jones and Pevzner, 2004; Money and Whelan, 2011; Mahmoody et al., 2012). This is even more relevant in comparative genomics. The utilisation of cloud computing, including high-throughput and high-performance computing, alongside the development of parallelized approaches to these problems is a way of mitigating these issues (Rognes and Seeberg, 2000; Zomaya, 2006; Vera et al., 2008; Luebke, 2008; Manavski and Valle, 2008). Parallelized algorithms allow analyses to be run on distributed computer hardware with much greater specifications and capacity, and thus greatly reduce runtime (Rognes and Seeberg, 2000; Manavski and Valle, 2008; Vera et al., 2008). Not all problems are capable of being optimised for parallelization however. This requires approaching things from a different perspective, including rethinking the underlying chemistry, computational models and infrastructure.

Next generation sequencing technologies are improving greatly as there is a great deal of competition to do so (Metzker, 2010; Niedringhaus et al., 2011; Zhang et al., 2011; Liu et al., 2012). The push for longer reads and higher throughput is a war of numbers however that often occurs when corporate entities compete. There needs to be more thought given to the purpose of the sequencing and the complexities involved in its analysis. More data is of no use of it is error prone and requires greater computational resources to process (Brenner, 1999; Devos and Valencia, 2001; Huttenhower and Hofmann, 2010; Hubisz et al., 2011). Less data of a higher quality would greatly improve bottlenecks in analyses downstream. Ultimately, a focus on developing and improving single molecule DNA sequencing is likely to be the most appropriate and productive approach towards overcoming these issues (Pettersson, 2009; Clarke et al., 2009; Pareek et al., 2011; Liu et al., 2012; Schneider and Dekker, 2012; McCoy et al., 2014). A focus on the hardware requirements is also necessary. Chaining together computing resources into vast server farms allows for a great deal of computational power to be harnessed for scientific analyses (Cuff et al., 2004; Bader et al., 2005; Schatz et al., 2010; Fusaro, 2011; Niemenmaa et al., 2012; Krampis et al., 2012). However, the financial and environmental impacts of these facilities make them unrealistic for use in the longer term (Carroll et al., 2011; Honee et al., 2012; Doyle and O'Mahony, 2014; Seegolam and Usmani, 2014). There needs to be a focus on developing more efficient, environmentally friendly and powerful hardware for the analyses of tomorrow’s data. This includes using differing architectures such as ARM (Balakrishnan, 2012; Rajovic et al., 2013a; Rajovic et al., 2013b; Rajovic et al., 2014) in the first instance, but there is also a focus on completely replacing semiconductor-based computing with natural computation via DNA computers or quantum computers for example (Ezziane, 2006; Castro, 2007; Kari and Rozenberg, 2008). Steps have already been taken towards storing biological data using DNA (Goldman et al., 2013). This will have profound implications on how we process and analyse biological data, particularly in large-scale comparative genomics.

Interdisciplinary nature of genome informatics

The amount of data produced by genome projects inevitably requires a multifaceted approach. The realisation that research is an interdisciplinary process that requires the integration of a large number of skilled professionals, not just academics, will greatly improve the outcomes of scientific endeavours. It is imperative that different disciplines integrate and work together to develop new infrastructure, procedures, and protocols to ensure the efficient and reliable processing of scientific data. Scientists desperately needs to see the bigger picture, collaborate more effectively and help to progress the development and adoption of emerging technologies if it is going to see sustained growth. This is particularly true in comparative genomics as the amount of data we see being produced by next generation sequencing technologies is set to rise to over 4000% by 2030 (Baker et al., 2010). The challenges facing the processing and analyses of this volume of data will make the current “big data” climate seem insignificant in comparison (Huttenhower and Hofmann, 2010; Schatz et al., 2010; Gross, 2011; Crosswell and Thornton, 2012; Michener and Jones, 2012; Hampton et al., 2013). This synthesis of disciplines needs to see biotechnologists and chemists working to improve the quality of sequencing; computer scientists writing more optimal algorithms and tools for analyses, alongside improving the distribution of raw data via the Semantic Web; electronics engineers and physicists working on producing more efficient and effective computer hardware; statisticians and mathematicians working on more robust means of highlighting meaningful information from these volumes of data; and of course biologists to apply their understanding of the biological systems throughout. Most importantly there needs to be more communication and interaction between all these groups (Hambrusch et al., 2009).

There also needs to be a change in the way that biology is taught from the ground up (Ranganathan, 2005; Qin, 2009; Schneider et al., 2012). Data science, as a label for individuals that can apply expert knowledge of computer science, mathematics and statistics, and substantive domain specific expertise to the processing and analytics of “big data”, is a growing trend. There is disagreement between the disciplines on what actually constitutes a “data scientist”, but regardless, what it represents is a synthesis of disciplines with a common goal. Various courses are appearing online and at academic institutions that are taking this multifaceted approach in training a new genre of “data scientists” to tackle the looming problems (Christensen et al., 2013; Waldrop, 2014). This needs to be the approach taken in biology too. A more interdisciplinary curriculum is required, with modules taught across academic disciplines (Hambrusch et al., 2009), in order to prepare biologists for future data analysis requirements. With data at the scale of exabytes or zettabytes not being uncommon within the next 15 years, this change cannot come too soon.

[This page is left intentionally blank]

REFERENCES

Acquaah, G. (2012) Principles of Plant Genetics and Breeding. Chichester, UK: John Wiley & Sons.

Adams MD, Rudner DZ and Rio DC. 1996. Biochemistry and regulation of pre-mRNA splicing. Current opinion in cell biology. 8:331-9.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., et al. (2000) The genome sequence of Drosophila melanogaster. Science. 287 (5461), 2185–2195.

Aldrup-MacDonald, M. & Sullivan, B. (2014) The Past, Present, and Future of Human Centromere Genomics. Genes. 5 (1), 33–50.

Alföldi, J. & Lindblad-Toh, K. (2013) Comparative genomics as a tool to understand evolution and disease. Genome research. 23 (7), 1063–1068.

Alkan, C., Sajjadian, S. & Eichler, E.E. (2011) Limitations of next-generation genome sequence assembly. Nature Methods. 8 (1), 61–65.

Altschul, S.F. & Lipman, D.J. (1990) Protein database searches for multiple alignments. Proceedings of the National Academy of Sciences. 87 (14), 5509–5513.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. Journal of Molecular Biology. 215 (3), 403–410.

Ames, R.M., Money, D., Ghatge, V.P., Whelan, S. & Lovell, S.C. (2012) Determining the evolutionary history of gene families. Bioinformatics (Oxford, England). 28 (1), 48–55.

Ames, R.M., Rash, B.M., Hentges, K.E., Robertson, D.L., Delneri, D. & Lovell, S.C. (2010) Gene duplication and environmental adaptation within yeast populations. Genome biology and evolution. 2 (0), 591–601.

Amid, C., Rehaume, L.M., Brown, K.L., Gilbert, J.G.R., Dougan, G., Hancock, R.E.W. & Harrow, J.L. (2009) Manual annotation and analysis of the defensin gene cluster in the C57BL/6J mouse reference genome. BMC Genomics. 10 (1), 606.

and Analysis Consortium, T.C.S. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 437 (7055), 69–87.

Antezana, E., Egaña, M., Blondé, W. & Illarramendi, A. (2009) The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process. Genome

Aparicio S et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science (New York, N.Y.). 297:1301-10.

Appasani, K. (2012) Epigenomics: From Chromatin Biology to Therapeutics.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Research. 29 (1), 37–40.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. & Yeh, L.-S.L. (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Research. 32 (Database issue), D115–9.

Arnold, M.L., Ballerini, E.S., Brothers, A.N., Hamlin, J.A.P., Ishibashi, C.D.A. & Zuellig, M.P. (2012) The genomics of natural selection and adaptation: Christmas past, present and future(?). Plant Ecology & Diversity. 5 (4), 451–456.

Arrais, J.P. & Oliveira, J.L. (2010) On the exploitation of cloud computing in bioinformatics. Audio, Transactions of the IRE Professional Group on. 1–4.

Ashburner, M. (2005) 'Ontologies for Biologists - A Community Model for the Annotation of Genomic Data', in Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE. [Online]. 1 January 2005 IEEE. p. 7.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., et al. (2000) Gene Ontology: tool for the unification of biology. Nature Genetics. 25 (1), 25–29.

Ashurst, J.L., Chen, C.-K., Gilbert, J.G.R., Jekosch, K., Keenan, S., Meidl, P., Searle, S.M., Stalker, J., Storey, R., Trevanion, S., Wilming, L. & Hubbard, T. (2005) The Vertebrate Genome Annotation (Vega) database. Nucleic Acids Research. 33 (Database issue), D459–65.

Assis, R. & Bachtrog, D. (2013) Neofunctionalization of young duplicate genes in Drosophila. Proceedings of the National Academy of Sciences of the United States of America. 110 (43), 17409–17414.

Assis, R. & Kondrashov, A.S. (2012) Nonallelic gene conversion is not GC-biased in Drosophila or primates. Molecular biology and evolution. 29 (5), 1291–1295.

Authors, T.I.D.R.W., Birney, E., Hudson, T.J., Green, E.D., Gunter, C., Eddy, S., Rogers, J., Harris, J.R., Ehrlich, S.D., Apweiler, R., Austin, C.P., Berglund, L., Bobrow, M., Bountra, C., Brookes, A.J., Cambon-Thomsen, A., Carter, N.P., Chisholm, R.L., Contreras, J.L., et al. (2009) Prepublication data sharing. Nature 461 (7261), 168–170.

Babenko, V.N., Rogozin, I.B., Mekhedov, S.L. & Koonin, E. V (2004) Prevalence of intron gain over intron loss in the evolution of paralogous gene families. Nucleic Acids Research. 32 (12), 3724–3733.

Bachtrog, D. (2013) Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nature Reviews Genetics. 14 (2), 113–124.

Bailey, J.A. (2002) Recent segmental duplications in the human genome. Science. 297 (5583), 1003–1007.

Baker, M. (2010) Next-generation sequencing: adjusting to data overload. Nature Methods. 7 (7), 495–499.

Baker, M. (2012) Quantitative data: learning to share. Nature Methods. 9 (1), 39–41.

Balakrishnan, N. (2012) Building and benchmarking a low power ARM cluster. Master's thesis.

Ballester, B., Johnson, N., Proctor, G. & Flicek, P. (2010) Consistent annotation of gene expression arrays. BMC Genomics. 11 (1), 294.

Bansal, M.S. & Eulenstein, O. (2008) An Omega(n2/ log n) speed-up of TBR heuristics for the gene-duplication problem. Computational Biology and Bioinformatics, IEEE/ACM Transactions on. 5 (4), 514–524.

Bao, Z. & Eddy, S.R. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome research. 12 (8), 1269–1276.

Barabási, A.-L. & Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. Nature Reviews Genetics. 5 (2), 101–113.

Barsh, G.S. & Andersson, L. (2013) Evolutionary genomics: Detecting selection. Nature. 495 (7441), 325–326.

Barsnes, H., Vizcaíno, J.A., Eidhammer, I. & Martens, L. (2009) PRIDE Converter: making proteomics data-sharing easy. Nat Biotech. 27 (7), 598–599.

Basenko, E., Topcu, Z. & McEachern, M.J. (2011) Recombination can either help maintain very short telomeres or generate longer telomeres in yeast cells with weak telomerase activity. Eukaryotic cell. 10 (8), 1131–1142.

Basu, M.K., Makalowski, W., Rogozin, I.B. & Koonin, E. V (2008) U12 intron positions are more strongly conserved between animals and plants than U2 intron positions. Biol Direct. 3 (1), 19.

Basu, M.K., Rogozin, I.B., Deusch, O., Dagan, T., Martin, W. & Koonin, E. V (2008) Evolutionary dynamics of introns in plastid-derived genes in plants: saturation nearly reached but slow intron gain continues. Molecular biology and evolution. 25 (1), 111–119.

Bateman, A. & Wood, M. (2009) Cloud computing. Bioinformatics (Oxford, England). 25 (12), 1475.

Beerli, P. (2006) Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics. 22 (3), 341–345.

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.-P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., Pachter, L., Myers, E. & Langley, C.H. (2007) Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. PLoS Biology. 5 (11), e310.

Belle, E.M.S., Duret, L., Galtier, N. & Eyre-Walker, A. (2004) The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. Journal of Molecular Evolution. 58 (6), 653–660.

BENNETT, M.D. (2004) Perspectives on polyploidy in plants - ancient and neo. Biological Journal of the Linnean Society. 82 (4), 411–423.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research. 27:573-580.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. (2005) GenBank. Nucleic Acids Research. 33 (Database issue), D34–8.

Benton MJ and Donoghue PCJ. 2007. Paleontological evidence to date the tree of life. Molecular biology and evolution. 24:26-53.

Berges, M., Hubwieser, P., Magenheim, J., Bender, E., Bröker, K., Margaritis-Kopecki, M., Neugebauer, J., Schaper, N., Schubert, S. & Ohrndorf, L. (2013) Developing a competency model for teaching computer science in schools. ITiCSE ’13: Proceedings of the 18th ACM conference on Innovation and technology in computer science education p.327.

Bergman CM and Quesneville H. 2007. Discovering and detecting transposable elements in genome sequences. Briefings in bioinformatics. 8:382-92.

Bernardi, G. (2012) The genome: an isochore ensemble and its evolution. Annals of the New York Academy of Sciences. 1267 (1), 31–34.

Berners-Lee, T. & Hendler, J. (2001a) Publishing on the semantic web. Nature

Berners-Lee, T. & Hendler, J. (2001b) The semantic web. Scientific ….

Berners-Lee, T., Connolly, D. & Swick, R.R. (1999) Web architecture: Describing and exchanging data. WWW-address: http://www w3 org/ ….

Berners-Lee, T., Fielding, R. & Frystyk, H. (1996) Hypertext transfer protocol--HTTP/1.0.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001) The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. Journal of Molecular Biology. 112 (3), 535–542.

Bianco, K., Caughey, A.B., Shaffer, B.L., Davis, R. & Norton, M.E. (2006) History of Miscarriage and Increased Incidence of Fetal Aneuploidy in Subsequent Pregnancy. Obstetrics & Gynecology. 107 (5), 1098–1102.

Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J.M., Mei, R., Scherer, S.W., Julian, C.G., Wilson, M.J., López Herráez, D., Brutsaert, T., Parra, E.J., Moore, L.G. & Shriver, M.D. (2010) Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data David J Begun (ed.). PLoS genetics. 6 (9), e1001116.

Bilofsky, H.S., Burks, C., Fickett, J.W., Goad, W.B., Lewitter, F.I., Rindone, W.P., Swindell, C.D. & Tung, C.S. (1986) The GenBank genetic sequence databank. Nucleic Acids Research. 14 (1), 1–4.

Birney, E., Andrews, T.D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyras, E., Fernández-Suárez, X.M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H.-R., Iyer, V., et al. (2004) An overview of Ensembl. Genome research. 14 (5), 925–928.

Blomberg, S.P. & Garland, T. (2002) Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. Journal of Evolutionary Biology. 15 (6), 899–910.

Blumenstiel, J.P., Hartl, D.L. & Lozovsky, E.R. (2002) Patterns of insertion and deletion in contrasting chromatin domains. Molecular biology and evolution. 19 (12), 2211–2225.

Bodenreider, O. & Stevens, R. (2006) Bio-ontologies: current trends and future directions. Briefings in bioinformatics. 7 (3), 256–274.

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

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F. & Corbeil, J. (2012) Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. 13 (12), R122.

Bon, E., Casaregola, S., Blandin, G., Llorente, B., Neuvéglise, C., Munsterkotter, M., Guldener, U., Mewes, H.-W., Van Helden, J., Dujon, B. & Gaillardin, C. (2003) Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. Nucleic Acids Research. 31 (4), 1121–1135.

Bonasio, R., Zhang, G., Ye, C., Mutti, N.S., Fang, X., Qin, N., Donahue, G., Yang, P., Li, Q., Li, C., Zhang, P., Huang, Z., Berger, S.L., Reinberg, D., Wang, J. & Liebig, J. (2010) Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator. Science. 329 (5995), 1068–1071.

Bornberg-Bauer, E., Huylmans, A.-K. & Sikosek, T. (2010) How do new proteins arise? Current opinion in structural biology. 20 (3), 390–396.

Bosco G, Campbell P, Leiva-Neto JT and Markow TA. 2007. Analysis of Drosophila species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics. 177:1277-90.

Bouck, J., Miller, W., Gorrell, J.H., Muzny, D. & Gibbs, R.A. (1998) Analysis of the quality and utility of random shotgun sequencing at low redundancies. Genome research. 8 (10), 1074–1084.

Boulesteix M, Weiss M and Biémont C. 2006. Differences in genome size between closely related species: the Drosophila melanogaster species subgroup. Molecular biology and evolution. 23:162-7.

Boyko, A.R. (2011) The domestic dog: man's best friend in the genomic era. Genome Biol. 12 (2), 216.

Bradnam, K.R. & Korf, I. (2008) Longer First Introns Are a General Property of Eukaryotic Gene Structure Alan Christoffels (ed.). PLoS One. 3 (8), e3093.

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, İ., 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., et al. (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Giga Science. 2 (1), 10.

Brenner, S.E. (1999) Errors in genome annotation. Trends in Genetics: TIG. 15 (4), 132–133.

Britten, R.J., Rowen, L., Williams, J. & Cameron, R.A. (2003) Majority of divergence between closely related DNA samples is due to indels. Proceedings of the National Academy of Sciences. 100 (8), 4661–4665.

Bromham, L. (2009) Why do species vary in their rate of molecular evolution? Biology Letters. 5 (3), 401–404.

Brooksbank, C., Cameron, G. & Thornton, J. (2010) The European Bioinformatics Institute’s data resources. Nucleic Acids Research. 38 (Database issue), D17–25.

Brown, T.A. (2006) Genomes 3. Garland Science.

Brunner, P.C., Torriani, S.F.F., Croll, D., Stukenbrock, E.H. & McDonald, B.A. (2014) Hitchhiking selection is driving intron gain in a pathogenic fungus. Molecular biology and evolution. 31 (7), 1741–1749.

Burnette JM, Miyamoto-Sato E, Schaub MA, Conklin J and Lopez AJ. 2005. Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. Genetics. 170:661-74.

Butler, M.A. & King, A.A. (2004) Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. The American naturalist. 164 (6), 683–695.

Buttazzo, G. (2000) 'Can a Machine Ever Become Self-aware?', in R. Aurich et al. (eds.) Artificial Humans, an Historical Retrospective of the Berlin International Film Festival. [Online]. Artificial Humans. pp. 45–49.

Cai, J., Zhao, R., Jiang, H. & Wang, W. (2008) De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics. 179 (1), 487–496.

Campos, R., Storz, J.F. & Ferrand, N. (2012) Copy number polymorphism in the α-globin gene cluster of European rabbit (Oryctolagus cuniculus). Heredity. 108 (5), 531–536.

Canceill, D., Viguera, E. & Ehrlich, S.D. (1999) Replication slippage of different DNA polymerases is inversely related to their strand displacement efficiency. J. Biol. Chem. 274 (39), 27481–27490.

Carmel, L., Rogozin, I.B., Wolf, Y.I. & Koonin, E. V (2007) Evolutionarily conserved genes preferentially accumulate introns. Genome research. 17 (7), 1045–1050.

Carneiro, M.O., Russ, C., Ross, M.G., Gabriel, S.B., Nusbaum, C. & DePristo, M.A. (2012) Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics. 13 (1), 375.

Carp, H. (2001) Karyotype of the abortus in recurrent miscarriage. Fertility and sterility. 75 (4), 678–682.

Carroll, R., Balasubramaniam, S., Botvich, D. & Donnelly, W. (2011) 'Dynamic Optimization Solution for Green Service Migration in Data Centres', in Communications (ICC), 2011 IEEE International Conference on. [Online]. 1 January 2011 IEEE. pp. 1–6.

Carvalho, A.B. & Clark, A.G. (1999) Genetic recombination: Intron size and natural selection. Nature. 401 (6751), 344.

Castillo-Davis, C.I., Bedford, T.B.C. & Hartl, D.L. (2004) Accelerated rates of intron gain/loss and protein evolution in duplicate genes in human and mouse malaria parasites. Molecular biology and evolution. 21 (7), 1422–1427.

Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E. V & Kondrashov, F.A. (2002) Selection for short introns in highly expressed genes. Nature genetics. 31 (4), 415–418.

Ceci, S.J. (1988) Scientists' attitudes toward data sharing. Science.

Chain, F.J.J. & Evans, B.J. (2006) Multiple Mechanisms Promote the Retained Expression of Gene Duplicates in the Tetraploid Frog Xenopus laevis. PLoS genetics. 2 (4), e56.

Chain, P., Kurtz, S., Ohlebusch, E. & Slezak, T. (2003) An applications-focused review of comparative genomics tools: Capabilities, limitations and future challenges. Briefings in bioinformatics. 4 (2), 105–123.

Chamary J-V and Hurst LD. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. Molecular biology and evolution. 21:1014-23.

Chao, D.-Y., Dilkes, B., Luo, H., Douglas, A., Yakubova, E., Lahner, B. & Salt, D.E. (2013) Polyploids exhibit higher potassium uptake and salinity tolerance in Arabidopsis. Science. 341 (6146), 658–659.

Charlesworth, B. (2009) Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics. 10 (3), 195–205.

Charlesworth, B. & Charlesworth, D. (2000) The degeneration of Y chromosomes. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. 355 (1403), 1563–1572.

Charlesworth, D. & Wright, S.I. (2001) Breeding systems and genome evolution. Current Opinion in Genetics & Development. 11 (6), 685–690.

Chassagnole, C., Rodriguez, J.C.A., Doncescu, A. & Yang, L.T. (2005) Parallel Computing for Bioinformatics and Computational Biology. Albert Y Zomaya (ed.). Hoboken, NJ, USA: John Wiley & Sons.

Chatterji, S. & Pachter, L. (2007) Patterns of gene duplication and intron loss in the ENCODE regions suggest a confounding factor. Genomics. 90 (1), 44–48.

Chaurasia, A., Tarallo, A., Bernà, L., Yagi, M., Agnisola, C. & D’Onofrio, G. (2014) Length and GC Content Variability of Introns among Teleostean Genomes in the Light of the Metabolic Rate Hypothesis Igor B Rogozin (ed.). PLoS One. 9 (8), e103889.

Chen, D., Lin, Y. & Zhang, H. (2008) Characterization and expression of two amphioxus DDAH genes originating from an amphioxus-specific gene duplication. Gene. 410 (1), 75–81.

Chen, F.-C., Chen, C.-J., Li, W.-H. & Chuang, T.-J. (2010) Gene family size conservation is a good indicator of evolutionary rates. Molecular biology and evolution. 27 (8), 1750–1758.

Chen, F., Mackey, A.J., Vermunt, J.K. & Roos, D.S. (2007) Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes Cecile Fairhead (ed.). PLoS One. 2 (4), e383.

Chen, J.M., Cooper, D.N., Chuzhanova, N. & Férec, C. (2007) Gene conversion: mechanisms, evolution and human disease. Nature Reviews ….

Chen, S., Gomes, R., Costa, V., Santos, P., Charneca, R., Zhang, Y., Liu, X., Wang, S., Bento, P., Nunes, J.-L., Buzgó, J., Varga, G., Anton, I., Zsolnai, A. & Beja-Pereira, A. (2013) How immunogenetically different are domestic pigs from wild boars: a perspective from single-nucleotide polymorphisms of 19 immunity-related candidate genes. Immunogenetics. 65 (10), 737–748.

Chen, Y., Cunningham, F., Rios, D., McLaren, W.M., Smith, J., Pritchard, B., Spudich, G.M., Brent, S., Kulesha, E., Marin-Garcia, P., Smedley, D., Birney, E. & Flicek, P. (2010) Ensembl variation resources. BMC Genomics. 11 (1), 293.

Chen, Z., Cheng, C.-H.C., Zhang, J., Cao, L., Chen, L., Zhou, L., Jin, Y., Ye, H., Deng, C., Dai, Z., Xu, Q., Hu, P., Sun, S., Shen, Y. & Chen, L. (2008) Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. Proceedings of the National Academy of Sciences of the United States of America. 105 (35), 12944–12949.

Chi, L.M. & Lam, S.L. (2005) Structural roles of CTG repeats in slippage expansion during DNA replication. Nucleic Acids Research. 33 (5), 1604–1617.

Choo, K.H. (1998) Why is the centromere so cold? Genome research. 8 (2), 81–82.

Chor, B. & Tuller, T. (2005) Maximum likelihood of evolutionary trees: hardness and approximation. Bioinformatics. 21 (suppl 1), i97–i106.

Chowdhury, S., Dent, T., Pashayan, N., Hall, A., Lyratzopoulos, G., Hallowell, N., Hall, P., Pharoah, P. & Burton, H. (2013) Incorporating genomics into breast and prostate cancer screening: assessing the implications. Genetics in Medicine. 15 (6), 423–432.

Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D. & Emanuel, E.J. (2013) The MOOC Phenomenon: Who Takes Massive Open Online Courses and Why? SSRN Electronic Journal.

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Hubbard, T., Kasprzyk, A., Keefe, D., Lehvaslaiho, H., et al. (2003) Ensembl 2002: accommodating comparative genomics. Nucleic Acids Research. 31 (1), 38–42.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., Pollard, D.A., Sackton, T.B., Larracuente, A.M., Singh, N.D., Abad, J.P., Abt, D.N., Adryan, B., Aguade, M., Akashi, H., et al. (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature. 450 (7167), 203–218.

Clark, S., Egan, R., Frazier, P.I. & Wang, Z. (2013) ALE: a Generic Assembly Likelihood Evaluation Framework for Assessing the Accuracy of Genome and Metagenome Assemblies. Bioinformatics. 29 (4), 435–443.

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

Coghlan A and Wolfe KH. 2004. Origins of recently gained introns in Caenorhabditis. Proceedings of the National Academy of Sciences of the United States of America. 101:11362-7.

Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H., Tokishita, S., Aerts, A., Arnold, G.J., Basu, M.K., Bauer, D.J., Cáceres, C.E., Carmel, L., Casola, C., Choi, J.-H., Detter, J.C., Dong, Q., Dusheyko, S., Eads, B.D., et al. (2011) The ecoresponsive genome of Daphnia pulex. Science (New York, N.Y.). 331 (6017), 555–561.

Comai, L. (2005) The advantages and disadvantages of being polyploid. Nature Reviews Genetics. 6 (11), 836–846.

Comeron, J.M. & Kreitman, M. (2000) The correlation between intron length and recombination in drosophila. Dynamic equilibrium between mutational and selective forces. Genetics. 156 (3), 1175–1190.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. & Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21 (18), 3674–3676.

Connallon, T. & Clark, A.G. (2010) Gene duplication, gene conversion and the evolution of the Y chromosome. Genetics. 186 (1), 277–286.

Conrad, B. & Antonarakis, S.E. (2007) Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease. dx.doi.org. 8 (1), 17–35.

Consortium, C. elegans S. & C elegans Sequencing Consortium, T. (1998) Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science. 282 (5396), 2012–2018.

Consortium, G.O. (2001) Creating the gene ontology resource: design and implementation. Genome research. 11 (8), 1425–1433.

Consortium, G.O., Blake, J.A., Dolan, M., Drabkin, H., Hill, D.P., Li, N., Sitnikov, D., Bridges, S., Burgess, S., Buza, T., McCarthy, F., Peddinti, D., Pillai, L., Carbon, S., Dietze, H., Ireland, A., Lewis, S.E., Mungall, C.J., Gaudet, P., et al. (2013) Gene Ontology annotations and resources. Nucleic Acids Research. 41 (Database issue), D530–5.

Consortium, I.C.G.S., Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A.M., Delany, M.E., Dodgson, J.B., Genome fingerprint map, sequence, assembly, Chinwalla, A.T., Cliften, P.F., Clifton, S.W., Delehaunty, K.D., Fronick, C., et al. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 432 (7018), 695–716.

Consortium, I.H.G.S., Collins, F.S., Lander, E.S., Rogers, J., Waterston, R.H. & Conso, I. (2004) Finishing the euchromatic sequence of the human genome. Nature 431 (7011), 931–945.

Consortium, M.G., Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature. 420 (6915), 520–562.

Consortium, R.M.G.S. and A., Gibbs, R.A., Rogers, J., Katze, M.G., Bumgarner, R., Weinstock, G.M., Mardis, E.R., Remington, K.A., Strausberg, R.L., Venter, J.C., Wilson, R.K., Batzer, M.A., Bustamante, C.D., Eichler, E.E., Hahn, M.W., Hardison, R.C., Makova, K.D., Miller, W., Milosavljevic, A., et al. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science. 316 (5822), 222–234.

Consortium, S.U.G.S., Sodergren, E., Weinstock, G.M., Davidson, E.H., Cameron, R.A., Gibbs, R.A., Angerer, R.C., Angerer, L.M., Arnone, M.I., Burgess, D.R., Burke, R.D., Coffman, J.A., Dean, M., Elphick, M.R., Ettensohn, C.A., Foltz, K.R., Hamdoun, A., Hynes, R.O., Klein, W.H., et al. (2006) The genome of the sea urchin Strongylocentrotus purpuratus. Science (New York, N.Y.). 314 (5801), 941–952.

Consortium, T. 1000 G.P. (2011) A map of human genome variation from population-scale sequencing. Nature. 473 (7348), 544.

Consortium, T. 1000 G.P., author, C., committee, S., Medicine, P. group B.C. of, BGI-Shenzhen, Harvard, B.I. of M.I.T. and, Illumina, Technologies, L., Genetics, M.P.I. for M., Science, R.A., Louis, W.U. in S., Institute, W.T.S., Technologies, A. group A., Medicine, B.C. of, College, B., Hospital, B. and W., Cardiff University, T.H.G.M.D., Laboratory, C.S.H., Universities, C. and S., et al. (2010) A map of human genome variation from population-scale sequencing. Nature. 467 (7319), 1061–1073.

Cook, S.A. (1971) 'The complexity of theorem-proving procedures', in the third annual ACM symposium. [Online]. 1 January 1971 New York, New York, USA: SIGACT, ACM Special Interest Group on Algorithms and Computation Theory. pp. 151–158.

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P.D., Grützner, F. & Kaessmann, H. (2014) Origins and functional evolution of Y chromosomes across mammals. Nature. 508 (7497), 488–493.

Costantini, M. & Bernardi, G. (2008) The short-sequence designs of isochores from the human genome. Proceedings of the National Academy of Sciences of the United States of America. 105 (37), 13971–13976.

Costantini, M., Alvarez-Valin, F., Costantini, S., Cammarano, R. & Bernardi, G. (2013) Compositional patterns in the genomes of unicellular eukaryotes. BMC Genomics. 14 (1), 755.

Courteau, J. (1991) Genome databases. Science. 254 (5029), 201–207.

Crick, F. (1979) Split genes and RNA splicing. Science. 204 (4390), 264–271.

Crosswell, L.C. & Thornton, J.M. (2012) ELIXIR: a distributed infrastructure for European biological data. Trends in Biotechnology. 30 (5), 241–242.

Crusoe, M.R. & Brown, C.T. (2013) Walking the talk: adopting and adapting sustainable scientific software development processes in a small biology lab.

Cuff, J.A. (2004) The Ensembl Computing Architecture. Genome research. 14 (5), 971–975.

Cui, L. (2006) Widespread genome duplications throughout the history of flowering plants. Genome research. 16 (6), 738–749.

Curwen, V., Eyras, E., Andrews, T.D., Clarke, L., Mongin, E., Searle, S.M.J. & Clamp, M. (2004) The Ensembl automatic gene annotation system. Genome research. 14 (5), 942–950.

Dai, L., Gao, X., Guo, Y., Xiao, J. & Zhang, Z. (2012) Bioinformatics clouds for big data manipulation. Biol Direct. 7 (1), 43–discussion 43.

De Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. (2006) CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 22 (10), 1269–1271.

De Koning, A.P.J., Gu, W., Castoe, T.A., Batzer, M.A. & Pollock, D.D. (2011) Repetitive Elements May Comprise Over Two-Thirds of the Human Genome Gregory P Copenhaver (ed.). PLoS genetics. 7 (12), e1002384.

Deaton, A.M. & Bird, A. (2011) CpG islands and the regulation of transcription. Genes & Development. 25 (10), 1010–1022.

DECASTRO, L. (2007) Fundamentals of natural computing: an overview. Physics of Life Reviews. 4 (1), 1–36.

Delihas, N. (2011) Impact of small repeat sequences on bacterial genome evolution. Genome biology and evolution. 3 (0), 959–973.

Demuth, J.P. & Hahn, M.W. (2009) The life and death of gene families. Bioessays. 31 (1), 29–39.

Demuth, J.P., De Bie, T., Stajich, J.E., Cristianini, N. & Hahn, M.W. (2006) The Evolution of Mammalian Gene Families Justin Borevitz (ed.). PLoS One. 1 (1), e85.

Derr, L.K. & Strathern, J.N. (1993) A role for reverse transcripts in gene conversion. Nature. 361 (6408), 170–173.

Deutsch M and Long M. 1999. Intron-exon structures of eukaryotic model organisms. Nucleic acids research. 27:3219-28.

Deutsch, M. & Long, M. (1999) Intron-exon structures of eukaryotic model organisms. Nucleic Acids Research. 27 (15), 3219–3228.

Devos, D. & Valencia, A. (2001) Intrinsic errors in genome annotation. Trends in Genetics. 17 (8), 429–431.

Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J.C., Hernandez-Boussard, T., Rees, C.A., Cherry, J.M., Botstein, D., Brown, P.O. & Alizadeh, A.A. (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Research. 31 (1), 219–223.

Dieterich, C., Clifton, S.W., Schuster, L.N., Chinwalla, A., Delehaunty, K., Dinkelacker, I., Fulton, L., Fulton, R., Godfrey, J., Minx, P., Mitreva, M., Roeseler, W., Tian, H., Witte, H., Yang, S.-P., Wilson, R.K. & Sommer, R.J. (2008) The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism. Nature genetics. 40 (10), 1193–1198.

Doolittle, W.F. (2013) Is junk DNA bunk? A critique of ENCODE. Proceedings of the National Academy of Sciences of the United States of America. 110 (14), 5294–5300.

Dopman, E.B. & Hartl, D.L. (2007) A portrait of copy-number polymorphism in Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America. 104 (50), 19920–19925.

Dowell, R.D., Jokerst, R.M., Day, A., Eddy, S.R. & Stein, L. (2001) The Distributed Annotation System. BMC bioinformatics. 2 (1), 7.

Doyle, J. & O'Mahony, D. (2014) 'Nihil: Computing Clouds with Zero Emissions', in Cloud Engineering (IC2E), 2014 IEEE International Conference on. [Online]. 1 January 2014 IEEE. pp. 331–336.

Du Plessis, L., Skunca, N. & Dessimoz, C. (2011) The what, where, how and why of gene ontology--a primer for bioinformaticians. Briefings in bioinformatics. 12 (6), 723–735.

Dubois, P.F. (2005) Maintaining correctness in scientific programs. Computing in Science & Engineering. 7 (3), 80–85.

Dudley, J.T. & Butte, A.J. (2009) A quick guide for developing effective bioinformatics programming skills Fran Lewitter (ed.). PLoS Computational Biology. 5 (12), e1000589.

Dudley, J.T., Pouliot, Y., Chen, R., Morgan, A.A. & Butte, A.J. (2010) Translational bioinformatics in the cloud: an affordable alternative. Genome Medicine. 2 (8), 51.

Dufresne, F. & Jeffery, N. (2011) A guided tour of large genome size in animals: what we know and where we are heading. Chromosome Research. 19 (7), 925–938.

Dumas, L., Kim, Y.H., Karimpour-Fard, A., Cox, M., Hopkins, J., Pollack, J.R. & Sikela, J.M. (2007) Gene copy number variation spanning 60 million years of human and primate evolution. Genome research. 17 (9), 1266–1277.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B.R., Landt, S.G., Lee, B.-K., Pauli, F., Rosenbloom, K.R., Sabo, P., Safi, A., Sanyal, A., et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489 p.57–74.

Duret, L. (2001) Why do genes have introns? Recombination might add a new piece to the puzzle. Trends in Genetics. 17 (4), 172–175.

Duret, L. & Galtier, N. (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. Annual Review of Genomics and Human Genetics. 10 (1), 285–311.

Duret, L. & Hurst, L.D. (2001) The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. Molecular biology and evolution. 18 (5), 757–762.

Duret, L., Eyre-Walker, A. & Galtier, N. (2006) A new perspective on isochore evolution. Gene. 38571–74.

Duret, L., Mouchiroud, D. & Gautier, C. (1995) Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. Journal of Molecular Evolution. 40 (3), 308–317.

Duret, L., Sémon, M., Piganeau, G., Mouchiroud, D. & Galtier, N. (2002) Vanishing GC-rich isochores in mammalian genomes. Genetics. 162 (4), 1837–1847.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. & Huber, W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 21 (16), 3439–3440.

Durinck, S., Spellman, P.T., Birney, E. & Huber, W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature protocols. 4 (8), 1184–1191.

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H.O.K., Buffalo, V., Zerbino, D.R., Diekhans, M., Nguyen, N., Ariyaratne, P.N., Sung, W.-K.K., Ning, Z., Haimel, M., Simpson, J.T., Fonseca, N.A., Birol, İ., Docking, T.R., et al. (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome research. 21 (12), 2224–2241.

Eastman, J.M., Alfaro, M.E., Joyce, P., Hipp, A.L. & Harmon, L.J. (2011) A NOVEL COMPARATIVE METHOD FOR IDENTIFYING SHIFTS IN THE RATE OF CHARACTER EVOLUTION ON TREES. Evolution. 65 (12), 3578–3589.

Eddy, S.R. (1998) Profile hidden Markov models. Bioinformatics. 14 (9), 755–763.

Eddy, S.R. (2012) The C-value paradox, junk DNA and ENCODE. Current Biology. 22 (21), R898–9.

Eddy, S.R. (2013) The ENCODE project: missteps overshadowing a success. Current biology : CB. 23 (7), R259–61.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics. 5 (1), 113.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 32 (5), 1792–1797.

Egan, C.M., Sridhar, S., Wigler, M. & Hall, I.M. (2007) Recurrent DNA copy number variation in the laboratory mouse. Nature Genetics.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., et al. (2009) Real-time DNA sequencing from single polymerase molecules. Science. 323 (5910), 133–138.

ELLEGREN, H. (2014) Genome sequencing and population genomics in non-model organisms. Trends in Ecology & Evolution. 29 (1), 51–63.

ELLEGREN, H., Smith, N.G.C. & Webster, M.T. (2003) Mutation rate variation in the mammalian genome. Current Opinion in Genetics & Development. 13 (6), 562–568.

EMBL Data Library and GenBank staff, T. (1987) A new system for direct submission of data to the nucleotide sequence data banks. Nucleic Acids Research. 15 (18), nil11–nil16.

Eng, C., Asthana, C., Aigle, B., Hergalant, S., Mari, J.-F. & Leblond, P. (2009) A New Data Mining Approach for the Detection of Bacterial Promoters Combining Stochastic and Combinatorial Methods. dx.doi.org. 16 (9), 1211–1225.

Enright, A.J., Van Dongen, S. & Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research. 30 (7), 1575–1584.

Esnault, C., Maestre, J. & Heidmann, T. (2000) Human LINE retrotransposons generate processed pseudogenes. Nature Genetics. 24 (4), 363–367.

Eyre-Walker, A. & Hurst, L.D. (2001) The evolution of isochores. Nature Reviews Genetics. 2 (7), 549–555.

Ezziane, Z. (2006) DNA computing: applications and challenges. Nanotechnology. 17 (2), R27–R39.

Falda, M., Toppo, S., Pescarolo, A., Lavezzo, E., Di Camillo, B., Facchinetti, A., Cilia, E., Velasco, R. & Fontana, P. (2012) Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. BMC bioinformatics. 13 Suppl 4 (Suppl 4), S14.

Fan, S., Elmer, K.R. & Meyer, A. (2012) Genomics of adaptation and speciation in cichlid fishes: recent advances and analyses in African and Neotropical lineages. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. 367 (1587), 385–394.

Farkash-Amar, S. & Simon, I. (2010) Genome-wide analysis of the replication program in mammals. Chromosome Research. 18 (1), 115–125.

Farrer, R.A., Henk, D.A., Garner, T.W.J., Balloux, F., Woodhams, D.C. & Fisher, M.C. (2013) Chromosomal copy number variation, selection and uneven rates of recombination reveal cryptic genome diversity linked to pathogenicity. Joseph Heitman (ed.). PLoS genetics. 9 (8), e1003703.

Fedorov, A., Merican, A.F. & Gilbert, W. (2002) Large-scale comparison of intron positions among animal, plant, and fungal genes. Proceedings of the National Academy of Sciences. 99 (25), 16128–16133.

Félix, M.-A. & Duveau, F. (2012) Population dynamics and habitat sharing of natural populations of Caenorhabditis elegans and C. briggsae. BMC Biology. 10 (1), 59.

Felsenstein, J. (1985) Phylogenies and the Comparative Method. The American naturalist. 125 (1), 1–15.

Fernald, G.H., Capriotti, E., Daneshjou, R., Karczewski, K.J. & Altman, R.B. (2011) Bioinformatics challenges for personalized medicine. Bioinformatics (Oxford, England). 27 (13), 1741–1748.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H. & Masinter, L. (1999) Hypertext transfer protocol–HTTP/1.1.

Fielding, R.T. (2000a) Architectural styles and the design of network-based software architectures.

Fielding, R.T. (2000b) Software architectural styles for network-based applications. Phase ii survey paper.

Fielding, R.T. & Taylor, R.N. (2000) Principled design of the modern Web architecture. International Conference on Software Engineering. Proceedings. 407–416.

Finotello, F., Lavezzo, E. & Fontana, P. (2011) Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. Briefings in ….

Fischer, I., Camus-Kulandaivelu, L., Allal, F. & Stephan, W. (2011) Adaptation to drought in two wild tomato species: the evolution of the Asr gene family. New Phytologist. 190 (4), 1032–1044.

Fisher, S.E. & Scharff, C. (2009) FOXP2 as a molecular window into speech and language. Trends in Genetics: TIG. 25 (4), 166–177.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. & al, et (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 269 (5223), 496–512.

Flicek P et al. 2010. Ensembl 2011. Nucleic acids research. 39:D800-806.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A.K., Keenan, S., et al. (2012) Ensembl 2013. Nucleic Acids Research. 41 (D1), gks1236–D55.

Flicek, P., Aken, B.L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Howe, K., Jenkinson, A., Johnson, N., et al. (2009) Ensembl’s 10th year. Nucleic Acids Research. 38 (Database), D557–D562.

Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S.C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Holland, R., et al. (2007) Ensembl 2008. Nucleic Acids Research. 36 (Database), D707–D714.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A.K., Keefe, D., Keenan, S., Kinsella, R., et al. (2011) Ensembl 2012. Nucleic Acids Research. 40 (Database issue), D84–D90.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., et al. (2010) Ensembl 2011. Nucleic Acids Research. 39 (Database), D800–D806.

Flynn, K.M., Vohr, S.H., Hatcher, P.J. & Cooper, V.S. (2010) Evolutionary rates and gene dispensability associate with replication timing in the archaeon Sulfolobus islandicus. Genome biology and evolution. 2 (0), 859–869.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L. & Postlethwait, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 151 (4), 1531–1545.

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., Hernandez-Boussard, T., Karimpour-Fard, A., Glueck, D., McGavran, L., Berry, R., Pollack, J. & Sikela, J.M. (2004) Lineage-Specific Gene Duplication and Loss in Human and Great Ape Evolution. PLoS Biology. 2 (7), e207.

Francino, M.P. (2005) An adaptive radiation model for the origin of new gene functions. Nature genetics. 37 (6), 573–578.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., Pasternak, S., Wheeler, D.A., Willis, T.D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature. 449 (7164), 851–861.

Freeman, J.L., Adeniyi, A., Banerjee, R., Dallaire, S., Maguire, S.F., Chi, J., Ng, B.L., Zepeda, C., Scott, C.E., Humphray, S., Rogers, J., Zhou, Y., Zon, L.I., Carter, N.P., Yang, F. & Lee, C. (2007) Definition of the zebrafish genome using flow cytometry and cytogenetic mapping. BMC Genomics. 8 (1), 195.

Frischmeyer PA and Dietz HC. 1999. Nonsense-mediated mRNA decay in health and disease. Human molecular genetics. 8:1893-900.

Fryxell, K.J. & Moon, W.-J. (2005) CpG mutation rates in the human genome are highly dependent on local GC content. Molecular biology and evolution. 22 (3), 650–658.

Fujita, M.K., Edwards, S. V & Ponting, C.P. (2011) The Anolis lizard genome: an amniote genome without isochores. Genome biology and evolution. 3 (0), 974–984.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T.R., Giardine, B.M., Harte, R.A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R.M., Learned, K., et al. (2010) The UCSC Genome Browser database: update 2011. Nucleic Acids Research. 39 (Database), D876–D882.

Fullerton, S.M., Bernardo Carvalho, A. & Clark, A.G. (2001) Local rates of recombination are positively correlated with GC content in the human genome. Molecular biology and evolution. 18 (6), 1139–1142.

Fusaro, V.A., Patil, P., Gafni, E., Wall, D.P. & Tonellato, P.J. (2011) Biomedical Cloud Computing With Amazon Web Services Fran Lewitter (ed.). PLoS Computational Biology. 7 (8), e1002147.

Gadagkar, R. (1997) The evolution of caste polymorphism in social insects: genetic release followed by diversifying evolution. Journal of Genetics. 76 (3), 167–179.

Gadau, J., Helmkampf, M., Nygaard, S., Roux, J. & al, et (2012) The genomic impact of 100 million years of social evolution in seven ant species. Trends in Genetics. 28 (1), 14–21.

Gaffney DJ and Keightley PD. 2006. Genomic selective constraints in murid noncoding DNA. PLoS genetics. 2:e204.

Galtier, N. (2003) Gene conversion drives GC content evolution in mammalian histones. Trends in Genetics. 19 (2), 65–68.

Galtier, N., Piganeau, G., Mouchiroud, D. & Duret, L. (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics. 159 (2), 907–911.

Gao, F. & Zhang, C.-T. (2006) Isochore structures in the chicken genome. The FEBS journal. 273 (8), 1637–1648.

Gao, F. & Zhang, C.-T. (2008) Prediction of replication time zones at single nucleotide resolution in the human genome. FEBS Letters. 582 (16), 2441–2444.

Gardiner-Garden, M. & Frommer, M. (1987) CpG islands in vertebrate genomes. Journal of Molecular Biology. 196 (2), 261–282.

Gazave E, Marqués-Bonet T, Fernando O, Charlesworth B and Navarro A. 2007. Patterns and rates of intron divergence between humans and chimpanzees. Genome biology. 8:R21.

Gazave, E., Darre, F., Morcillo-Suarez, C., Petit-Marty, N., Carreno, A., Marigorta, U.M., Ryder, O.A., Blancher, A., Rocchi, M., Bosch, E., Baker, C., Marques-Bonet, T., Eichler, E.E. & Navarro, A. (2011) Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. Genome research. 21 (10), 1626–1639.

Gentleman, R. (2005) Reproducible Research: A Bioinformatics Case Study. Statistical applications in genetics and molecular biology. 4 (1), .

Ghodsi, M., Hill, C.M., Astrovskaya, I., Lin, H., Sommer, D.D., Koren, S. & Pop, M. (2013) De novo likelihood-based measures for comparing genome assemblies. BMC Research Notes. 6 (1), 334.

Gibson, G. & Visscher, P.M. (2013) From personalized to public health genomics. Genome Medicine. 5 (7), 60.

Gilbert, D. (2003) Shopping in the genome market with EnsMart. Briefings in bioinformatics. 4 (3), 292–296.

Goble, C.A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P. & De Roure, D. (2010) myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic Acids Research. 38 (Web Server issue), W677–82.

Goidts, V., Cooper, D.N., Armengol, L., Schempp, W., Conroy, J., Estivill, X., Nowak, N., Hameister, H. & Kehrer-Sawatzki, H. (2006) Complex patterns of copy number variation at sites of segmental duplications: an important category of structural variation in the human genome. Human genetics. 120 (2), 270–284.

Goldberg, D., Grunwald, D., Lewis, C., Feld, J., Donley, K. & Edbrooke, O. (2013) Addressing 21st century skills by embedding computer science in K-12 classes. Technical symposium on computer science education p.637–638.

Goldberg, D., Grunwald, D., Lewis, C., Feld, J., Donley, K. & Edbrooke, O. (2013) 'Addressing 21st century skills by embedding computer science in K-12 classes', in Proceeding of the 44th ACM technical symposium. [Online]. 1 January 2013 New York, New York, USA: SIGCSE, ACM Special Interest Group on Computer Science Education. pp. 637–638.

Goldman, N., Bertone, P., Chen, S. & Dessimoz, C. (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature

Gong, Z., Liu, X., Tang, D., Yu, H., Yi, C., Cheng, Z. & Gu, M. (2011) Non-homologous chromosome pairing and crossover formation in haploid rice meiosis. Chromosoma. 120 (1), 47–60.

Goodstadt, L. & Ponting, C.P. (2006) Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Computational Biology. 2 (9), e133.

Gorlova, O., Fedorov, A., Logothetis, C., Amos, C. & Gorlov, I. (2014) Genes with a large intronic burden show greater evolutionary conservation on the protein level. BMC Evolutionary Biology. 14 (1), 50.

Graubert, T.A., Cahan, P., Edwin, D., Selzer, R.R., Richmond, T.A., Eis, P.S., Shannon, W.D., Li, X., McLeod, H.L., Cheverud, J.M. & Ley, T.J. (2007) A High-Resolution Map of Segmental DNA Copy Number Variation in the Mouse Genome. PLoS genetics. 3 (1), e3.

Graur, D., Zheng, Y., Price, N., Azevedo, R.B.R., Zufall, R.A. & Elhaik, E. (2013) On the immortality of television sets: 'function' in the human genome according to the evolution-free gospel of ENCODE. Genome biology and evolution. 5 (3), 578–590.

Green, P. (2007) 2x genomes Does depth matter? Genome research. 17 (11), 1547–1549.

Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., Du, L., Egholm, M., Rothberg, J.M., Paunovic, M. & Pääbo, S. (2006) Analysis of one million base pairs of Neanderthal DNA. Nature. 444 (7117), 330–336.

Gregory, T.R. (2004) The Evolution of the Genome. T Ryan Gregory (ed.). Academic Press.

Grigoriev, A. (1998) Analyzing genomes with cumulative skew diagrams. Nucleic Acids Research. 26 (10), 2286–2290.

Gross, M. (2011) Riding the wave of biological data. Current Biology. 21 (6), R204–R206.

Gu, X., Wang, Y. & Gu, J. (2002) Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nature Genetics. 31 (2), 205–209.

Guryev, V., Saar, K., Adamovic, T., Verheul, M., van Heesch, S.A.A.C., Cook, S., Pravenec, M., Aitman, T., Jacob, H., Shull, J.D., Hübner, N. & Cuppen, E. (2008) Distribution and functional impact of DNA copy number variation in the rat. Nature genetics. 40 (5), 538–545.

Haddrill, P.R., Charlesworth, B., Halligan, D.L. & Andolfatto, P. (2005) Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content. Genome Biol. 6 (8), R67.

Haffter P et al. 1996. The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. Development. 123:1-36.

Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C., al, et & Cristianini, N. (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome research. 15 (8), 1153–1160.

Hahn, M.W., Demuth, J.P. & Han, S.-G.G. (2007) Accelerated rate of gene gain and loss in primates. Genetics. 177 (3), 1941–1949.

Hahn, M.W., Han, M. V & Han, S.-G. (2007) Gene Family Evolution across 12 Drosophila Genomes. PLoS genetics. 3 (11), e197.

Hambrusch, S., Hoffmann, C., Korb, J.T., Haugan, M. & Hosking, A.L. (2009) A multidisciplinary approach towards computational thinking for science majors. ACM SIGCSE Bulletin 41 (1) p.183.

Hamm, G.H. & Cameron, G.N. (1986) The EMBL data library. Nucleic Acids Research. 14 (1), 5–9.

Han, M. V, Demuth, J.P., McGrath, C.L., Casola, C. & Hahn, M.W. (2009) Adaptive evolution of young gene duplicates in mammals. Genome research. 19 (5), 859–867.

Han, M. V, Thomas, G.W.C., Lugo-Martinez, J. & Hahn, M.W. (2013) Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Molecular biology and evolution. 30 (8), 1987–1997.

Hankeln T, Friedl H, Ebersberger I, Martin J and Schmidt ER. 1997. A variable intron distribution in globin genes of Chironomus: evidence for recent intron gain. Gene. 205:151-60.

Hankeln, T., Friedl, H., Ebersberger, I., Martin, J. & Schmidt, E.R. (1997) A variable intron distribution in globin genes of Chironomus: evidence for recent intron gain. Gene. 205 (1-2), 151–160.

Hansen, T.F., Pienaar, J. & Orzack, S.H. (2008) A COMPARATIVE METHOD FOR STUDYING ADAPTATION TO A RANDOMLY EVOLVING ENVIRONMENT. Evolution. 62 (8), 1965–1977.

Hardison, R.C. (2003) Comparative Genomics. PLoS Biology. 1 (2), E58–E58.

Harris, E.E. (2008) Searching the genome for our adaptations. Evolutionary Anthropology: Issues.

Harvey, P.H. & Pagel, M.D. (1991) The Comparative Method in Evolutionary Biology. Oxford University Press, USA.

Hatton AR, Subramaniam V and Lopez AJ. 1998. Generation of Alternative Ultrabithorax Isoforms and Stepwise Removal of a Large Intron by Resplicing at Exon–Exon Junctions. Molecular Cell. 2:787-796.

Hawkins JS, Kim H, Nason JD, Wing RA and Wendel JF. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome research. 16:1252-61.

He, X. & Zhang, J. (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics. 169 (2), 1157–1164.

Hedges, S.B., Dudley, J. & Kumar, S. (2006) TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics (Oxford, England). 22 (23), 2971–2972.

Hey, J. & Kliman, R.M. (2002) Interactions between natural selection, recombination and gene density in the genes of Drosophila. Genetics. 160 (2), 595–608.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., Hillman-Jackson, J., Kuhn, R.M., Pedersen, J.S., Pohl, A., Raney, B.J., Rosenbloom, K.R., Siepel, A., Smith, K.E., Sugnet, C.W., et al. (2006) The UCSC Genome Browser Database: update 2006. Nucleic Acids Research. 34 (Database issue), D590–8.

Hinsen, K. (2011) A data and code model for reproducible research and executable papers. Procedia Computer Science. 4579–588.

Hiratani, I. & Gilbert, D.M. (2009) Replication timing as an epigenetic mark. Epigenetics : official journal of the DNA Methylation Society. 4 (2), 93–97.

Hiratani, I., Takebayashi, S., Lu, J. & Gilbert, D.M. (2009) Replication timing and transcriptional control: beyond cause and effect—part II. Current Opinion in Genetics & Development. 19 (2), 142–149.

Hobolth, A., Dutheil, J.Y., Hawks, J., Schierup, M.H. & Mailund, T. (2011) Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome research. 21 (3), 349–356.

Hoffa, C., Mehta, G., Freeman, T., Deelman, E., Keahey, K., Berriman, B. & Good, J. (2008) On the Use of Cloud Computing for Scientific Workflows. Audio, Transactions of the IRE Professional Group on. 640–645.

Hofmann, C.M. & Carleton, K.L. (2009) Gene duplication and differential gene expression play an important role in the diversification of visual pigments in fish. Integrative and Comparative Biology. 49 (6), 630–643.

Honee, C., Hedin, D., St-Laurent, J. & Froling, M. (2012) Environmental performance of data centres - A case study of the Swedish National Insurance Administration. Electronics Goes Green 2012+ (EGG), 2012. 1–6.

Hong, X., Scofield, D.G. & Lynch, M. (2006) Intron size, abundance, and distribution within untranslated regions of genes. Molecular biology and evolution. 23 (12), 2392–2404.

Hooks, K.B., Delneri, D. & Griffiths-Jones, S. (2014) Intron Evolution in Saccharomycetaceae. Genome biology and evolution. 6 (9), 2543–2556.

Hosouchi, T. (2002) Physical Map-Based Sizes of the Centromeric Regions of Arabidopsis thaliana Chromosomes 1, 2, and 3. DNA Research. 9 (4), 117–121.

Hothorn, T. & Leisch, F. (2011) Case studies in reproducibility. Briefings in Bioinformatics. 12 (3), 288–300.

Howison, M., Zapata, F. & Dunn, C.W. (2013) Toward a statistically explicit understanding of de novo sequence assembly. Bioinformatics (Oxford, England). 29 (23), 2959–2963.

Huang, Y.-T., Chao, K.-M. & Chen, T. (2005) An approximation algorithm for haplotype inference by maximum parsimony. Journal of computational biology : a journal of computational molecular cell biology. 12 (10), 1261–1274.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., et al. (2002) The Ensembl genome database project. Nucleic Acids Research. 30 (1), 38–41.

Hubbard, T.J.P., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., et al. (2009) Ensembl 2009. Nucleic Acids Research. 37 (Database), D690–D697.

Hubisz, M.J., Lin, M.F., Kellis, M. & Siepel, A. (2011) Error and Error Mitigation in Low-Coverage Genome Assemblies Thomas Mailund (ed.). PLoS One. 6 (2), e17034.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R. & Bollback, J.P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science (New York, N.Y.). 294 (5550), 2310–2314.

Hughes AL and Yeager M. 1997. Comparative evolutionary rates of introns and exons in murine rodents. Journal of Molecular Evolution. 45:125-130.

Hunt, C., Moore, K., Xiang, Z., Hurst, S.M., McDougall, R.C., Rajandream, M.-A.A. le, Barrell, B.G., Gwilliam, R., Wood, V., Lyne, M.H. & Aves, S.J. (2001) Subtelomeric sequence from the right arm ofSchizosaccharomyces pombe chromosome I contains seven permease genes. Yeast. 18 (4), 355–361.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M. & Otto, T.D. (2013) REAPR: a universal tool for genome assembly evaluation. Genome Biol. 14 (5), R47.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Fraser, M., Gough, J., et al. (2011)

InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Research. 40 (Database issue), D306–12.

Hurst, L.D. (2013) Open questions: A logic (or lack thereof) of genome organization. BMC Biology. 11 (1), 58.

Husby, A., Ekblom, R. & Qvarnström, A. (2011) Let’s talk turkey: immune competence in domestic and wild fowl. Heredity. 107 (2), 103–104.

Huynen, M.A. & van Nimwegen, E. (1998) The frequency distribution of gene family sizes in complete genomes. Molecular biology and evolution. 15 (5), 583–589.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. & Lee, C. (2004) Detection of large-scale variation in the human genome. Nature genetics. 36 (9), 949–951.

Ilie, L. & Molnar, M. (2013) RACER: Rapid and Accurate Correction of Errors in Reads. Bioinformatics. 29 (19), btt407–2493.

Iliopoulos, I., Tsoka, S., Andrade, M.A., Enright, A.J., Carroll, M., Poullet, P., Promponas, V., Liakopoulos, T., Palaios, G., Pasquier, C., Hamodrakas, S., Tamames, J., Yagnik, A.T., Tramontano, A., Devos, D., Blaschke, C., Valencia, A., Brett, D., Martin, D., et al. (2003) Evaluation of annotation strategies using an entire genome sequence. Bioinformatics. 19 (6), 717–726.

Initiative, T.A.G. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 408 (6814), 796–815.

Irimia, M. & Roy, S.W. (2008) Spliceosomal introns as tools for genomic and evolutionary analysis. Nucleic Acids Research. 36 (5), 1703–1712.

Irimia, M., Rukov, J.L., Penny, D., Vinther, J., Garcia-Fernàndez, J. & Roy, S.W. (2008a) Origin of introns by ‘intronization’ of exonic sequences. Trends in Genetics. 24 (8), 378–381.

Irimia, M., Rukov, J.L., Penny, D., Vinther, J., Garcia-Fernàndez, J. & Roy, S.W. (2008b) Origin of introns by ‘intronization’ of exonic sequences. Trends in Genetics. 24 (8), 378–381.

Iskow, R.C., Gokcumen, O. & Lee, C. (2012) Exploring the role of copy number variants in human adaptation. Trends in Genetics. 28 (6), 245–257.

Iwamoto M, Maekawa M, Saito A, Higo H and Higo K. 1998. Evolutionary relationship of plant catalase genes inferred from exon-intron structures: isozyme divergence after the separation of monocots and dicots. TAG Theoretical and Applied Genetics. 97:9-19.

Jabbari, K. (2009) Encyclopedia of life sciences. Chichester, UK: John Wiley & Sons, Ltd.

Jaillon O et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature. 431:946-57.

Jaillon, O., Aury, J.-M. & Wincker, P. (2009) 'Changing by doubling', the impact of Whole Genome Duplications in the evolution of eukaryotes. Comptes rendus biologies. 332 (2-3), 241–253.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyère, C., et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 449 (7161), 463–467.

Jenkinson, A.M., Albrecht, M., Birney, E., Blankenburg, H., Down, T., Finn, R.D., Hermjakob, H., Hubbard, T.J.P., Jimenez, R.C., Jones, P., Kähäri, A., Kulesha, E., Macías, J.R., Reeves, G.A. & Prlić, A. (2008) Integrating biological data – the Distributed Annotation System. BMC bioinformatics. 9 (Suppl 8), S3.

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., Appels, R., Pfeifer, M., Tao, Y., Zhang, X., Jing, R., Zhang, C., Ma, Y., Gao, L., Gao, C., Spannagl, M., Mayer, K.F.X., Li, D., Pan, S., et al. (2013) Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496 (7443), 91–95.

Jiang, K. & Goertzen, L.R. (2011) Spliceosomal intron size expansion in domesticated grapevine (Vitis vinifera). BMC Research Notes. 4 (1), 52.

Jiang, S.-Y.Y., Ma, A., Ramamoorthy, R. & Ramachandran, S. (2013) Genome-Wide Survey on Genomic Variation, Expression Divergence, and Evolution in Two

Contrasting Rice Genotypes under High Salinity Stress. Genome biology and evolution. 5 (11), 2032–2050.

Jiang, Z., Hubley, R., Smit, A. & Eichler, E.E. (2008) DupMasker: a tool for annotating primate segmental duplications. Genome research. 18 (8), 1362–1368.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., et al. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*. 473 (7345), 97–100.

Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M.C., Myers, R.M., Miller, C.T., Summers, B.R., Knecht, A.K., et al. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 484 (7392), 55–61.

Jones, M.B., Schildhauer, M.P. & Reichman, O.J. (2006) The new bioinformatics: integrating ecological data from the gene to the biosphere. … Review of Ecology.

Jones, N.C. & Pevzner, P.A. (2004) An Introduction to Bioinformatics Algorithms. MIT Press.

Jurka J et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research. 110:462-7.

Kalari, K.R., Casavant, M., Bair, T.B., Keen, H.L., Comeron, J.M., Casavant, T.L. & Scheetz, T.E. (2006) First exons and introns--a survey of GC content and gene structure in the human genome. In silico biology. 6 (3), 237–242.

Kanehisa, M. & Bork, P. (2003) Bioinformatics in the post-sequence era. Nature Genetics. 33 (3s), 305–310.

Karev, G.P., Berezovskaya, F.S. & Koonin, E. V (2005) Modeling genome evolution with a diffusion approximation of a birth-and-death process. arXiv.org. q-bio.GN.

Karev, G.P., Wolf, Y.I. & Koonin, E. V (2003) Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? Bioinformatics. 19 (15), 1889–1900.

Karev, G.P., Wolf, Y.I., Berezovskaya, F.S. & Koonin, E. V (2004) Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. BMC Evolutionary Biology. 4 (1), 32.

Kari, L. & Rozenberg, G. (2008) The many facets of natural computing. Communications of the ACM. 51 (10), 72–83.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D., Kent, W.J. & University of California, S.C. (2003) The UCSC Genome Browser Database. Nucleic Acids Research. 31 (1), 51–54.

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., Harte, R.A., Heitner, S., Hinrichs, A.S., Learned, K., Lee, B.T., Li, C.H., Raney, B.J., Rhead, B., Rosenbloom, K.R., et al. (2014) The UCSC Genome Browser database: 2014 update. Nucleic Acids Research. 42 (Database issue), D764–70.

Kasprzyk, A. (2011) BioMart: driving a paradigm change in biological data management. Database. 2011 (0), bar049–bar049.

Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T. & Birney, E. (2004) EnsMart: a generic system for fast and flexible access to biological data. Genome research. 14 (1), 160–169.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research. 30 (14), 3059–3066.

Kaye, J., Heeney, C., Hawkins, N. & De Vries, J. (2009) Data sharing in genomics—re-shaping scientific practice. Nature Reviews ….

Keahey, K. & Freeman, T. (2008) Contextualization: Providing One-Click Virtual Clusters. Audio, Transactions of the IRE Professional Group on. 301–308.

Kejnovsky, E., Leitch, I.J. & Leitch, A.R. (2009) Contrasting evolutionary dynamics between angiosperm and mammalian genomes. Trends in Ecology & Evolution. 24 (10), 572–582.

Kent, W.J. (2002) BLAT---The BLAST-Like Alignment Tool. Genome research. 12 (4), 656–664.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. & Haussler, D. (2002) The human genome browser at UCSC. Genome research. 12 (6), 996–1006.

Kersey, P.J., Lawson, D., Birney, E., Derwent, P.S., Haimel, M., Herrero, J., Keenan, S., Kerhornou, A., Koscielny, G., Kahari, A., Kinsella, R.J., Kulesha, E., Maheswari, U., Megy, K., Nuhn, M., Proctor, G., Staines, D., Valentin, F., Vilella, A.J., et al. (2010) Ensembl Genomes: extending Ensembl across the taxonomic space. Nucleic Acids Research. 38 (Database issue), D563–9.

Kersey, P.J., Staines, D.M., Lawson, D., Kulesha, E., Derwent, P., Humphrey, J.C., Hughes, D.S.T., Keenan, S., Kerhornou, A., Koscielny, G., Langridge, N., McDowall, M.D., Megy, K., Maheswari, U., Nuhn, M., Paulini, M., Pedro, H., Toneva, I., Wilson, D., et al. (2011) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. Nucleic Acids Research. 40 (Database issue), D91–7.

Khaja, R., Zhang, J., MacDonald, J.R., He, Y., Joseph-George, A.M., Wei, J., Rafiq, M.A., Qian, C., Shago, M., Pantano, L., Aburatani, H., Jones, K., Redon, R., Hurles, M., Armengol, L., Estivill, X., Mural, R.J., Lee, C., Scherer, S.W., et al. (2006) Genome assembly comparison identifies structural variants in the human genome. Nature genetics. 38 (12), 1413–1418.

Kim E, Magen A and Ast G. 2007. Different levels of alternative splicing among eukaryotes. Nucleic acids research. 35:125-31.

Kim, Y.-J. & Han, K. (2015) Endogenous retrovirus-mediated genomic variations in chimpanzees. Mobile Genetic Elements. 4 (6), 1–4.

Kimura T et al. 2004. Large-scale isolation of ESTs from medaka embryos and its application to medaka developmental genetics. Mechanisms of development. 121:915-32.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. 16 (2), 111–120.

Kinsella, R.J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P. & Flicek, P. (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database. 2011 (0), bar030–bar030.

Kjeldbjerg, A.L., Villesen, P., Aagaard, L. & Pedersen, F.S. (2008) Gene conversion and purifying selection of a placenta-specific ERV-V envelope gene during simian evolution. BMC Evolutionary Biology. 8 (1), 266.

Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J.G., Easton, B.C., Eaton, M., Hamady, M., Lindsay, H., Liu, Z., Lozupone, C., McDonald, D., Robeson, M., Sammut, R., Smit, S., Wakefield, M.J., Widmann, J., Wikman, S., Wilson, S., et al. (2007) PyCogent: a toolkit for making sense from sequence. Genome Biol. 8 (8), R171.

Knowles, D.G. & McLysaght, A. (2009) Recent de novo origin of human protein-coding genes. Genome research. 19 (10), 1752–1759.

Kong, A., Barnard, J., Gudbjartsson, D.F., Thorleifsson, G., Jonsdottir, G., Sigurdardottir, S., Richardsson, B., Jonsdottir, J., Thorgeirsson, T., Frigge, M.L., Lamb, N.E., Sherman, S., Gulcher, J.R. & Stefansson, K. (2004) Recombination rate and reproductive success in humans. Nature Genetics. 36 (11), 1203–1206.

Koonin, E. V, Csürös, M. & Rogozin, I.B. (2013) Whence genes in pieces: reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. Wiley Interdisciplinary Reviews: RNA. 4 (1), 93–105.

Koonin, E. V, Wolf, Y.I. & Karev, G.P. (2002) The structure of the protein universe and genome evolution. Nature. 420 (6912), 218–223.

Kopena, J.B. & Regli, W.C. (2003) 'DAMLJessKB: A Tool for Reasoning with the Semantic Web BT - The Semantic Web - ISWC 2003', in The Semantic Web - ISWC 2003. [Online]. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 628–643.

Korbel, J.O., Kim, P.M., Chen, X., Urban, A.E., Weissman, S., Snyder, M. & Gerstein, M.B. (2008) The current excitement about copy-number variation: how it relates to gene duplications and protein families. Current opinion in structural biology. 18 (3), 366–374.

Kordiš, D., Kokošar, J., x161, K.K., an, x161, K.K. & ar, J. (2012) What Can Domesticated Genes Tell Us about the Intron Gain in Mammals? International journal of evolutionary biology. 2012 (1), 1–7.

Krampis, K., Booth, T., Chapman, B., Tiwari, B., Bicak, M., Field, D. & Nelson, K.E. (2012) Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC bioinformatics. 13 (1), 42.

Kruskal, W.H. & Wallis, W.A. (2012) Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association. 47 (260), 583–621.

Kuhn, R.M., Haussler, D. & Kent, W.J. (2012) The UCSC genome browser and associated tools. Briefings in bioinformatics. 14 (2), bbs038–161.

Kuhn, R.M., Karolchik, D., Zweig, A.S., Trumbower, H., Thomas, D.J., Thakkapallayil, A., Sugnet, C.W., Stanke, M., Smith, K.E., Siepel, A., Rosenbloom, K.R., Rhead, B., Raney, B.J., Pohl, A., Pedersen, J.S., Hsu, F., Hinrichs, A.S., Harte, R.A., Diekhans, M., et al. (2007) The UCSC genome browser database: update 2007. Nucleic Acids Research. 35 (Database issue), D668–D673.

Kuonen, D. (2003) Challenges in bioinformatics for statistical data miners. Bulletin of the Swiss Statistical Society.

Labarga, A., Valentin, F., Anderson, M. & Lopez, R. (2007) Web services at the European bioinformatics institute. Nucleic Acids Research. 35 (Web Server issue), W6–11.

Lancia, G., Bafna, V., Istrail, S., Lippert, R. & Schwartz, R. (2001) 'SNPs Problems, Complexity, and Algorithms BT - The Semantic Web - ISWC 2003', in The Semantic Web - ISWC 2003. [Online]. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 182–193.

Lander ES et al. 2001. Initial sequencing and analysis of the human genome. Nature. 409:860-921.

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., et al. (2001) Initial sequencing and analysis of the human genome. Nature. 409 (6822), 860–921.

Langhammer, P.F., Lips, K.R., Burrowes, P.A., Tunstall, T., Palmer, C.M. & Collins, J.P. (2013) A fungal pathogen of amphibians, Batrachochytrium dendrobatidis, attenuates in pathogenicity with in vitro passages. Matthew Mat Â Charles Fisher (ed.). PLoS One. 8 (10), e77630.

Lartillot, N. (2013) Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. Molecular biology and evolution. 30 (3), 489–502.

Lathe, W., Williams, J., Mangan, M. & Karolchik, D. (2008) Genomic data resources: challenges and promises. Nature Education. 1 (3), 2.

Lattorff, H.M.G. & Moritz, R.F.A. (2013) Genetic underpinnings of division of labor in the honeybee (Apis mellifera). Trends in Genetics: TIG. 29 (11), 641–648.

Le, H.-S., Schulz, M.H., McCauley, B.M., Hinman, V.F. & Bar-Joseph, Z. (2013) Probabilistic error correction for RNA sequencing. Nucleic Acids Research. 41 (10), gkt215–e109.

Lee, A.S., Gutiérrez-Arcelus, M., Perry, G.H., Vallender, E.J., Johnson, W.E., Miller, G.M., Korbel, J.O. & Lee, C. (2008) Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. Human Molecular Genetics. 17 (8), 1127–1136.

Lee, T.B., Cailliau, R., Groff, J.F. & Pollermann, B. (2013) World ‐ wide web: the information universe. Internet Research. 20 (4), 461 ‐ 471.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., et al. (2010) The European Nucleotide Archive. Nucleic Acids Research. 39 (Database), gkq967–D31.

Lercher, M.J. & Hurst, L.D. (2002) Human SNP variability and mutation rate are higher in regions of high recombination. Trends in Genetics: TIG. 18 (7), 337–340.

Leroy, S., Fargette, M., Morand, S. & Bouamer, S. (2007) Genome size of plant-parasitic nematodes. Nematology. 9 (3), 449–450.

Levasseur, A. & Pontarotti, P. (2011) The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. Biol Direct.

Levine, M.T., Jones, C.D., Kern, A.D., Lindfors, H.A. & Begun, D.J. (2006) Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proceedings of the National Academy of Sciences. 103 (26), 9935–9939.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F. & Denisov, G. (2007) The diploid genome sequence of an individual human. PLoS Biology. 5 (10), 1–32.

Li W, Tucker AE, Sung W, Thomas WK and Lynch M. 2009. Extensive, recent intron gains in Daphnia populations. Science (New York, N.Y.). 326:1260-2.

Li, H., Liu, G. & Xia, X. (2009) Correlations between recombination rate and intron distributions along chromosomes of C. elegans. Progress in Natural Science. 19 (4), 517–522.

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

Li, W., Tucker, A.E., Sung, W., Thomas, W.K. & Lynch, M. (2009) Extensive, recent intron gains in Daphnia populations. Science (New York, N.Y.). 326 (5957), 1260–1262.

Libbrecht, R., Oxley, P.R., Kronauer, D.J.C. & Keller, L. (2013) Ant genomics sheds light on the molecular regulation of social organization. Genome Biol. 14 (7), 212.

Librado Sanz, P., Vieira, F.G. & Rozas Liras, J.A. (2011) BadiRate Software.

Librado, P., Vieira, F.G. & Rozas, J. (2012) BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics (Oxford, England). 28 (2), 279–281.

Lin C-F, Mount SM, Jarmołowski A and Makałowski W. 2010. Evolutionary dynamics of U12-type spliceosomal introns. BMC Evolutionary Biology. 10:47.

Lin, H., Zhu, W., Silva, J.C., Gu, X. & Buell, C.R. (2006) Intron gain and loss in segmentally duplicated genes in rice. Genome Biol. 7 (5), R41.

Lin, J. & Ryaboy, D. (2013) Scaling big data mining infrastructure: the twitter experience. ACM SIGKDD Explorations Newsletter. 14 (2), 6–19.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L.D., Lowe, C.B., Holloway, A.K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 478 (7370), 476–482.

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., et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438 (7069), 803–819.

Lipinski, K.J., Farslow, J.C., Fitzpatrick, K.A., Lynch, M., Katju, V. & Bergthorsson, U. (2011) High Spontaneous Rate of Gene Duplication in Caenorhabditis elegans. Current Biology. 21 (4), 306–310.

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V., Tsai, I.J., Bergman, C.M., Bensasson, D., O'Kelly, M.J.T., van Oudenaarden, A., Barton, D.B.H., Bailes, E., Nguyen, A.N., Jones, M., et al. (2009) Population genomics of domestic and wild yeasts. Nature. 458 (7236), 337–341.

Liu, G.E., Hou, Y., Zhu, B., Cardone, M.F., Jiang, L., Cellamare, A., Mitra, A., Alexander, L.J., Coutinho, L.L., Dell'Aquila, M.E., Gasbarre, L.C., Lacalandra, G., Li, R.W., Matukumalli, L.K., Nonneman, D., de A Regitano, L.C., Smith, T.P.L., Song, J., Sonstegard, T.S., et al. (2010) Analysis of copy number variations among diverse cattle breeds. Genome research. 20 (5), 693–703.

Liu, L., Li, Y., Li, S., Hu, N., He, Y. & Pong, R. (2012) Comparison of next-generation sequencing systems. BioMed Research ….

Liu, L., Yu, L., Kalavacharla, V. & Liu, Z. (2011) A Bayesian model for gene family evolution. BMC bioinformatics. 12 (1), 426.

Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Molecular biology and evolution. 13 (5), 660–665.

Locke, D.P., Hillier, L.W., Warren, W.C., Worley, K.C., Nazareth, L. V, Muzny, D.M., Yang, S.-P., Wang, Z., Chinwalla, A.T., Minx, P., Mitreva, M., Cook, L., Delehaunty, K.D., Fronick, C., Schmidt, H., Fulton, L.A., Fulton, R.S., Nelson, J.O., Magrini, V., et al. (2011) Comparative and demographic analysis of orang-utan genomes. Nature. 469 (7331), 529–533.

Locke, D.P., Segraves, R., Carbone, L., Archidiacono, N., Albertson, D.G., Pinkel, D. & Eichler, E.E. (2003) Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. Genome research. 13 (3), 347–357.

López-Bigas N, Audit B, Ouzounis C, Parra G and Guigó R. 2005. Are splicing mutations the most frequent cause of hereditary disease? FEBS letters. 579:1900-3.

Lord, P., Bechhofer, S., Wilkinson, M.D., Schiltz, G., Gessler, D., Hull, D., Goble, C. & Stein, L. (2004) 'Applying Semantic Web Services to Bioinformatics: Experiences Gained, Lessons Learnt BT - The Semantic Web – ISWC 2004', in The Semantic Web – ISWC 2004. [Online]. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 350–364.

Loveland, J. (2005) VEGA, the genome browser with a difference. Briefings in bioinformatics. 6 (2), 189–193.

Löytynoja, A., Vilella, A.J. & Goldman, N. (2012) Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. Bioinformatics (Oxford, England). 28 (13), 1684–1691.

Luebke, D. (2008) 'CUDA: Scalable parallel programming for high-performance scientific computing', in 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Proceedings, ISBI. [Online]. 1 January 2008 IEEE. pp. 836–838.

Lunt, D. (2013) Calculating intron density.

Luscombe, N.M., Qian, J., Zhang, Z., Johnson, T. & Gerstein, M. (2002) The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome biology. 3 (8), RESEARCH0040.

Lynch M and Conery JS. 2003. The origins of genome complexity. Science (New York, N.Y.). 302:1401-4.

Lynch, M. (2002) The evolution of spliceosomal introns. Current Opinion in Genetics & Development. 12 (6), 701–710.

Lynch, M. (2002a) Genomics. Gene duplication and evolution. Science (New York, N.Y.). 297 (5583), 945–947.

Lynch, M. (2002b) Intron evolution as a population-genetic process. Proceedings of the National Academy of Sciences. 99 (9), 6118–6123.

Lynch, M. (2007) The Origins of Genome Architeture. Sinauer Associates Incorporated.

Lynch, M. (2011) Statistical Inference on the Mechanisms of Genome Evolution Nancy A Moran (ed.). PLoS genetics. 7 (6), e1001389.

Lynch, M. (2012) The evolution of multimeric protein assemblages. Molecular biology and evolution. 29 (5), 1353–1366.

Lynch, M. & Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. Science. 290 (5494), 1151–1155.

Lynch, M. & Conery, J.S. (2003) The evolutionary demography of duplicate genes. Journal of structural and functional genomics. 3 (1), 35–44.

Lynch, M. & Conery, J.S. (2003) The origins of genome complexity. Science (New York, N.Y.). 302 (5649), 1401–1404.

Lynch, M. & Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics. 154 (1), 459–473.

Lynch, M. & Katju, V. (2004) The altered evolutionary trajectories of gene duplicates. Trends in Genetics. 20 (11), 544–549.

Macmanes, M.D. & Eisen, M.B. (2013) Improving transcriptome assembly through error correction of high-throughput sequence reads. arXiv.org. q-bio.GN (R116), e113.

Macrini, T.E. (2004) Monodelphis domestica. Mammalian Species. 7601–8.

Maeso, I., Roy, S.W. & Irimia, M. (2012) Widespread Recurrent Evolution of Genomic Features. Genome biology and evolution. 4 (4), 486–500.

Mahmoody, A., Kahn, C.L. & Raphael, B.J. (2012) Reconstructing genome mixtures from partial adjacencies. BMC bioinformatics. 13 Suppl 1 (Suppl 19), S9.

Majewski J and Ott J. 2002. Distribution and characterization of regulatory elements in the human genome. Genome research. 12:1827-36.

Mak, H.C. (2011) Next-generation sequence analysis. Nat Biotech. 29 (1), 45–46.

Malik, H.S. & Henikoff, S. (2001) Adaptive evolution of Cid, a centromere-specific histone in Drosophila. Genetics. 157 (3), 1293–1298.

Malik, H.S. & Henikoff, S. (2009) Major Evolutionary Transitions in Centromere Complexity. Cell. 138 (6), 1067–1082.

Manavski, S.A. & Valle, G. (2008) CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment. BMC bioinformatics. 9 (Suppl 2), S10–9.

Marais G, Nouvellet P, Keightley PD and Charlesworth B. 2005. Intron size and exon evolution in Drosophila. Genetics. 170:481-5.

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

Marques-Bonet, T. & Eichler, E.E. (2009) The evolution of human segmental duplications and the core duplicon hypothesis. Cold Spring Harbor Symposia on Quantitative Biology. 74 (0), 355–362.

Marques-Bonet, T., Girirajan, S. & Eichler, E.E. (2009) The origins and impact of primate segmental duplications. Trends in Genetics. 25 (10), 443–454.

Marques-Bonet, T., Kidd, J.M., Ventura, M., Graves, T.A., Cheng, Z., Hillier, L.W., Jiang, Z., Baker, C., Malfavon-Borja, R., Fulton, L.A., Alkan, C., Aksay, G., Girirajan, S., Siswara, P., Chen, L., Cardone, M.F., Navarro, A., Mardis, E.R., Wilson, R.K., et al. (2009) A burst of segmental duplications in the genome of the African great ape ancestor. Nature. 457 (7231), 877–881.

Marsolier-Kergoat, M.-C. & Yeramian, E. (2009) GC content and recombination: reassessing the causal effects for the Saccharomyces cerevisiae genome. Genetics. 183 (1), 31–38.

Martin, A.P. (1999) Increasing Genomic Complexity by Gene Duplication and the Origin of Vertebrates. The American naturalist. 154 (2), 111–128.

Martin, A.P. & Palumbi, S.R. (1993) Body size, metabolic rate, generation time, and the molecular clock. Proceedings of the National Academy of Sciences. 90 (9), 4087–4091.

Martins, E.P. (2000) Adaptation and the comparative method. Trends in Ecology & Evolution. 15 (7), 296–299.

Martins, E.P. & Hansen, T.F. (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. American Naturalist.

Marygold, S.J., Leyland, P.C., Seal, R.L., Goodman, J.L., Thurmond, J., Strelets, V.B., Wilson, R.J. & consortium, the F. (2012) FlyBase: improvements to the bibliography. Nucleic Acids Research. 41 (D1), gks1024–D757.

Mason, C.E. & Elemento, O. (2012) Faster sequencers, larger datasets, new challenges. Genome Biol. 13 (3), 314.

Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Molecular biology and evolution. 21 (9), 1781–1791.

McCoy, R.C., Taylor, R.W., Blauwkamp, T.A., Kelley, J.L., Kertesz, M., Pushkarev, D., Petrov, D.A. & Fiston-Lavier, A.-S. (2014) Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements Nadia Singh (ed.). PLoS One. 9 (9), e106689.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M.A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 20 (9), 1297–1303.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. & Cunningham, F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 26 (16), 2069–2070.

McLysaght, A., Enright, A.J., Skrabanek, L. & Wolfe, K.H. (2000) Estimation of Synteny Conservation and Genome Compaction Between Pufferfish &#x28;Fugu&#x29; and Human. International Journal of Genomics. 1 (1), 22–36.

McLysaght, A., Hokamp, K. & Wolfe, K.H. (2002) Extensive genomic duplication during early chordate evolution. *Nature Genetics*. 31 (2), 200–204.

McVean, G.A., Altshuler Co-Chair, D.M., Durbin Co-Chair, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., Gibbs, R.A., Green, E.D., Hurles, M.E., Knoppers, B.M., Korbel, J.O., Lander, E.S., Lee, C., Lehrach, H., et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature. 491 (7422), 56–65.

McVean, G.A.T., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R. & Donnelly, P. (2004) The fine-scale structure of recombination rate variation in the human genome. Science (New York, N.Y.). 304 (5670), 581–584.

Medvedev, P. & Brudno, M. (2009) Maximum likelihood genome assembly. Journal of Computational Biology. 16 (8), 1101–1116.

Medvedev, P., Scott, E., Kakaradov, B. & Pevzner, P. (2011) Error correction of high-throughput sequencing datasets with non-uniform coverage. Bioinformatics (Oxford, England). 27 (13), i137–41.

Mele, M., Javed, A., Pybus, M., Zalloua, P., Haber, M., Comas, D., Netea, M.G., Balanovsky, O., Balanovska, E., Jin, L., Yang, Y., Pitchappan, R.M., Arunkumar, G., Parida, L., Calafell, F., Bertranpetit, J. & Consortium, G. (2012) Recombination gives a new insight in the effective population size and the history of the old world human populations. Molecular biology and evolution. 29 (1), 25–30.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L., Marshall, W.F., Qu, L.-H., Nelson, D.R., Sanderfoot, A.A., Spalding, M.H., Kapitonov, V. V, Ren, Q., Ferris, P., Lindquist, E., et al. (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science. 318 (5848), 245–250.

Metzker, M.L. (2010) Sequencing technologies—the next generation. Nature Reviews Genetics.

Meunier, J. & Duret, L. (2004) Recombination drives the evolution of GC-content in the human genome. Molecular biology and evolution. 21 (6), 984–990.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., Raney, B.J., Pohl, A., Malladi, V.S., Li, C.H., Lee, B.T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., et al. (2013) The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Research. 41 (Database issue), D64–9.

Michel, B. (2000) Replication fork arrest and DNA recombination. Trends in biochemical sciences. 25 (4), 173–178.

Mighell, A.J., Smith, N.R., Robinson, P.A. & Markham, A.F. (2000) Vertebrate pseudogenes. FEBS Letters. 468 (2-3), 109–114.

Mikkelsen, T.S., Wakefield, M.J., Aken, B., Amemiya, C.T., Chang, J.L., Duke, S., Garber, M., Gentles, A.J., Goodstadt, L., Heger, A., Jurka, J., Kamal, M., Mauceli, E., Searle, S.M.J., Sharpe, T., Baker, M.L., Batzer, M.A., Benos, P. V, Belov, K., et al.

(2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature 447 (7141), 167–177.

Milinkovitch, M.C., Helaers, R., Depiereux, E., Tzika, A.C. & Gabaldón, T. (2010) 2x genomes--depth does matter. Genome Biol. 11 (2), R16.

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

Mills RE, Bennett EA, Iskow RC and Devine SE. 2007. Which transposable elements are active in the human genome? Trends in genetics : TIG. 23:183-91.

Mironov AA, Fickett JW and Gelfand MS. 1999. Frequent alternative splicing of human genes. Genome research. 9:1288-93.

Money, D. & Whelan, S. (2012) Characterizing the Phylogenetic Tree-Search Problem. Systematic Biology. 61 (2), 228–239.

Montoya-Burgos, J.I., Boursot, P. & Galtier, N. (2003) Recombination explains isochores in mammalian genomes. Trends in Genetics. 19 (3), 128–130.

Moore, G.E. (1965) Cramming more components onto integrated circuits. Electronics. 114–117.

Moore, J.H., Asselbergs, F.W. & Williams, S.M. (2010) Bioinformatics challenges for genome-wide association studies. Bioinformatics (Oxford, England). 26 (4), 445–455.

Morgulis A, Gertz EM, Schäffer AA and Agarwala R. 2006. WindowMasker: window-based masker for sequenced genomes. Bioinformatics (Oxford, England). 22:134-41.

Morris, R.T. & Drouin, G. (2011) Ectopic Gene Conversions in the Genome of Ten Hemiascomycete Yeast Species. International journal of evolutionary biology. 2011 (14), 1–11.

Moss, S.P., Joyce, D. a, Humphries, S., Tindall, K.J. & Lunt, D.H. (2011) Comparative analysis of teleost genome sequences reveals an ancient intron size expansion in the zebrafish lineage. Genome biology and evolution. 31187–96.

Mulder, N.J., Kersey, P., Pruess, M. & Apweiler, R. (2007) In Silico Characterization of Proteins: UniProt, InterPro and Integr8. Molecular biotechnology. 38 (2), 165–177.

N, R., A, M.-B., A, G., C, N., J, F., J, W., Cotton, V. & A, H. (2009) The role of recombination in telomere length maintenance. Biochemical Society Transactions. 37 (3), 589.

Nabiyouni, M., Prakash, A. & Fedorov, A. (2013) Vertebrate codon bias indicates a highly GC-rich ancestral genome. Gene. 519 (1), 113–119.

Nagaki, K., Cheng, Z., Ouyang, S., Talbert, P.B., Kim, M., Jones, K.M., Henikoff, S., Buell, C.R. & Jiang, J. (2004) Sequencing of a rice centromere uncovers active genes. Nature Genetics. 36 (2), 138–145.

Nagarajan, N. & Pop, M. (2009) Parametric Complexity of Sequence Assembly: Theory and Applications to Next Generation Sequencing. Journal of computational biology : a journal of computational molecular cell biology. 16 (7), 897–908.

Nam, K. & ELLEGREN, H. (2012) Recombination Drives Vertebrate Genome Contraction Dmitri A Petrov (ed.). PLoS genetics. 8 (5), e1002680.

Neafsey DE and Palumbi SR. 2003. Genome size evolution in pufferfish: a comparative analysis of diodontid and tetraodontid pufferfish genomes. Genome research. 13:821-30.

Nelson, B. (2009) Data sharing: Empty archives. Nature. 461 (7261), 160–163.

Neuvéglise, C., Marck, C. & Gaillardin, C. (2011) The intronome of budding yeasts. Comptes rendus biologies. 334 (8-9), 662–670.

Nicholas, T.J., Baker, C., Eichler, E.E. & Akey, J.M. (2011) A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog. BMC Genomics. 12414.

Niedringhaus, T.P., Milanova, D., Kerby, M.B., Snyder, M.P. & Barron, A.E. (2011) Landscape of Next-Generation Sequencing Technologies. Analytical Chemistry. 83 (12), 4327–4341.

Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., J Sninsky, J., Adams, M.D. & Cargill, M. (2005) A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. PLoS Biology. 3 (6), e170.

Niemenmaa, M., Kallio, A., Schumacher, A., Klemela, P., Korpelainen, E. & Heljanko, K. (2012) Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. Bioinformatics. 28 (6), 876–877.

Niu, D.-K. & Jiang, L. (2013) Can ENCODE tell us how much junk DNA we carry in our genome? Biochemical and biophysical research communications. 430 (4), 1340–1343.

Noble, W.S. (2009) A Quick Guide to Organizing Computational Biology Projects Fran Lewitter (ed.). PLoS Computational Biology. 5 (7), e1000424.

Noonan, J.P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Paabo, S., Pritchard, J.K. & Rubin, E.M. (2006) Sequencing and analysis of Neanderthal genomic DNA. Science. 314 (5802), 1113–1118.

Norman, J.D., Danzmann, R.G., Glebe, B. & Ferguson, M.M. (2011) The genetic basis of salinity tolerance traits in Arctic charr (Salvelinus alpinus). BMC Genetics. 12 (1), 81.

Norman, J.D., Ferguson, M.M. & Danzmann, R.G. (2014) Transcriptomics of salinity tolerance capacity in Arctic charr (Salvelinus alpinus): a comparison of gene expression profiles between divergent QTL genotypes. Physiological …. 46 (4), 123–137.

Norman, J.D., Robinson, M., Glebe, B., Ferguson, M.M. & Danzmann, R.G. (2012) Genomic arrangement of salinity tolerance QTLs in salmonids: a comparative analysis of Atlantic salmon (Salmo salar) with Arctic charr (Salvelinus alpinus) and rainbow trout (Oncorhynchus mykiss). BMC Genomics. 13420.

Novozhilov, A.S., Karev, G.P. & Koonin, E. V (2006) Biological applications of the theory of birth-and-death processes. Briefings in bioinformatics. 7 (1), 70–85.

Nygaard, S., Zhang, G., Schiøtt, M., Li, C., Wurm, Y., Hu, H., Zhou, J., Ji, L., Qiu, F., Rasmussen, M., Pan, H., Hauser, F., Krogh, A., Grimmelikhuijzen, C.J.P., Wang, J.

& Boomsma, J.J. (2011) The genome of the leaf-cutting ant Acromyrmex echinatior suggests key adaptations to advanced social life and fungus farming. Genome research. 21 (8), 1339–1348.

Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., et al. (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497 (7451), 579–584.

Oh, D.-H., Dassanayake, M., Bohnert, H.J. & Cheeseman, J.M. (2012) Life at the extreme: lessons from the genome. Genome Biol. 13 (3), 241.

Ohno, S. (1970) Evolution by Gene Duplication. Springer.

Oleksyk, T.K., Smith, M.W. & O'Brien, S.J. (2009) Genome-wide scans for footprints of natural selection. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. 365 (1537), 185–205.

Oliver, J.L., Bernaola-Galván, P., Carpena, P. & Román-Roldán, R. (2001) Isochore chromosome maps of eukaryotic genomes. Gene. 276 (1-2), 47–56.

Olsen, K.M. & Wendel, J.F. (2013) A bountiful harvest: genomic insights into crop domestication phenotypes. Annual Review of Plant Biology. 6447–70.

Oostlander, A.E., Meijer, G.A. & Ylstra, B. (2004) Microarray-based comparative genomic hybridization and its applications in human genetics. Clinical Genetics. 66 (6), 488–495.

Ophir, R. & Graur, D. (1997) Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene. 205 (1-2), 191–202.

Organ, C.L., Shedlock, A.M., Meade, A., Pagel, M. & Edwards, S. V (2007) Origin of avian genome size and structure in non-avian dinosaurs. Nature. 446 (7132), 180–184.

Otto, S.P. & Whitton, J. (2000) Polyploid incidence and evolution. Annual Review of Genetics. 34 (1), 401–437.

Paces, J., Zíka, R., Paces, V., Pavlícek, A., Clay, O. & Bernardi, G. (2004) Representing GC variation along eukaryotic chromosomes. Gene. 333135–141.

Pagani, I., Liolios, K., Jansson, J., Chen, I.-M.A., Smirnova, T., Nosrat, B., Markowitz, V.M. & Kyrpides, N.C. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Research. 40 (Database issue), D571–9.

Pagel, M. (1994) 'Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters', in Proceedings of the Royal Society of …. [Online]. 1 January 1994

Pagel, M. (1999a) Inferring the historical patterns of biological evolution. Nature.

Pagel, M. (1999b) The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Systematic Biology.

Paradis, E., Claude, J. & Strimmer, K. (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 20 (2), 289–290.

Pareek, C.S., Smoczynski, R. & Tretyn, A. (2011) Sequencing technologies and genome sequencing. Journal of applied genetics. 52 (4), 413–435.

Parfrey, L.W., Lahr, D.J.G. & Katz, L.A. (2008) The Dynamic Nature of Eukaryotic Genomes. Molecular biology and evolution. 25 (4), 787–794.

Paszkiewicz, K. & Studholme, D.J. (2010) De novo assembly of short sequence reads. Briefings in Bioinformatics. 11 (5), 457–472.

Patel, A.A. & Steitz, J.A. (2003) Splicing double: insights from the second spliceosome. Nature Reviews Molecular Cell Biology. 4 (12), 960–970.

Patel, A.A., McCarthy, M. & Steitz, J.A. (2002a) The splicing of U12 ‐ type introns can be a rate ‐ limiting step in gene expression. The EMBO Journal. 21 (14), 3804 – 3815.

Patel, A.A., McCarthy, M. & Steitz, J.A. (2002b) The splicing of U12-type introns can be a rate-limiting step in gene expression. The EMBO Journal. 21 (14), 3804–3815.

Paten, B., Zerbino, D.R., Hickey, G. & Haussler, D. (2013) A Unifying Model of Genome Evolution Under Parsimony. arXiv.org. q-bio.GN.

Paterson, A.H., Freeling, M., Tang, H. & Wang, X. (2010) Insights from the Comparison of Plant Genome Sequences. Annual Review of Plant Biology. 61 (1), 349–372.

Paudel, Y., Madsen, O., Megens, H.-J., Frantz, L.A.F., Bosse, M., Bastiaansen, J.W.M., Crooijmans, R.P.M.A. & Groenen, M.A.M. (2013) Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. BMC Genomics. 14449.

Pearson, W.R. (2013) An introduction to sequence similarity ('homology') searching. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]. Chapter 3Unit3.1–3.1.8.

Pedersen, R.A. (1971) DNA content, ribosomal gene multiplicity, and cell size in fish. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution. 177 (1), 65–78.

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., Carter, N.P., Lee, C. & Stone, A.C. (2007) Diet and the evolution of human amylase gene copy number variation. Nature genetics. 39 (10), 1256–1260.

Perry, G.H., Tchinda, J., McGrath, S.D., Zhang, J., Picker, S.R., Cáceres, A.M., Iafrate, A.J., Tyler-Smith, C., Scherer, S.W., Eichler, E.E., Stone, A.C. & Lee, C. (2006) Hotspots for copy number variation in chimpanzees and humans. Proceedings of the National Academy of Sciences. 103 (21), 8006–8011.

Perry, G.H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A.S., Hyland, C., Stone, A.C., Hurles, M.E., Tyler-Smith, C., Eichler, E.E., Carter, N.P., Lee, C. & Redon, R. (2008) Copy number variation and evolution in humans and chimpanzees. Genome research. 18 (11), 1698–1710.

Petre, M. & Wilson, G. (2013) PLOS/Mozilla Scientific Code Review Pilot: Summary of Findings. arXiv.org. cs.SE.

Petronella, N. & Drouin, G. (2014) Purifying selection against gene conversions in the folate receptor genes of primates. Genomics. 103 (1), 40–47.

Petrov, D.A. (2002) Mutational Equilibrium Model of Genome Size Evolution. Theoretical Population Biology. 61 (4), 531–544.

Pettersson, E., Lundeberg, J. & Ahmadian, A. (2009) Generations of sequencing technologies. Genomics. 93 (2), 105–111.

Pevzner, P. & Shamir, R. (2011) Bioinformatics for Biologists.

Piast, M., Kustrzeba-Wojcicka, I. & Matusiewicz, M. (2007) Bioinformatics: From arduous beginnings to molecular databases. … IN CLINICAL AND ….

Piwowar, H.A., Day, R.S. & Fridsma, D.B. (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate John Ioannidis (ed.). PLoS One. 2 (3), e308.

Pop, M., Salzberg, S.L. & Shumway, M. (2002) Genome sequence assembly: algorithms and issues. Computer. 35 (7), 47–54.

Pope, B.D., Hiratani, I. & Gilbert, D.M. (2010) Domain-wide regulation of DNA replication timing during mammalian development. Chromosome Research. 18 (1), 127–136.

Potter, S.C., Clarke, L., Curwen, V., Keenan, S., Mongin, E., Searle, S.M.J., Stabenau, A., Storey, R. & Clamp, M. (2004) The Ensembl Analysis Pipeline. Genome research. 14 (5), 934–941.

Powers, J.G., Weigman, V.J., Shu, J., Pufky, J.M., Cox, D. & Hurban, P. (2013) Efficient and accurate whole genome assembly and methylome profiling of E. coli. BMC Genomics. 14 (1), 675.

Prachumwat, A., DeVincentis, L. & Palopoli, M.F. (2004) Intron size correlates positively with recombination rate in Caenorhabditis elegans. Genetics. 166 (3), 1585–1590.

Prado-Martinez, J., Hernando-Herraez, I., Lorente-Galdos, B., Dabad, M., Ramirez, O., Baeza-Delgado, C., Morcillo-Suarez, C., Alkan, C., Hormozdiari, F., Raineri, E., Estellé, J., Fernandez-Callejo, M., Valles, M., Ritscher, L., Schöneberg, T., de la Calle-Mustienes, E., Casillas, S., Rubio-Acero, R., Mele, M., et al. (2013) The genome sequencing of an albino Western lowland gorilla reveals inbreeding in the wild. BMC Genomics. 14 (1), 363.

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A.E., Malig, M., Hernandez-Rodriguez, J., et al. (2013) Great ape genetic diversity and population history. Nature 499 (7459), 471–475.

Price, M.N., Dehal, P.S. & Arkin, A.P. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular biology and evolution. 26 (7), 1641–1650.

Prlić, A. & Procter, J.B. (2012) Ten simple rules for the open development of scientific software. PLoS Computational Biology. 8 (12), e1002802.

Pronk, J.C., Frants, R.R., Jansen, W., Eriksson, A.W. & Tonino, G.J.M. (1982) Evidence of duplication of the human salivary amylase gene. Human genetics. 60 (1), 32–35.

Proost, S., Pattyn, P., Gerats, T. & Van de Peer, Y. (2011) Journey through the past: 150 million years of plant genome evolution. The Plant Journal. 66 (1), 58–65.

Protocols, C. (2002) Current Protocols in Bioinformatics. Current. Unit 6.2.

Proulx, S.R. (2012) Multiple routes to subfunctionalization and gene duplicate specialization. Genetics. 190 (2), 737–751.

Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J.R., Walenz, B., Koren, S., Sutton, G., Kodira, C., Winer, R., Knight, J.R., Mullikin, J.C., Meader, S.J., Ponting, C.P., Lunter, G., Higashino, S., Hobolth, A., Dutheil, J., Karakoç, E., et al. (2012) The bonobo genome compared with the chimpanzee and human genomes. Nature. 486 (7404), 527–531.

Pushkarev, D., Neff, N.F. & Quake, S.R. (2009) Single-molecule sequencing of an individual human genome. Nature Publishing Group. 27 (9), 847–850.

Pushker, R., Mira, A. & Rodríguez-Valera, F. (2004) Comparative genomics of gene-family size in closely related bacteria. Genome Biol. 5 (4), R27.

Qin, H. & Qin, H. (2009) Teaching computational thinking through bioinformatics to biology students. ACM SIGCSE Bulletin. 41 (1), 188–191.

Qiu, W.-G., Schisler, N. & Stoltzfus, A. (2004) The evolutionary gain of spliceosomal introns: sequence and phase preferences. Molecular biology and evolution. 21 (7), 1252–1263.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. & Lopez, R. (2005) InterProScan: protein domains identifier. Nucleic Acids Research. 33 (Web Server issue), W116–W120.

Rahman, A. & Pachter, L. (2013) CGAL: computing genome assembly likelihoods. Genome Biol.

Rajovic, N., Carpenter, P.M., Gelado, I., Puzovic, N., Ramirez, A. & Valero, M. (2013) 'Supercomputing with commodity CPUs', in the International Conference for High Performance Computing, Networking, Storage and Analysis. [Online]. 1 January 2013 New York, New York, USA: SIGARCH, ACM Special Interest Group on Computer Architecture. pp. 1–12.

Rajovic, N., Rico, A., Puzovic, N., Adeniyi-Jones, C. & Ramirez, A. (2014) Tibidabo: Making the case for an ARM-based HPC system. Future Generation Computer Systems. 36322–334.

Rajovic, N., Vilanova, L., Villavieja, C., Puzovic, N. & Ramirez, A. (2013) The low power architecture approach towards exascale computing. Journal of Computational Science. 4 (6), 439–443.

Rands, C.M., Meader, S., Ponting, C.P. & Lunter, G. (2014) 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. Mikkel H Schierup (ed.). PLoS genetics. 10 (7), e1004525.

Ranganathan, S. (2005) Bioinformatics Education—Perspectives and Challenges. PLoS Computational Biology. 1 (6), e52.

Rannala, B. & Yang, Z. (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164 (4), 1645–1656.

Rastogi, S. & Liberles, D.A. (2005) Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evolutionary Biology. 5 (1), 28.

Reenskaug, T.M.H. (1979) The original MVC reports.

Rhee, S.Y., Wood, V., Dolinski, K. & Draghici, S. (2008) Use and misuse of the gene ontology annotations. Nature Reviews Genetics. 9 (7), 509–515.

Ricker, N., Qian, H. & Fulthorpe, R.R. (2012) The limitations of draft assemblies for understanding prokaryotic adaptation and evolution. Genomics. 100 (3), 167–175.

Riethman, H.C., Xiang, Z., Paul, S., Morse, E., Hu, X.L., Flint, J., Chi, H.C., Grady, D.L. & Moyzis, R.K. (2001) Integration of telomere sequences with the draft human genome sequence. Nature. 409 (6822), 948–951.

Rios, D., McLaren, W.M., Chen, Y., Birney, E., Stabenau, A., Flicek, P. & Cunningham, F. (2010) A database and API for variation, dense genotyping and resequencing data. BMC bioinformatics. 11 (1), 238.

Rison, S.C.G., Hodgman, T.C. & Thornton, J.M. (2000) Comparison of functional annotation schemes for genomes. Functional & Integrative Genomics. 1 (1), 56–69.

Roberts, R.J., Carneiro, M.O. & Schatz, M.C. (2013) The advantages of SMRT sequencing. Genome Biol. 14 (7), 405.

Roch, S. (2006) A Short Proof that Phylogenetic Tree Reconstruction by Maximum Likelihood Is Hard. Computational Biology and Bioinformatics, IEEE/ACM Transactions on. 3 (1), 92–94.

Rocha, E.P.C. (2004) The replication-related organization of bacterial genomes. Microbiology (Reading, England). 150 (Pt 6), 1609–1627.

Rockman, M. V & Kruglyak, L. (2009) Recombinational landscape and population genomics of Caenorhabditis elegans. Molly Przeworski (ed.). PLoS genetics. 5 (3), e1000419.

Roest Crollius H and Weissenbach J. 2005. Fish genomics and biology. Genome research. 15:1675-82.

Rogers J. 1989. How were introns inserted into nuclear genes? Trends in Genetics. 5:213-216.

Rogers, J.H. (1989) How were introns inserted into nuclear genes? Trends in Genetics: TIG. 5 (7), 213–216.

Rognes, T. & Seeberg, E. (2000) Six-fold speed-up of Smith–Waterman sequence database searches using parallel processing on common microprocessors. Bioinformatics. 16 (8), 699–706.

Rogozin, I.B., Carmel, L., Csuros, M. & Koonin, E. V (2012) Origin and evolution of spliceosomal introns. Biol Direct.

Rogozin, I.B., Carmel, L., Csürös, M. & Koonin, E. V (2012) Origin and evolution of spliceosomal introns. Biol Direct. 7 (1), 11.

Rogozin, I.B., Wolf, Y.I., Sorokin, A. V, Mirkin, B.G. & Koonin, E. V (2003) Remarkable Interkingdom Conservation of Intron Positions and Massive, Lineage-Specific Intron Loss and Gain in Eukaryotic Evolution. Current Biology. 13 (17), 1512–1517.

Rokas, A. & Abbot, P. (2009) Harnessing genomics for evolutionary insights. Trends in Ecology & Evolution.

Romiguier, J., Ranwez, V., Douzery, E.J.P. & Galtier, N. (2010) Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome research. 20 (8), 1001–1009.

Roth, D.B., Porter, T.N. & Wilson, J.H. (1985) Mechanisms of nonhomologous recombination in mammalian cells. Molecular and Cellular Biology. 5 (10), 2599–2607.

Roy SW and Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. Nature reviews. Genetics. 7:211-21.

Roy, S.W. (2009) Intronization, de-intronization and intron sliding are rare in Cryptococcus. BMC Evolutionary Biology. 9 (1), 192.

Roy, S.W. & Gilbert, W. (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. Nature Reviews Genetics. 7 (3), 211–221.

Roy, S.W. & Irimia, M. (2009a) Mystery of intron gain: new data and new models. Trends in Genetics. 25 (2), 67–73.

Roy, S.W. & Irimia, M. (2009b) Splicing in the eukaryotic ancestor: form, function and dysfunction. Trends in Ecology & Evolution. 24 (8), 447–455.

Roy, S.W., Fedorov, A. & Gilbert, W. (2002) The signal of ancient introns is obscured by intron density and homolog number. Proceedings of the National Academy of Sciences of the United States of America. 99 (24), 15513–15517.

Ruiz, A. & Rueda-Almonacid, J.V. (2008) Batrachochytrium dendrobatidis and Chytridiomycosis in Anuran Amphibians of Colombia. EcoHealth. 5 (1), 27–33.

Ryba, T., Battaglia, D., Pope, B.D., Hiratani, I. & Gilbert, D.M. (2011) Genome-scale analysis of replication timing: from bench to bioinformatics. Nature protocols. 6 (6), 870–895.

Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T.C., Robins, A.J., Dalton, S. & Gilbert, D.M. (2010) Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome research. 20 (6), 761–770.

Sakudoh, T., Nakashima, T., Kuroki, Y., Fujiyama, A., Kohara, Y., Honda, N., Fujimoto, H., Shimada, T., Nakagaki, M., Banno, Y. & Tsuchida, K. (2011) Diversity in copy number and structure of a silkworm morphogenetic gene as a result of domestication. Genetics. 187 (3), 965–976.

Salier, J.-P.P. (2000) Chromosomal location, exon/intron organization and evolution of lipocalin genes. Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology. 1482 (1-2), 25–34.

Samuelson, L.C., Wiebauer, K., Snow, C.M. & Meisler, M.H. (1990) Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. Molecular and Cellular Biology. 10 (6), 2513–2520.

Sánchez, D., Ganfornina, M.D., Gutiérrez, G. & Marín, A. (2003) Exon-intron structure and evolution of the Lipocalin gene family. Molecular biology and evolution. 20 (5), 775–783.

Sanderson, M.J. (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics. 19 (2), 301–302.

Sandve, S.R., Rudi, H., Asp, T. & Rognli, O.A. (2008) Tracking the evolution of a cold stress associated gene family in cold tolerant grasses. BMC Evolutionary Biology. 8 (1), 245.

Sanger, F., Air, G.M., Barrell, B.G. & Brown, N.L. (1977) Nucleotide sequence of bacteriophage X174 DNA. Nature.

Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchison III, C.A., Slocombe, P.M. & Smith, M. (1978) The nucleotide sequence of bacteriophage ϕX174. Journal of Molecular Biology. 125 (2), 225–246.

Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. & Petersen, G.B. (1982) Nucleotide sequence of bacteriophage λ DNA. Journal of Molecular Biology. 162 (4), 729–773.

Sankoh, O. & IJsselmuiden, C. (2011) Sharing research data to improve public health: a perspective from the global south. The Lancet. 378 (9789), 401–402.

Santos, J.C. (2012) Fast molecular evolution associated with high active metabolic rates in poison frogs. Molecular biology and evolution. 29 (8), 2001–2018.

Savage, C.J. & Vickers, A.J. (2009) Empirical study of data sharing by authors publishing in PLoS journals. Chris Mavergames (ed.). PLoS One. 4 (9), e7078–e7078.

Saxonov, S., Berg, P. & Brutlag, D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences. 103 (5), 1412–1417.

Saze, H., Kitayama, J., Takashima, K., Miura, S., Harukawa, Y., Ito, T. & Kakutani, T. (2013) Mechanism for full-length RNA processing of Arabidopsis genes containing intragenic heterochromatin. Nature Communications. 4.

Scally, A., Dutheil, J.Y., Hillier, L.W., Jordan, G.E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S.H., Schwalie, P.C., Tang, Y.A., Ward, M.C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L.N., et al. (2012) Insights into hominid evolution from the gorilla genome sequence. Nature. 483 (7388), 169–175.

Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L. & Nolan, G.P. (2010) Computational solutions to large-scale data management and analysis. Nature Reviews Genetics. 11 (9), 647–657.

Schatz, M.C., Langmead, B. & Salzberg, S.L. (2010) Cloud computing and the DNA data race. Nat Biotech. 28 (7), 691–693.

Schluter, D., Marchinko, K.B., Barrett, R.D.H. & Rogers, S.M. (2010) Natural selection and the genetics of adaptation in threespine stickleback. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. 365 (1552), 2479–2486.

Schmegner, C., Hameister, H., Vogel, W. & Assum, G. (2007) Isochores and replication time zones: a perfect match. Cytogenetic and genome research. 116 (3), 167–172.

Schmid, R. & Blaxter, M.L. (2008) annot8r: GO, EC and KEGG annotation of EST datasets. BMC bioinformatics 9 (1) p.180.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., et al. (2010) Genome sequence of the palaeopolyploid soybean. *Nature*. 463 (7278), 178–183.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., et al. (2010) Genome sequence of the palaeopolyploid soybean. Nature. 463 (7278), 178–183.

Schneider, G.F. & Dekker, C. (2012) DNA sequencing with nanopores. Nat Biotech. 30 (4), 326–328.

Schneider, M.V., Watson, J., Attwood, T., Rother, K., Budd, A., McDowall, J., Via, A., Fernandes, P., Nyronen, T., Blicher, T., Jones, P., Blatter, M.-C., De Las Rivas, J., Judge, D.P., van der Gool, W. & Brooksbank, C. (2010) Bioinformatics training: a review of challenges, actions and support requirements. Briefings in bioinformatics. 11 (6), bbq021–551.

Schofield, P.N., Bubela, T., Weaver, T., Portilla, L., Brown, S.D., Hancock, J.M., Einhorn, D., Tocchini-Valentini, G., de Angelis, M.H. & Rosenthal, N. (2009) Post-publication sharing of data and tools. Nature. 461 (7261), 171–173.

Schueler, M.G., Higgins, A.W., Rudd, M.K., Gustashaw, K. & Willard, H.F. (2001) Genomic and genetic definition of a functional human centromere. Science. 294 (5540), 109–115.

Schw, T., Schwander, T., er, Lo, N., Beekman, M., Oldroyd, B.P. & Keller, L. (2010) Nature versus nurture in social insect caste differentiation. Trends in Ecology & Evolution. 25 (5), 275–282.

Schwaiger, M., Stadler, M.B., Bell, O., Kohler, H., Oakeley, E.J. & Schübeler, D. (2009) Chromatin state marks cell-type- and gender-specific replication of the Drosophila genome. Genes & Development. 23 (5), 589–601.

Schwartz SH et al. 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome research. 18:88-103.

Searle, S.M., Gilbert, J., Iyer, V. & Clamp, M. (2004) The otter annotation system. Genome research. 14 (5), 963–970.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T.C., Trask, B., Patterson, N., et al. (2004) Large-scale copy number polymorphism in the human genome. Science (New York, N.Y.). 305 (5683), 525–528.

Seegolam, A. & Usmani, K.A. (2014) 'Understanding the maturity of EU code of conduct on data centres: A Mauritian case study explained', in IST-Africa Conference Proceedings, 2014. [Online]. 1 January 2014 IEEE. pp. 1–16.

Seehausen, O., Terai, Y., Magalhaes, I.S., Carleton, K.L., Mrosso, H.D.J., Miyagi, R., van der Sluijs, I., Schneider, M. V, Maan, M.E., Tachida, H., Imai, H. & Okada, N.

(2008) Speciation through sensory drive in cichlid fish. Nature. 455 (7213), 620–626.

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:R59.

Sella, G., Petrov, D.A., Przeworski, M. & Andolfatto, P. (2009) Pervasive Natural Selection in the Drosophila Genome? Michael W Nachman (ed.). PLoS genetics. 5 (6), e1000495.

Sen, K. & Ghosh, T.C. (2013) Pseudogenes and their composers: delving in the 'debris' of human genome. Briefings in functional genomics. 12 (6), 536–547.

Shapiro, M.D., Kronenberg, Z., Li, C., Domyan, E.T., Pan, H., Campbell, M., Tan, H., Huff, C.D., Hu, H., Vickrey, A.I., Nielsen, S.C.A., Stringham, S.A., Hu, H., Willerslev, E., Gilbert, M.T.P., Yandell, M., Zhang, G. & Wang, J. (2013) Genomic diversity and evolution of the head crest in the rock pigeon. Science (New York, N.Y.). 339 (6123), 1063–1067.

Sharp PA. 1985. On the origin of RNA splicing and introns. Cell. 42:397-400.

Sharp, P.A. (1985) On the origin of RNA splicing and introns. Cell. 42 (2), 397–400.

Sharpton TJ, Neafsey DE, Galagan JE and Taylor JW. 2008. Mechanisms of intron gain and loss in Cryptococcus. Genome biology. 9:R24.

Sharpton, T.J., Stajich, J.E. & Rounsley, S.D. (2009) Comparative genomic analyses of the human fungal pathogens Coccidioides and their relatives. Genome.

Sheik, C.S., Beasley, W.H., Elshahed, M.S., Zhou, X., Luo, Y. & Krumholz, L.R. (2011) Effect of warming and drought on grassland microbial communities. ISME Journal. 5 (10), 1692–1700.

Shepard S, McCreary M and Fedorov A. 2009. The peculiarities of large intron splicing in animals. A. Christoffels, ed. PloS one. 4:e7853.

Shi, J., Wolf, S.E., Burke, J.M., Presting, G.G., Ross-Ibarra, J. & Dawe, R.K. (2010) Widespread Gene Conversion in Centromere Cores Harmit S Malik (ed.). PLoS Biology. 8 (3), e1000327.

Shiu, S.-H., Byrnes, J.K., Pan, R., Zhang, P. & Li, W.-H. (2006) Role of positive selection in the retention of duplicate genes in mammalian genomes. Proceedings of the National Academy of Sciences. 103 (7), 2232–2236.

Simonson, T.S., Yang, Y., Huff, C.D., Yun, H., Qin, G., Witherspoon, D.J., Bai, Z., Lorenzo, F.R., Xing, J., Jorde, L.B., Prchal, J.T. & Ge, R. (2010) Genetic evidence for high-altitude adaptation in Tibet. Science (New York, N.Y.). 329 (5987), 72–75.

Simpson, J.T. (2013) Exploring Genome Characteristics and Sequence Quality Without a Reference. arXiv.org. q-bio.GN.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. & Birol, İ. (2009) ABySS: a parallel assembler for short read sequence data. Genome research. 19 (6), 1117–1123.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., Chinwalla, A., Delehaunty, A., Delehaunty, K., Du, H., Fewell, G., Fulton, L., Fulton, R., Graves, T., Hou, S.-F., et al. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 423 (6942), 825–837.

Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Current Opinion in Genetics & Development. 9:657-663.

Smit AFA, Hubley R and Green P. 2004. RepeatMasker Open-3.0.

Smit, A.F.A., Hubley, R. & Green, P. (2004) RepeatMasker Open-3.0.

Smith, A., Willassen, N.P. & Våge, D.I. (2013) 'ELIXIR: A distributed life sciences infrastructure supporting innovation in marine sciences BT - Marine biotechnology in the European research area Challenges and opportunities for Europe. Final CSA MarineBiotech Conference. Royal Flemish Academy of Belg', in J Mees (ed.) Marine biotechnology in the European research area Challenges and opportunities for Europe. Final CSA MarineBiotech Conference. Royal Flemish Academy of Belgium for Science and the Arts, Brussels, Belgium, - March . VLIZ Special Publication. [Online]. VLIZ. p. 57.

Smith, A.F.A., Hubley, R. & Green, P. (2010) RepeatMasker.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N., Whetzel, P.L. & Lewis, S. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotech. 25 (11), 1251–1255.

Soltis, D.E. & Soltis, P.S. (1999) Polyploidy: recurrent formation and genome evolution. Trends in Ecology & Evolution. 14 (9), 348–352.

Spady, T.C., Parry, J.W.L., Robinson, P.R., Hunt, D.M., Bowmaker, J.K. & Carleton, K.L. (2006) Evolution of the cichlid visual palette through ontogenetic subfunctionalization of the opsin gene arrays. Molecular biology and evolution. 23 (8), 1538–1547.

Spingola, M., Grate, L., Haussler, D. & Ares, M. (1999) Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. RNA. 5 (2), 221–234.

Spudich, G., Fernández-Suárez, X.M., Birney, E., Fernandez-Suarez, X.M. & Birney, E. (2007) Genome browsing with Ensembl: a practical overview. Briefings in Functional Genomics and Proteomics. 6 (3), 202–219.

Spudich, G.M. & Fernández-Suárez, X.M. (2010) Touring Ensembl: a practical guide to genome browsing. BMC Genomics. 11 (1), 295.

Stabenau A et al. 2004. The Ensembl core software libraries. Genome research. 14:929-33.

Stabenau, A., McVicker, G., Melsopp, C., Proctor, G., Clamp, M. & Birney, E. (2004) The Ensembl core software libraries. Genome research. 14 (5), 929–933.

Staden, R. (1977) Sequence data handling by computer. Nucleic Acids Research. 4 (11), 4037–4051.

Staden, R. (1978) Further procedures for sequence analysis by computer. Nucleic Acids Research. 5 (3), 1013–1016.

Staden, R. (1979) A strategy of DNA sequencing employing computer programs. Nucleic Acids Research. 6 (7), 2601–2610.

Staden, R. (1996) The Staden sequence analysis package. Molecular biotechnology. 5 (3), 233–241.

Staden, R., Beal, K.F. & Bonfield, J.K. (1999) 'The Staden Package, 1998 BT - Bioinformatics Methods and Protocols', in Bioinformatics Methods and Protocols. [Online]. New Jersey: Humana Press. pp. 115–130.

Stajich JE et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome research. 12:1611-8.

Stajich JE, Dietrich FS and Roy SW. 2007. Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. Genome biology. 8:R223.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E. & al, et (2002) The Bioperl Toolkit: Perl Modules for the Life Sciences. Genome 1–8.

Stajich, J.E., Dietrich, F.S. & Roy, S.W. (2007) Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. Genome Biol. 8 (10), R223–R223.

Stalker, J., Gibbins, B., Meidl, P., Smith, J., Spooner, W., Hotz, H.-R. & Cox, A. V (2004) The Ensembl Web site: mechanics of a genome browser. Genome research. 14 (5), 951–955.

Stamatakis, A., Ludwig, T. & Meier, H. (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 21 (4), 456–463.

Stamatoyannopoulos, J.A., Adzhubei, I. & Thurman, R.E. (2009) Human mutation rate associated with DNA replication timing. Nature

Stapley, J., Reger, J., Feulner, P.G.D., Smadja, C., Galindo, J., Ekblom, R., Bennison, C., Ball, A.D., Beckerman, A.P. & Slate, J. (2010) Adaptation genomics: the next generation. Trends in Ecology & Evolution. 25 (12), 705–712.

Stein LD et al. 2003. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. PLoS biology. 1:E45.

Stein, L. (2002) Creating a bioinformatics nation. Nature. 417 (6885), 119–120.

Stein, L.D. (2008) Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. Nature Reviews Genetics. 9 (9), 678–688.

Stein, L.D. (2010) The case for cloud computing in genome informatics. Genome Biol.

Stein, L.D. (2010) The case for cloud computing in genome informatics. Genome Biol. 11 (5), 207.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A. & Lewis, S. (2002) The generic genome browser: a building block for a model organism system database. Genome research. 12 (10), 1599–1610.

Steinbock, L.J., Lucas, A., Otto, O. & Keyser, U.F. (2012) Voltage-driven transport of ions and DNA through nanocapillaries Victor M Ugaz (ed.). Electrophoresis. 33 (23), 3480–3487.

Steiper, M.E. & Seiffert, E.R. (2012) Evidence for a convergent slowdown in primate molecular rates and its implications for the timing of early primate evolution. Proceedings of the National Academy of Sciences of the United States of America. 109 (16), 6006–6011.

Stevens, R., Goble, C.A. & Bechhofer, S. (2000) Ontology-based knowledge representation for bioinformatics. Briefings in bioinformatics. 1 (4), 398–414.

Stevens, R.D., Robinson, A.J. & Goble, C.A. (2003) myGrid: personalised bioinformatics on the information grid. Bioinformatics. 19 (suppl 1), i302–i304.

Stodden, V., Guo, P. & Ma, Z. (2013) Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals Dmitri Zaykin (ed.). PLoS One. 8 (6), e67111.

Strozzi, F. & Aerts, J. (2011) A Ruby API to query the Ensembl database for genomic features. Bioinformatics. 27 (7), 1013–1014.

Sudmant, P.H., Huddleston, J., Catacchio, C.R., Malig, M., Hillier, L.W., Baker, C., Mohajeri, K., Kondova, I., Bontrop, R.E., Persengiev, S., Antonacci, F., Ventura, M., Prado-Martinez, J., Project, G.A.G., Marques-Bonet, T. & Eichler, E.E. (2013) Evolution and diversity of copy number variation in the great ape lineage. Genome research. 23 (9), 1373–1382.

Sullivan, A.E., Silver, R.M., LaCoursiere, D.Y., Porter, T.F. & Branch, D.W. (2004) Recurrent Fetal Aneuploidy and Recurrent Miscarriage. Obstetrics & Gynecology. 104 (4), 784–788.

Sun, X., Wahlstrom, J. & Karpen, G. (1997) Molecular structure of a functional Drosophila centromere. Cell. 91 (7), 1007–1019.

Sun, Y., Whittle, C.A., Corcoran, P. & Johannesson, H. (2015) Intron evolution in Neurospora: the role of mutational bias and selection. Genome research. 25 (1), 100–110.

Swanson-Wagner, R.A., Eichten, S.R., Kumari, S., Tiffin, P., Stein, J.C., Ware, D. & Springer, N.M. (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome research. 20 (12), 1689–1699.

Swanson, W.J. (2003) Adaptive evolution of genes and gene families. Current Opinion in Genetics & Development. 13 (6), 617–622.

Symington, L.S. & Petes, T.D. (1988) Meiotic recombination within the centromere of a yeast chromosome. Cell. 52 (2), 237–240.

Taft, R.J., Pheasant, M. & Mattick, J.S. (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays. 29 (3), 288–299.

Talbert, P.B. & Henikoff, S. (2010) Centromeres Convert but Don’t Cross. PLoS Biology. 8 (3), e1000326.

Tan, T.W., Tong, J.C., Khan, A.M., de Silva, M., Lim, K.S. & Ranganathan, S. (2010) Advancing standards for bioinformatics activities: persistence, reproducibility, disambiguation and Minimum Information About a Bioinformatics investigation (MIABi). BMC Genomics. 11 (Suppl 4), S27.

Tateno, Y., Imanishi, T. & Miyazaki, S. (2002) DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic acids ….

Team, R.C. (2012) R: A Language and Environment for Statistical Computing.

Teshima, K.M. & Innan, H. (2004) The effect of gene conversion on the divergence between duplicated genes. Genetics. 166 (3), 1553–1560.

Thomas, J.H. (2008) Genome evolution in Caenorhabditis. Briefings in Functional Genomics and Proteomics. 7 (3), 211–216.

Tiley, G.P. & Burleigh, G. (2015) The relationship of recombination rate, genome structure, and patterns of molecular evolution across angiosperms. *BMC Evolutionary Biology*. 15 (1), 194.

Tirmizi, S.H., Aitken, S. & Moreira, D.A. (2011) Mapping between the OBO and OWL ontology languages. J Biomedical ….

Turner, T.L., Bourne, E.C., Von Wettberg, E.J., Hu, T.T. & Nuzhdin, S. V (2010) Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. Nature Genetics. 42 (3), 260–263.

Vaishali Katju, U.B. (2010) Genomic and Population-Level Effects of Gene Conversion in Caenorhabditis Paralogs. Genes. 1 (3), 452–468.

Van Dongen, S. (2008) Graph Clustering Via a Discrete Uncoupling Process. SIAM Journal on Matrix Analysis and Applications. 30 (1), 121.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., Pruss, D., Pindo, M., Fitzgerald, L.M., Vezzulli, S., Reid, J., Malacarne, G., Iliev, D., Coppola, G., Wardell, B., Micheletti, D., Macalma, T., Facci, M., Mitchell, J.T., Perazzolli, M., et al. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. Brian Dilkes (ed.). PLoS One. 2 (12), e1326–e1326.

Venter JC et al. 2001. The sequence of the human genome. Science (New York, N.Y.). 291:1304-51.

Venter, J.C. (2001) The sequence of the human genome. Science. 291 (5507), 1304–1351.

Vera, G., Jansen, R.C. & Suppi, R.L. (2008) R/parallel – speeding up bioinformatics analysis with R. BMC Bioinformatics 9 (1) p.390.

Viguera, E., Canceill, D. & Ehrlich, S.D. (2001) Replication slippage involves DNA polymerase pausing and dissociation. The EMBO Journal. 20 (10), 2587–2595.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. & Birney, E. (2008) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome research. 19 (2), 327–335.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. & Birney, E. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome research. 19 (2), 327–335.

Vinogradov, A.E. (1999) Intron–Genome Size Relationship on a Large Evolutionary Scale. Journal of Molecular Evolution. 49 (3), 376–384.

Voelker, R.B., Erkelenz, S., Reynoso, V., Schaal, H. & Berglund, J.A. (2012) Frequent Gain and Loss of Intronic Splicing Regulatory Elements during the Evolution of Vertebrates. Genome biology and evolution. 4 (7), 659–674.

Vogel, K.J. & Moran, N.A. (2013) Functional and evolutionary analysis of the genome of an obligate fungal symbiont. Genome biology and evolution. 5 (5), 891–904.

Wagner A. 2005. Energy constraints on the evolution of gene expression. Molecular biology and evolution. 22:1365-74.

Wagner, A. (2005) Energy constraints on the evolution of gene expression. Molecular biology and evolution. 22 (6), 1365–1374.

Wajid, B. & Serpedin, E. (2012) Review of General Algorithmic Features for Genome Assemblers for Next Generation Sequencers. Genomics, Proteomics & Bioinformatics. 10 (2), 58–73.

Walker, J.A., Konkel, M.K., Ullmer, B. & Monceaux, C.P. (2012) Orangutan Alu quiescence reveals possible source element: support for ancient backseat drivers. Mob DNA.

Wall, D.P., Kudtarkar, P., Fusaro, V.A., Pivovarov, R., Patil, P. & Tonellato, P.J. (2010) Cloud computing for comparative genomics. BMC bioinformatics. 11 (1), 259.

Wang, D. & Yu, J. (2011) Both size and GC-content of minimal introns are selected in human populations. Thomas Mailund (ed.). PLoS One. 6 (3), e17945.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., et al.

(2008) The diploid genome sequence of an Asian individual. Nature 456 (7218), 60–65.

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., Yue, Z., Cong, L., Shang, H., Zhu, S., Zou, C., Li, Q., Yuan, Y., Lu, C., Wei, H., Gou, C., Zheng, Z., Yin, Y., Zhang, X., et al. (2012) The draft genome of a diploid cotton Gossypium raimondii. Nature genetics. 44 (10), 1098–1103.

WANG, L. & Jiang, T. (1994) On the Complexity of Multiple Sequence Alignment. dx.doi.org. 1 (4), 337–348.

Wang, X.V., Blades, N., Ding, J., Sultana, R. & Parmigiani, G. (2012) Estimation of sequencing error rates in short reads. BMC bioinformatics. 13 (1), 185.

Ward, R.M., Schmieder, R., Highnam, G. & Mittelman, D. (2013) Big data challenges and opportunities in high-throughput sequencing. Systems Biomedicine. 1 (1), 29–34.

Warren, R.L., Sutton, G.G., Jones, S.J.M. & Holt, R.A. (2007) Assembling millions of short DNA sequences using SSAKE. Bioinformatics (Oxford, England). 23 (4), 500–501.

Watanabe, M., Hiraide, K. & Okada, N. (2007) Functional diversification of kir7.1 in cichlids accelerated by gene duplication. Gene. 399 (1), 46–52.

Watanabe, Y., Abe, T., Ikemura, T. & Maekawa, M. (2009) Relationships between replication timing and GC content of cancer-related genes on human chromosomes 11q and 21q. Gene. 433 (1-2), 26–31.

Weadick, C.J. & Chang, B.S.W. (2012) Complex patterns of divergence among green-sensitive (RH2a) African cichlid opsins revealed by Clade model analyses. BMC Evolutionary Biology. 12206.

Weber, C.C., Boussau, B., Romiguier, J., Jarvis, E.D. & ELLEGREN, H. (2014) Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. Genome biology. 15 (12), 549.

Wehe, A., Bansal, M.S., Burleigh, J.G. & Eulenstein, O. (2008) DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics (Oxford, England). 24 (13), 1540–1541.

Wehe, A., Chang, W.-C., Eulenstein, O. & Aluru, S. (2010) A scalable parallelization of the gene duplication problem. Journal of Parallel and Distributed Computing. 70 (3), 237–244.

Weinstock, G.M., Robinson, G.E., Gibbs, R.A., Weinstock, G.M., Weinstock, G.M., Robinson, G.E., Worley, K.C., Evans, J.D., Maleszka, R., Robertson, H.M., Weaver, D.B., Beye, M., Bork, P., Elsik, C.G., Evans, J.D., Hartfelder, K., Hunt, G.J., Robertson, H.M., Robinson, G.E., et al. (2006) Insights into social insects from the genome of the honeybee Apis mellifera. Nature. 443 (7114), 931–949.

Weiss, M.M., Hermsen, M.A., Meijer, G.A., van Grieken, N.C., Baak, J.P., Kuipers, E.J. & van Diest, P.J. (1999) Comparative genomic hybridisation. Molecular Pathology. 52 (5), 243–251.

Wendel, J.F. (2000) 'Genome evolution in polyploids BT - Plant Molecular Evolution', in Plant Molecular Evolution. [Online]. Dordrecht: Springer Netherlands. pp. 225–249.

Weterings, E. & van Gent, D.C. (2004) The mechanism of non-homologous end-joining: a synopsis of synapsis. DNA repair. 3 (11), 1425–1435.

Wickham, H. (2007) Reshaping data with the reshape package. Journal of Statistical Software.

Wickham, H. (2009) ggplot2. New York, NY: Springer Science & Business Media.

Wickham, H. (2011) The split-apply-combine strategy for data analysis. Journal of Statistical Software.

Wicks, P., Massagli, M., Frost, J., Brownstein, C., Okun, S., Vaughan, T., Bradley, R. & Heywood, J. (2010) Sharing Health Data for Better Outcomes on PatientsLikeMe. Journal of Medical Internet Research. 12 (2), e19.

Willett, C.S. (2013) Gene conversion yields novel gene combinations in paralogs of GOT1 in the copepod Tigriopus californicus. BMC Evolutionary Biology. 13 (1), 148.

Wilming, L.G., Gilbert, J.G.R., Howe, K., Trevanion, S., Hubbard, T. & Harrow, J.L. (2008) The vertebrate genome annotation (Vega) database. Nucleic Acids Research. 36 (Database issue), D753–60.

Wilson, G. (2013) Software Carpentry: Lessons Learned. arXiv.org. cs.GL.

Wilson, G., Aruliah, D.A., Brown, C.T., Hong, N.P.C., Davis, M., Guy, R.T., Haddock, S.H.D., Huff, K., Mitchell, I.M., Plumbley, M., Waugh, B., White, E.P. & Wilson, P. (2012) Best Practices for Scientific Computing. arXiv.org. cs.MS.

Wilson, M.D., Cheung, J., Martindale, D.W., Scherer, S.W. & Koop, B.F. (2006) Comparative analysis of the paired immunoglobulin-like receptor (PILR) locus in six mammalian genomes: duplication, conversion, and the birth of new genes. Physiological …. 27 (3), 201–218.

Witherspoon, D.J. & Robertson, H.M. (2003) Neutral Evolution of Ten Types of mariner Transposons in the Genomes of Caenorhabditis elegans and Caenorhabditis briggsae. Journal of Molecular Evolution. 56 (6), 751–769.

Wolfe, K.H. (2015) Origin of the Yeast Whole-Genome Duplication. *PLoS Biology*. 13 (8), e1002221.

Wolfe, K.H. & Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature.

Wong, S. & Wolfe, K.H. (2005) Birth of a metabolic gene cluster in yeast by adaptive gene relocation. Nature genetics. 37 (7), 777–782.

Woodfine, K., Fiegler, H., Beare, D.M., Collins, J.E., McCann, O.T., Young, B.D., Debernardi, S., Mott, R., Dunham, I. & Carter, N.P. (2004) Replication timing of the human genome. Human Molecular Genetics. 13 (2), 191–202.

Wright, S.I. & Andolfatto, P. (2008) The Impact of Natural Selection on the Genome: Emerging Patterns in Drosophilaand Arabidopsis. Annu. Rev. Ecol. Evol. Syst. 39 (1), 193–213.

Wu, H.-J., Zhang, Z., Wang, J.-Y., Oh, D.-H., Dassanayake, M., Liu, B., Huang, Q., Sun, H.-X., Xia, R., Wu, Y., Wang, Y.Y.-N., Yang, Z., Liu, Y., Zhang, W., Zhang, H., Chu, J., Yan, C., Fang, S., Zhang, J., et al. (2012) Insights into salt tolerance from the genome of Thellungiella salsuginea. Proceedings of the National Academy of Sciences of the United States of America. 109 (30), 12219–12224.

Wu, J., Wang, Z., Shi, Z., Zhang, S., Ming, R., Zhu, S., Khan, M.A., Tao, S., Korban, S.S., Wang, H., Chen, N.J., Nishio, T., Xu, X., Cong, L., Qi, K., Huang, X., Wang, Y., Zhao, X., Wu, J., et al. (2013) The genome of the pear (Pyrus bretschneideri Rehd.). Genome research. 23 (2), 396–408.

Wu, X.-S., Xin, L., Yin, W.-X., Shang, X.-Y., Lu, L., Watt, R.M., Cheah, K.S.E., Huang, J.-D., Liu, D.-P. & Liang, C.-C. (2005) Increased efficiency of oligonucleotide-mediated gene repair through slowing replication fork progression. Proceedings of the National Academy of Sciences. 102 (7), 2508–2513.

Wurm, Y., Wang, J., Riba-Grognuz, O. & al, et (2011) The genome of the fire ant Solenopsis invicta. Proceedings of the

Xu, G., Guo, C., Shan, H. & Kong, H. (2012) Divergence of duplicate genes in exon-intron structure. Proceedings of the National Academy of Sciences of the United States of America. 109 (4), 1187–1192.

Yandell M et al. 2006. Large-scale trends in the evolution of gene structures within 11 animal genomes. PLoS Computational Biology. 2:e15.

Yandell, M. & Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics. 13 (5), 329–342.

Yandell, M., Mungall, C.J., Smith, C., Prochnik, S., Kaminker, J., Hartzell, G., Lewis, S. & Rubin, G.M. (2006) Large-Scale Trends in the Evolution of Gene Structures within 11 Animal Genomes. PLoS Computational Biology. 2 (3), e15.

Yang, Z. & Huang, J. (2011) De novo origin of new genes with introns in Plasmodium vivax. FEBS Letters. 585 (4), 641–644.

Yang, Z. & Yoder, A.D. (2003) Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene Loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Systematic Biology. 52 (5), 705–716.

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G.R.S., Ruffier, M., Taylor, K., Vullo, A. & Flicek, P. (2014) The Ensembl REST API: Ensembl Data for Any Language. Bioinformatics. 31 (1), btu613–145.

Ye, L., Hillier, L.W., Minx, P., Thane, N., Locke, D.P., Martin, J.C., Chen, L., Mitreva, M., Miller, J.R., Haub, K. V, Dooling, D.J., Mardis, E.R., Wilson, R.K., Weinstock, G.M. & Warren, W.C. (2011) A vertebrate case study of the quality of assemblies derived from next-generation sequences. Genome biology. 12 (3), R31.

Yenerall, P. & Zhou, L. (2012) Identifying the mechanisms of intron gain: progress and trends. Biol Direct. 7 (1), 29.

Yip, K.Y., Cheng, C. & Gerstein, M. (2013) Machine learning and genome annotation: a match meant to be? Genome Biol. 14 (5), 205.

Yoon, S.-R., Dubeau, L., de Young, M., Wexler, N.S. & Arnheim, N. (2003) Huntington disease expansion mutations in humans can occur before meiosis is completed. Proceedings of the National Academy of Sciences. 100 (15), 8834–8838.

Young, J.M., Endicott, R.M., Parghi, S.S., Walker, M., Kidd, J.M. & Trask, B.J. (2008) Extensive copy-number variation of the human olfactory receptor gene family. American journal of human genetics. 83 (2), 228–242.

Yu, J., Yang, Z., Kibukawa, M., Paddock, M., Passey, D.A. & Wong, G.K.S. (2002) Minimal introns are not ‘junk’. Genome research. 12 (8), 1185–1189.

Zerbino, D.R. & 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. & Birney, E. (2009) Pebble and Rock Band: Heuristic Resolution of Repeats and Scaffolding in the Velvet Short-Read de Novo Assembler Steven L Salzberg (ed.). PLoS One. 4 (12), e8407.

Zhang, F., Gu, W., Hurles, M.E. & Lupski, J.R. (2009) Copy Number Variation in Human Health, Disease, and Evolution. Annual Review of Genomics and Human Genetics. 10 (1), 451–481.

Zhang, J. (2003) Evolution by gene duplication: an update. Trends in Ecology & Evolution. 18 (6), 292–298.

Zhang, J., Zhang, Y. & Rosenberg, H.F. (2002) Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nature genetics. 30 (4), 411–415.

Zhang, K.K., Arabnia, H.R., Wang, Y. & Deng, Y. (2013) The current trend of genomics research for human diseases. BMC medical genomics. 6 Suppl 1S1.

Zhang, Q. & Edwards, S. V (2012) The Evolution of Intron Size in Amniotes: A Role for Powered Flight? Genome biology and evolution. 4 (10), 1033–1043.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J. & Shen, B. (2011) A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies I King Jordan (ed.). PLoS One. 6 (3), e17915–e17915.

Zhang, Z. & Gerstein, M. (2003) Identification and characterization of over 100 mitochondrial ribosomal protein pseudogenes in the human genome☆. Genomics. 81 (5), 468 – 480.

Zhu L et al. 2009. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC genomics. 10:47.

Zhu, L., Zhang, Y., Zhang, W., Yang, S., Chen, J.-Q. & Tian, D. (2009) Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics. 10 (1), 47.

Zhu, T., Niu, D.-K. & Tao Zhu, D.-K.N. (2013) Mechanisms of intron loss and gain in the fission yeast Schizosaccharomyces. Jürg Bähler (ed.). PLoS One. 8 (4), e61683.

Zhuang YA, Goldstein AM and Weiner AM. 1989. UACUAAC is the preferred branch site for mammalian mRNA splicing. Proceedings of the National Academy of Sciences of the United States of America. 86:2752-6.

Zid, M. & Drouin, G. (2013) Gene conversions are under purifying selection in the carcinoembryonic antigen immunoglobulin gene families of primates. Genomics. 102 (4), 301–309.

Zozulya, S., Echeverri, F. & Nguyen, T. (2001) The human olfactory receptor repertoire. Genome Biol. 2 (6), RESEARCH0018.

[This page is left intentionally blank]

APPENDICES

Appendix 1

Appendix 1.1 - The Reniform Reflecting Superposition Compound Eyes of Nephrops norvegicus: Optics, Susceptibility to Light-Induced Damage, Electrophysiology and a Ray Tracing Model

Please find the article - part of the *Advances in Marine Biology series (Volume 64) "The Ecology and Biology of Nephrops norvegicus"* - at the following URL:

http://www.sciencedirect.com/science/article/pii/B9780124104662000042

Appendix 2

Appendix 2.1 - HTTP methods

The HTTP POST, GET, PUT and DELETE methods relate to the four basic functions of persistent storage; CREATE, READ, UPDATE and DELETE (CRUD)(Martin, 1983). They allow for the creation of new data, the retrieval of existing data, the updating of existing data, or removal of existing data. Without a persistent storage mechanism the data would be stored temporarily in memory and lost when the power was removed.

Appendix 2.2 - Format of URI conventions

Use URI conventions in the following format via the GET method:

```
/get_adaptor/human/core/gene/fetch_by_stable_id/ENSG0000013
9618/description
/get_adaptor/mouse/core/transcript/list_stable_ids
/get_adaptor/gorilla/core/translation/list_dbIDs
/get_adaptor/zebrafish/core/exon/fetch_by_dbID/235872
```

The get_adaptor method takes three arguments (`/species/group/adaptor/`) and can call null ID methods such as `/list_stable_ids`, single ID methods such as `/fetch_by_stable_id/ENSG00000139618` or a method on a method id e.g. `/fetch_by_stable_id/ENSG00000139618/description`. This will return the information you require from the Ensembl databases in a format depending on the requested Content-Type e.g. `application/json`, `text/xml` or `text/html`.

You can use the POST method for more advanced queries that may require an array of elements or objects as input to a function, using JSON. The POST method only accepts and returns JSON.

You should use the following URI convention:

```
/get_adaptor/multi/ontology/ontologyterm/fetch_all_by_dbID_
list
```

Pass the JSON data (list of dbIDs) in using the following format:

```
{"args" : ["123", "234", "345", "456"]}
```

Some methods take objects as arguments. Such as:

```
/get_adaptor/human/core/sequence/fetch_by_Slice_start_end_s
trand
/get_adaptor/human/core/assemblymapper/fetch_by_CoordSystem
s
```

You can use the GET method to retrieve an object and then pass that back in using JSON in the following format (`$json_object` represents the retrieved object):

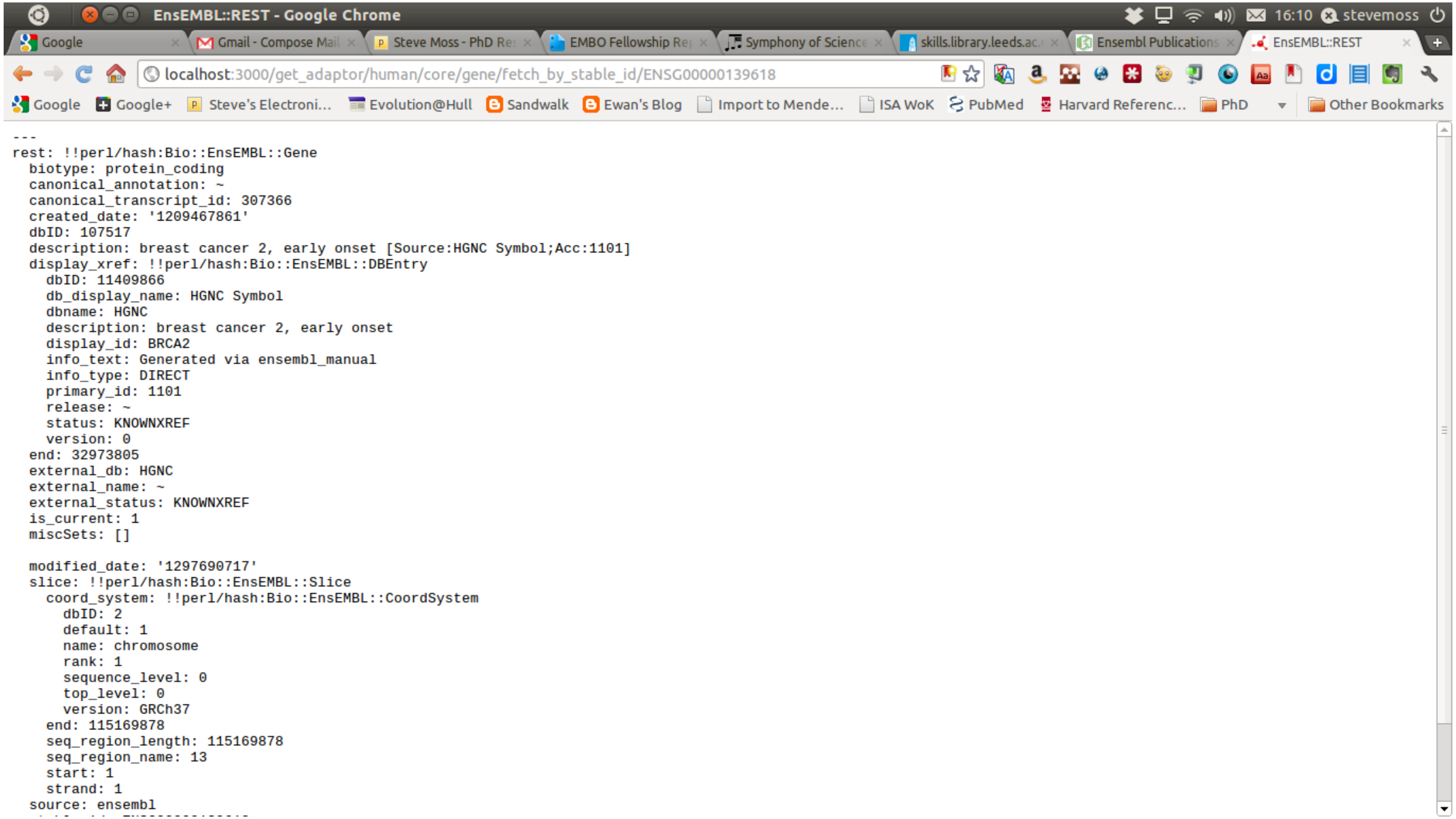
```
# representation of the slice object
my $json_object = "{"seq_region_name" : "20", "strand" : 1,
"coord_system" : {"dbID" : "2", "top_level" : 0, "version" :
"GRCh37",   "name"   :   "chromosome",   "default"   :   1,
"sequence_level" : 0, "rank" : "1", "seq_region_length" :
"63025520", "end" : "63025520", "start" : 1}}";

# json query to be passed
my $json_text = "{"args" : [$json_object, "1", "1000", "-
1"]}";
```

or:

```
# pass the arguments in this format if using a REST browser
plugin, command line
# or other non-programmatic method
{"args"  :  [{"seq_region_name"  :  "20",  "strand"  :  1,
"coord_system" : {"dbID" : "2", "top_level" : 0, "version" :
"GRCh37",   "name"   :   "chromosome",   "default"   :   1,
"sequence_level" : 0, "rank" : "1", "seq_region_length" :
"63025520", "end" : "63025520", "start" : 1}}, "1", "1000",
"-1"]}
```

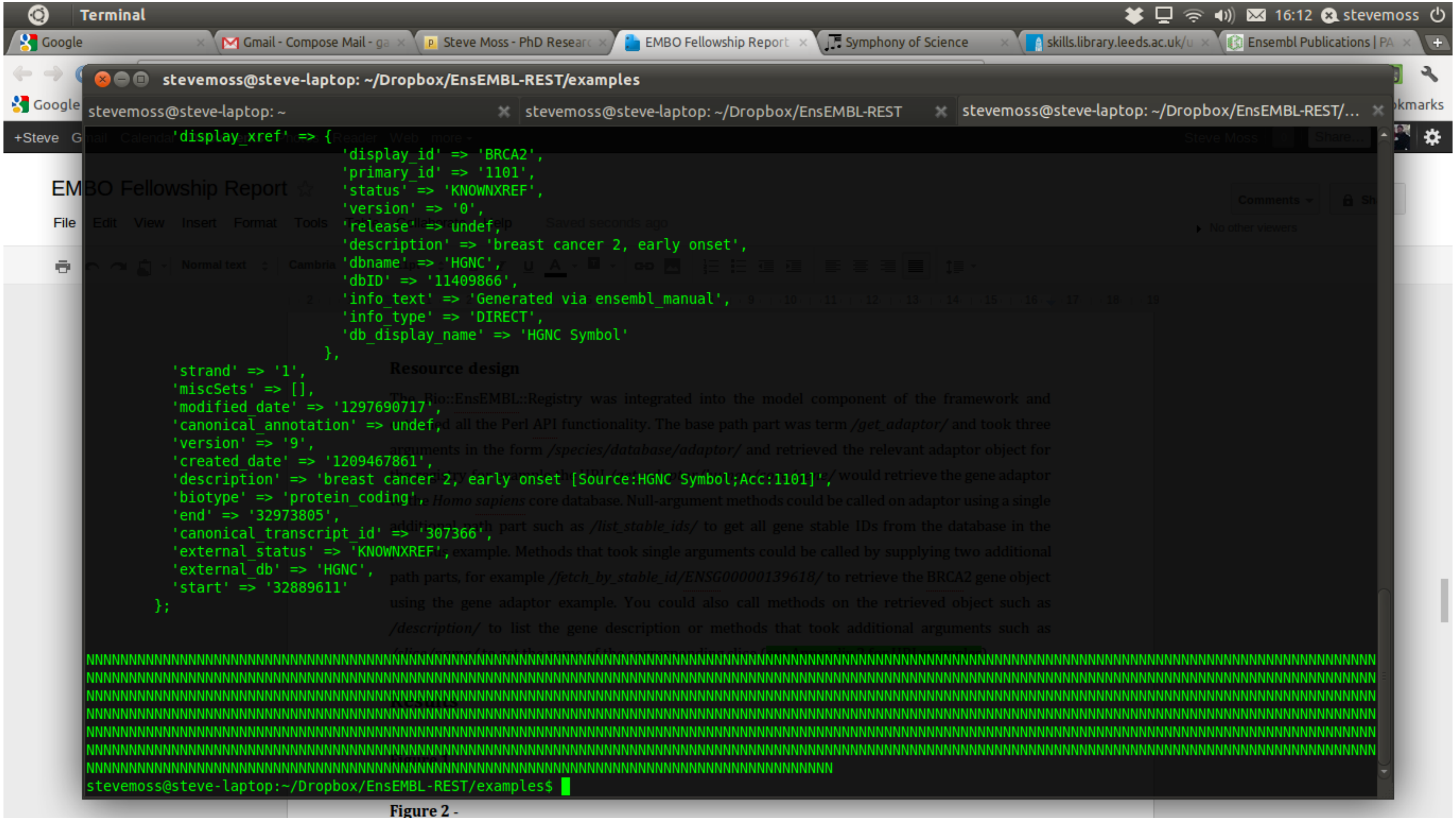
Because the GET method converts the Ensembl object into a null type JSON object during serialization, the Ensembl RESTful WSF retrieves a new object based on the JSON data during the POST process and passes that for processing instead.

Appendix 2.3 – Screenshots highlighting examples of the Ensembl RESTful web service framework



```
---
rest: !!perl/hash:Bio::EnsEMBL::Gene
  biotype: protein_coding
  canonical_annotation: ~
  canonical_transcript_id: 307366
  created_date: '1209467861'
  dbID: 107517
  description: breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101]
  display_xref: !!perl/hash:Bio::EnsEMBL::DBEntry
    dbID: 11409866
    db_display_name: HGNC Symbol
    dbname: HGNC
    description: breast cancer 2, early onset
    display_id: BRCA2
    info_text: Generated via ensembl_manual
    info_type: DIRECT
    primary_id: 1101
    release: ~
    status: KNOWNXREF
    version: 0
  end: 32973805
  external_db: HGNC
  external_name: ~
  external_status: KNOWNXREF
  is_current: 1
  miscSets: []

  modified_date: '1297690717'
  slice: !!perl/hash:Bio::EnsEMBL::Slice
    coord_system: !!perl/hash:Bio::EnsEMBL::CoordSystem
      dbID: 2
      default: 1
      name: chromosome
      rank: 1
      sequence_level: 0
      top_level: 0
      version: GRCh37
    end: 115169878
    seq_region_length: 115169878
    seq_region_name: 13
    start: 1
    strand: 1
  source: ensembl
```

Supplementary Figure 2.1 - The Ensembl RESTful web service being used via the browser to return a gene object from the Ensembl MySQL core database for the BRCA2 gene in the YAML format.


Terminal
16:12 stevemoss
Google
Gmail - Compose Mail - ga
Steve Moss - PhD Researc
EMBO Fellowship Report
Symphony of Science
skills.library.leeds.ac.uk/u
Ensembl Publications | PA
stevemoss@steve-laptop: ~/Dropbox/EnsEMBL-REST/examples
stevemoss@steve-laptop: ~
stevemoss@steve-laptop: ~/Dropbox/EnsEMBL-REST
stevemoss@steve-laptop: ~/Dropbox/EnsEMBL-REST/...
EMBO Fellowship Report
File
Edit
View
Insert
Format
Tools

```
          'display_xref' => {
                              'display_id' => 'BRCA2',
                              'primary_id' => '1101',
                              'status' => 'KNOWNXREF',
                              'version' => '0',
                              'release' => undef,
                              'description' => 'breast cancer 2, early onset',
                              'dbname' => 'HGNC',
                              'dbID' => '11409866',
                              'info_text' => 'Generated via ensembl_manual',
                              'info_type' => 'DIRECT',
                              'db_display_name' => 'HGNC Symbol'
                            },
          'strand' => '1',
          'miscSets' => [],
          'modified_date' => '1297690717',
          'canonical_annotation' => undef,
          'version' => '9',
          'created_date' => '1209467861',
          'description' => 'breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101]',
          'biotype' => 'protein_coding',
          'end' => '32973805',
          'canonical_transcript_id' => '307366',
          'external_status' => 'KNOWNXREF',
          'external_db' => 'HGNC',
          'start' => '32889611'
        };

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
stevemoss@steve-laptop:~/Dropbox/EnsEMBL-REST/examples$ 
```

Supplementary Figure 2.2 - A Perl example script that retrieves the BRCA2 gene using the HTTP GET method and the HTTP Content-Type header application/json. The slice of the gene object is returned to the Ensembl RESTful web service using HTTP POST and the sequence of the gene from 1 to 1,000 bases from the anti-sense strand is returned

Appendix 3

Appendix 3.1 - Comparative Analysis Of Teleost Genome Sequences Reveals An Ancient Intron Size Expansion In The Zebrafish Lineage Supplementary Information

Please find the article, published within the journal *Genome Biology and Evolution*, at the following URL:

http://gbe.oxfordjournals.org/content/3/1187.full

Appendix 4

Appendix 4.1 – Table of resources used in assessing assembly and annotation quality in primates

Supplementary Table 4.1 - Resources used in assessing assembly and annotation quality of the primates.

Species	Ensembl	GenBank	Consortium
Callithrix jacchus	Callithrix jacchus-3.2.1	GCF_000004665.1	The Genome Institute at Washington University
Gorilla gorilla	gorGor3.1	GCF_000151905.1	Wellcome Trust Sanger Institute
Homo sapiens	GRCh37.p6	GCA_000001405.7	Genome Reference Consortium
Macaca mulatta	MMUL 1.0	GCF_000002255.2	Macaque Genome Sequencing Consortium
Microcebus murinus	micMur1	GCA_000165445.1	Mammalian Genome Project
Nomascus leucogenys	Nleu1.0	GCF_000146795.1	Gibbon Genome Sequencing Consortium
Otolemur garnettii	OtoGar3	GCF_000181295.1	Broad Institute of MIT and Harvard
Pan troglodytes	CHIMP2.1.4	GCF_000001515.5	Chimpanzee Sequencing and Analysis Consortium
Pongo abelii	PPYG2	GCF_000001545.4	Orangutan Genome Sequencing Consortium
Tarsius syrichta	tarSyr1	GCA_000164805.1	Mammalian Genome Project
Tupaia belangeri	tupBel1	GCA_000181375.1	Mammalian Genome Project

Appendix 4.2 – Table of resources used in assessing assembly and annotation quality in rodents

Supplementary Table 4.2 - Resources used in assessing assembly and annotation quality of the rodents.

Species	Ensembl	GenBank	Consortium
Cavia porcellus	cavPor3	GCF_000151735.1	Mammalian Genome Project
Dipodomys ordii	dipOrd1	GCA_000151885.1	Mammalian Genome Project
Mus musculus	NCBIM37	GCF_000001635.18	Genome Reference Consortium
Oryctolagus cuniculus	oryCun2	GCF_000003625.2	Mammalian Genome Project
Rattus norvegicus	RGSC 3.4	GCF_000001895.3	Rat Genome Project
Ictidomys tridecemlineatus	speTri1	GCA_000181315.1	Mammalian Genome Project

Appendix 4.3 - CAFE consensus lambda values for individual gene family search

See the `primates_varied_indiv.lambda` file available at the following web address https://gist.github.com/gawbul/7119183.

Appendix 4.4 – Release 66 significant gene family descriptions

See the `r66_unique_family_descriptions.csv` file available at the following web address https://gist.github.com/gawbul/3e0ddd8ef60507223055.

Appendix 4.5 - Release 67 significant gene family descriptions

See the `r67_unique_family_descriptions.csv` file available at the following web address https://gist.github.com/gawbul/bd3c4b70fea477b224e8.

Appendix 4.6 – Table of matching gene families recovered from the Dumas analyses

Supplementary Table 4.3 - Breakdown of Ensembl annotated function for the 25 gene family IDs recovered as part of the Dumas comparison with the release 67 raw gene family data.

Ensembl Family ID	Ensembl Family Description
ENSFM00250000000002	TELOMERIC REPEAT BINDING FACTOR 1 TTAGGG REPEAT BINDING FACTOR 1
ENSFM00250000000099	NUCLEAR ENVELOPE PORE MEMBRANE POM 121 NUCLEOPORIN NUP121 PORE MEMBRANE OF 121 KDA
ENSFM00250000000393	PRAME FAMILY MEMBER
ENSFM00250000000661	ALPHA AMYLASE PRECURSOR EC_3.2.1.1 1 4 ALPHA D GLUCAN GLUCANOHYDROLASE
ENSFM00250000001212	N LYSINE METHYLTRANSFERASE SETD8 EC_2.1.1.- HISTONE LYSINE N METHYLTRANSFERASE SETD8 EC_2.1.1.43 SET DOMAIN CONTAINING 8
ENSFM00250000001425	FRG1
ENSFM00250000001738	LIM AND SENESCENT CELL ANTIGEN CONTAINING DOMAIN 2 PARTICULARLY INTERESTING NEW CYS HIS 2 PINCH 2
ENSFM00250000002195	RANBP2 AND GRIP DOMAIN CONTAINING RAN BINDING 2 RANBP2 RANB
ENSFM00250000002588	CUTANEOUS T CELL LYMPHOMA ASSOCIATED ANTIGEN CTAGE
ENSFM00250000002759	TRIPARTITE MOTIF CONTAINING 16
ENSFM00250000003039	GENERAL TRANSCRIPTION FACTOR IIH SUBUNIT 2 GENERAL TRANSCRIPTION FACTOR IIH POLYPEPTIDE 2
ENSFM00250000004017	KERATIN TYPE I CYTOSKELETAL CYTOKERATIN CK KERATIN
ENSFM00250000004074	RNA BINDING MOTIF PROTEIN X HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN G [CONTAINS RNA BINDING MOTIF PROTEIN X N TERMINALLY PROCESSED]
ENSFM00250000004772	CLASS I HISTOCOMPATIBILITY ANTIGEN ALPHA CHAIN PRECURSOR
ENSFM00400000131757	OLFACTORY RECEPTOR OLFACTORY RECEPTOR
ENSFM00500000269589	NEUROGENIC LOCUS NOTCH HOMOLOG PRECURSOR NOTCH [CONTAINS NOTCH EXTRACELLULAR TRUNCATION; NOTCH INTRACELLULAR
ENSFM00500000270385	FARNESYL PYROPHOSPHATE SYNTHASE FPP SYNTHASE FPS EC_2.5.1.10 2E 6E FARNESYL DIPHOSPHATE SYNTHASE DIMETHYLALLYLTRANSTRANSFERASE EC_2.5.1.- 1 FARNESYL DIPHOSPHATE SYNTHASE GERANYLTRANSTRANSFERASE
ENSFM00500000270422	DELETED IN AZOOSPERMIA DAZ
ENSFM00500000270455	GAMMA GLUTAMYLTRANSPEPTIDASE PRECURSOR GGT EC_2.3.2.2 GAMMA GLUTAMYLTRANSFERASE GLUTATHIONE HYDROLASE EC_3.4.19.- 13 [CONTAINS GAMMA GLUTAMYLTRANSPEPTIDASE HEAVY CHAIN; GAMMA GLUTAMYLTRANSPEPTIDASE LIGHT CHAIN]

Supplementary Table 4.3 continued

Ensembl Family ID	Ensembl Family Description
ENSFM00560000771007	PROLINE DEHYDROGENASE 1 MITOCHONDRIAL PRECURSOR EC_1.5.99.8 PROLINE OXIDASE
ENSFM00560000771165	ZINC FINGER
ENSFM00610000952844	TBC1 DOMAIN FAMILY MEMBER
ENSFM00650001140038	COBW DOMAIN CONTAINING COBALAMIN SYNTHASE W DOMAIN CONTAINING
ENSFM00660001157182	SERINE/THREONINE KINASE SMG1 SMG 1 EC_2.7.11.1
ENSFM00670001235658	OLFACTORY RECEPTOR OLFACTORY RECEPTOR

Appendix 4.7 – Table of noteworthy matching gene families recovered from the Dumas analyses

Supplementary Table 4.4 - Breakdown of Ensembl annotated function for the 2 gene family IDs recovered as part of the Dumas comparison with the release 67 fixed lambda CAFE data.

Ensembl Family ID	Ensembl Family Description
ENSFM00250000000661	GAMMA GLUTAMYLTRANSPEPTIDASE PRECURSOR GGT EC_2.3.2.2 GAMMA GLUTAMYLTRANSFERASE GLUTATHIONE HYDROLASE EC_3.4.19.- 13 [CONTAINS GAMMA GLUTAMYLTRANSPEPTIDASE HEAVY CHAIN; GAMMA GLUTAMYLTRANSPEPTIDASE LIGHT CHAIN]
ENSFM00250000000661	GAMMA GLUTAMYLTRANSPEPTIDASE PRECURSOR GGT EC_2.3.2.2 GAMMA GLUTAMYLTRANSFERASE GLUTATHIONE HYDROLASE EC_3.4.19.- 13 [CONTAINS GAMMA GLUTAMYLTRANSPEPTIDASE HEAVY CHAIN; GAMMA GLUTAMYLTRANSPEPTIDASE LIGHT CHAIN]

Appendix 4.8 – Divergence times in primates

Supplementary Table 4.5 - Median divergence times (Mya) for species in the primates dataset relative to *Homo sapiens* taken from TimeTree.org (Hedges *et al.*, 2006)

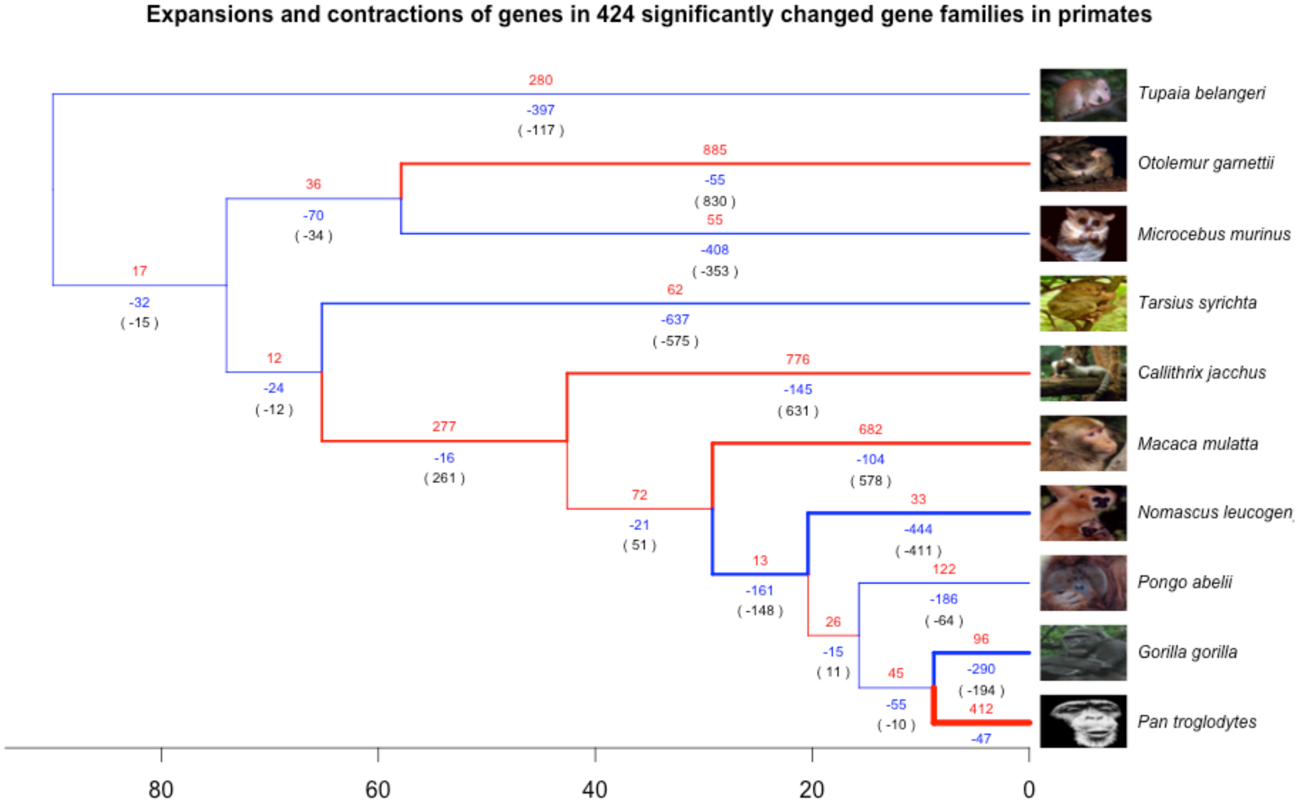
Primates	Divergence Time
Homo sapiens	-
Pan troglodytes	6.1
Gorilla gorilla gorilla	8.0
Pongo abelii	15.1
Nomascus leucogenys	19.4
Macaca mulatta	26.8
Callithrix jacchus	40.1
Tarsius syrichta	58.0
Microcebus murinus	77.5
Otolemur garnettii	77.5
Tupaia belangeri	89.0

Appendix 4.9 – Divergence times in rodents

Supplementary Table 4.6 - Median divergence times (Mya) for species in the rodents dataset relative to *Mus musculus* taken from TimeTree.org (Hedges *et al.*, 2006)

Rodents	Divergence Time
Mus musculus	-
Rattus norvegicus	22.0
Dipodomys ordii	73.0
Ictidomys tridecemlineatus	74.3
Cavia porcellus	77.0
Oryctolagus cuniculus	86.1

Appendix 4.10 – No human gene family expansions and contractions tree



Expansions and contractions of genes in 424 significantly changed gene families in primates

Supplementary Figure 4.1 - Expansions and contractions of genes along the branches of the primate phylogenetic tree with human data trimmed. Blue coloured branches depict overall contraction, while red coloured branches depict overall expansion. Black branches would represent equal or no change. Branch thickness represents the number of gene copy number changes weighted by the time to the ancestral node for each branch as a proportion of the time to the root node.

Appendix 5

Appendix 5.1 – Breakdown of intron count information for all 61 species available in release 70 of the EnsEMBL databases. Intron data were trimmed so all genes had at least 1 intron

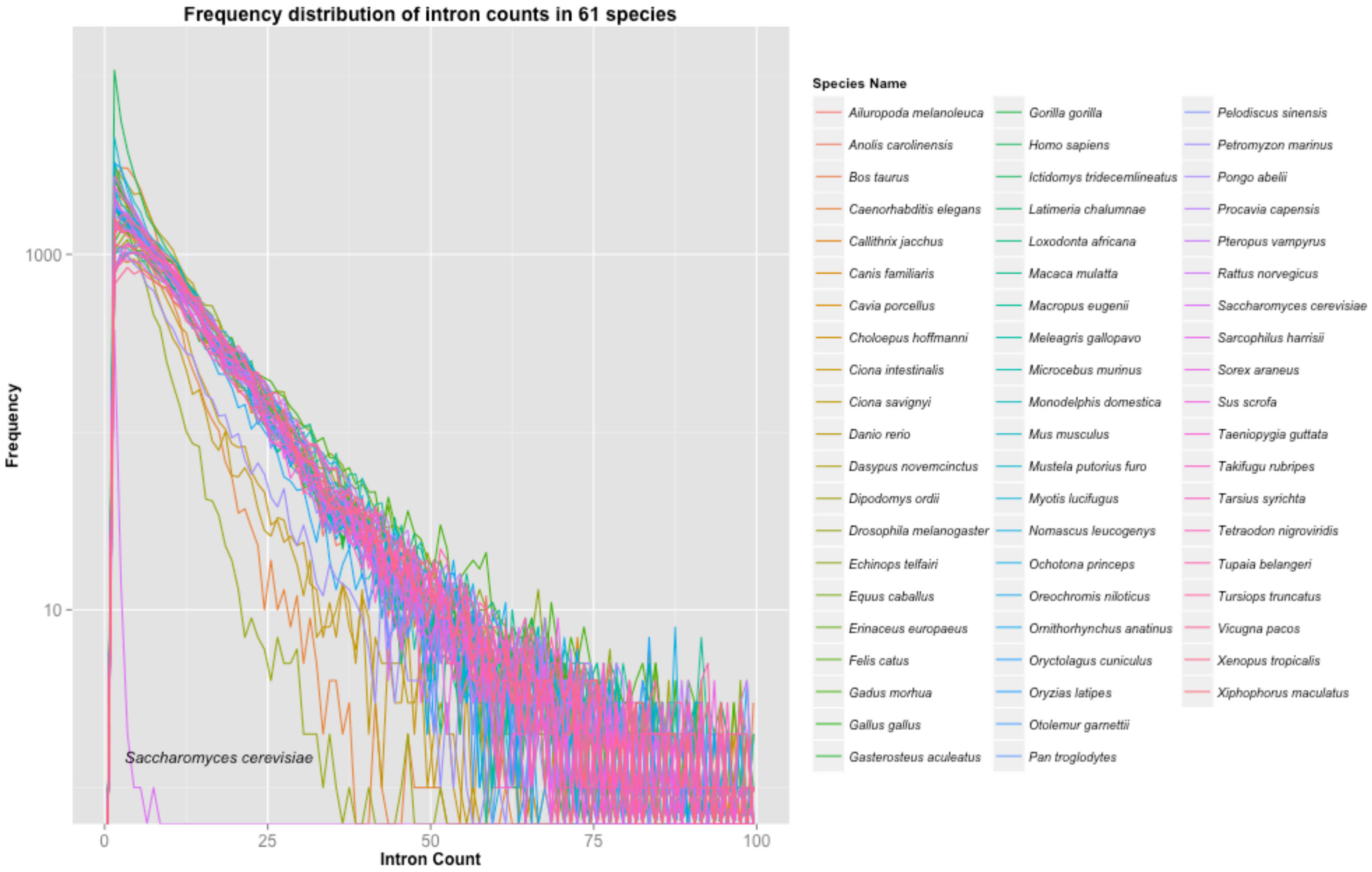
Supplementary Table 5.1 - Breakdown of intron count information for all 61 species available in release 70 of the EnsEMBL databases. Intron data were trimmed so all genes had at least 1 intron.

Species Name	Max	Mean	Median	Mode
ailuropoda_melanoleuca	314	9.388855	7	1
anolis_carolinensis	151	9.170396	6	1
bos_taurus	316	9.665572	7	1
caenorhabditis_elegans	65	5.238211	4	2
callithrix_jacchus	355	8.904024	6	1
canis_familiaris	344	9.544023	7	1
cavia_porcellus	147	9.89385	7	2
choloepus_hoffmanni	378	11.40097	8	4
ciona_intestinalis	131	6.119172	4	1
ciona_savignyi	126	6.674546	5	1
danio_rerio	229	8.516419	6	1
dasypus_novemcinctus	368	11.59302	8	3
dipodomys_ordii	331	12.0776	9	5
drosophila_melanogaster	81	3.864209	3	1
echinops_telfairi	278	11.85999	9	3
equus_caballus	357	8.859638	6	1
erinaceus_europaeus	261	12.28845	9	3
felis_catus	315	9.609164	7	1
gadus_morhua	254	12.04534	8	2
gallus_gallus	144	9.347318	6	1
gasterosteus_aculeatus	212	9.904965	7	1
gorilla_gorilla	346	8.903558	6	1
homo_sapiens	311	6.343056	3	1
ictidomys_tridecemlineatus	154	9.604691	7	1
latimeria_chalumnae	161	9.842294	7	1
loxodonta_africana	283	9.508163	7	1
macaca_mulatta	156	8.28536	5	1
macropus_eugenii	367	12.60975	9	4
meleagris_gallopavo	137	10.29303	7	2
microcebus_murinus	232	11.67309	9	3

Supplementary Table 5.1 continued

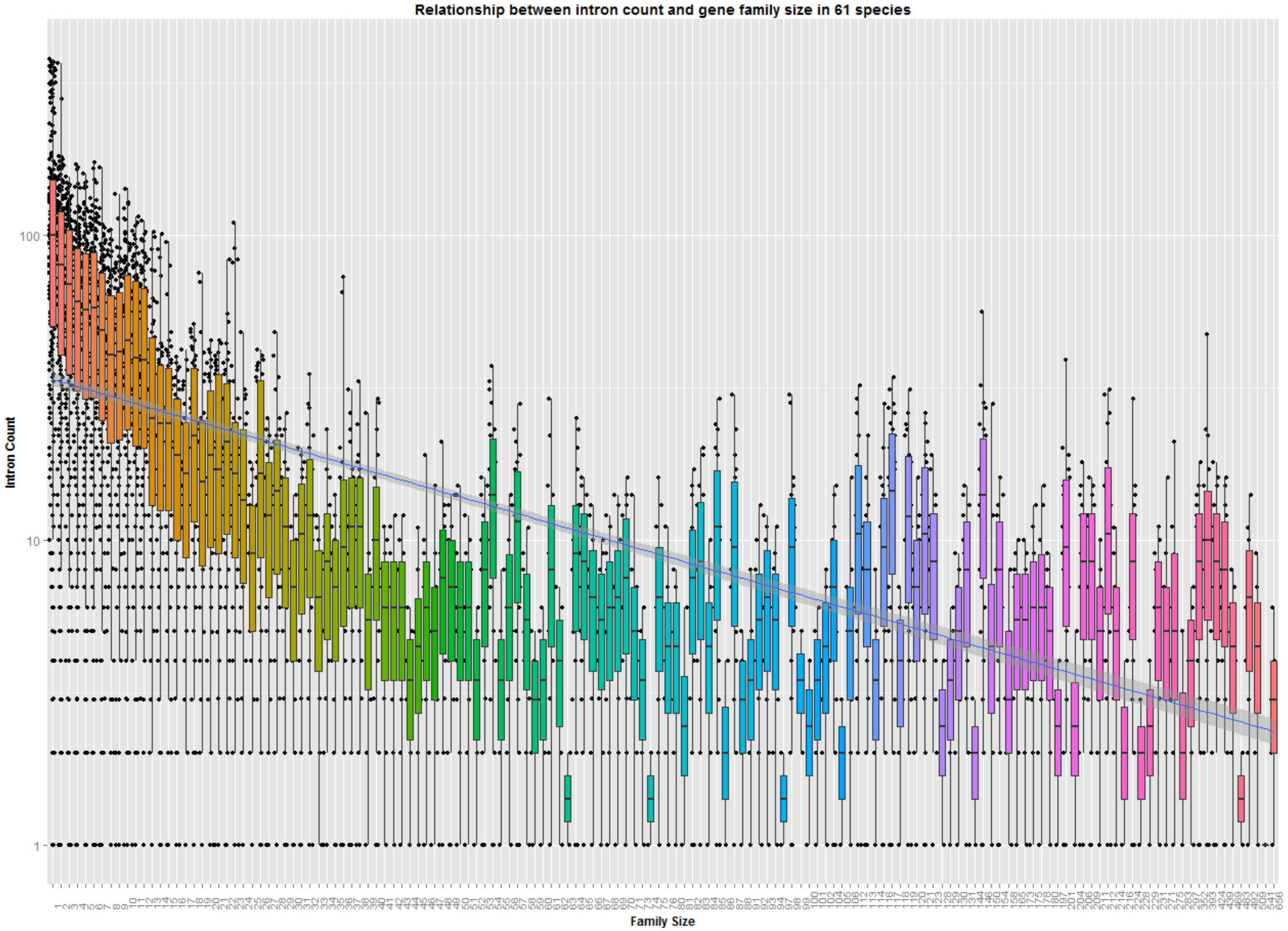
Species Name	Max	Mean	Median	Mode
monodelphis_domestica	152	9.227022	6	1
mus_musculus	311	7.818095	5	1
mustela_putorius_furo	316	9.553409	7	1
myotis_lucifugus	348	8.913681	6	1
nomascus_leucogenys	151	9.364024	7	1
ochotona_princeps	319	12.37557	9	5
oreochromis_niloticus	147	10.54295	8	3
ornithorhynchus_anatinus	152	7.131029	4	1
oryctolagus_cuniculus	281	9.382012	7	1
oryzias_latipes	218	9.803657	7	1
otolemur_garnettii	150	9.497065	7	1
pan_troglodytes	313	9.572439	7	1
pelodiscus_sinensis	357	9.390303	7	1
petromyzon_marinus	143	8.19893	6	1
pongo_abelii	291	9.567673	7	1
procavia_capensis	343	12.31972	9	3
pteropus_vampyrus	369	11.88527	9	5
rattus_norvegicus	117	8.527337	6	1
saccharomyces_cerevisiae	7	1.078283	1	1
sarcophilus_harrisii	299	9.705794	7	1
sorex_araneus	251	11.78891	9	3
sus_scrofa	117	8.414934	6	1
taeniopygia_guttata	148	8.679421	6	1
takifugu_rubripes	151	10.43109	8	1
tarsius_syrichta	338	11.37114	8	4
tetraodon_nigroviridis	159	9.89281	7	3
tupaia_belangeri	271	12.01076	9	1
tursiops_truncatus	364	11.84556	9	3
vicugna_pacos	352	11.59147	8	3
xenopus_tropicalis	294	10.26406	7	1
xiphophorus_maculatus	157	10.42504	8	3

Appendix 5.2 – Frequency distribution of intron counts in all 61 species.



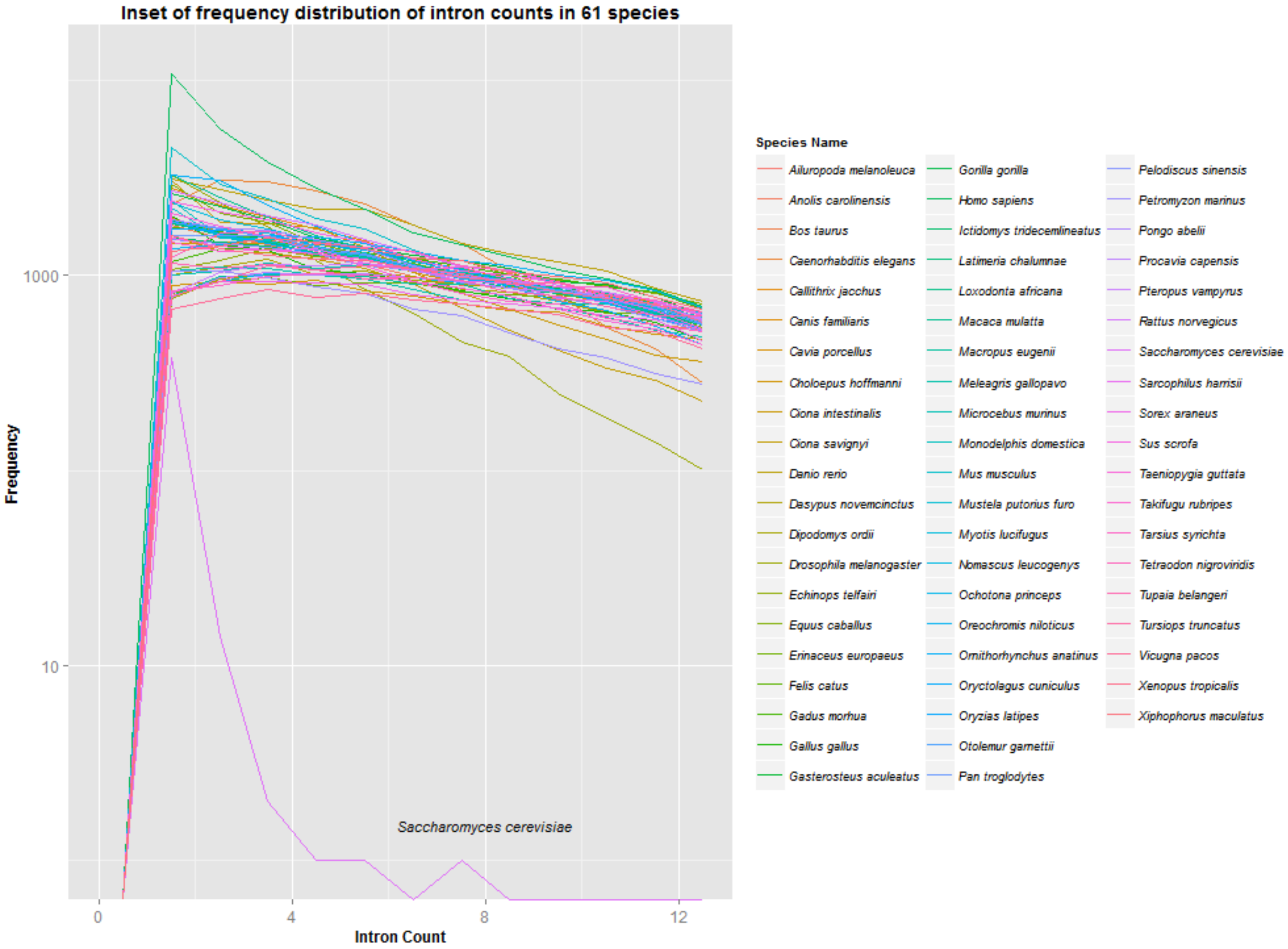
Supplementary Figure 5.1 - Frequency distribution of intron counts in all 61 species. Intron count is trimmed to 100, which represents the majority of the data. This figure represent the right skew in the data, with the mean, median and mode all being approximately <= 10. The maximum intron count in these species is 378. See Appendix 5.1 for inset of 0 to 12.5 intron count.

Appendix 5.3 – Boxplot displaying the relationship between gene family size and intron count.



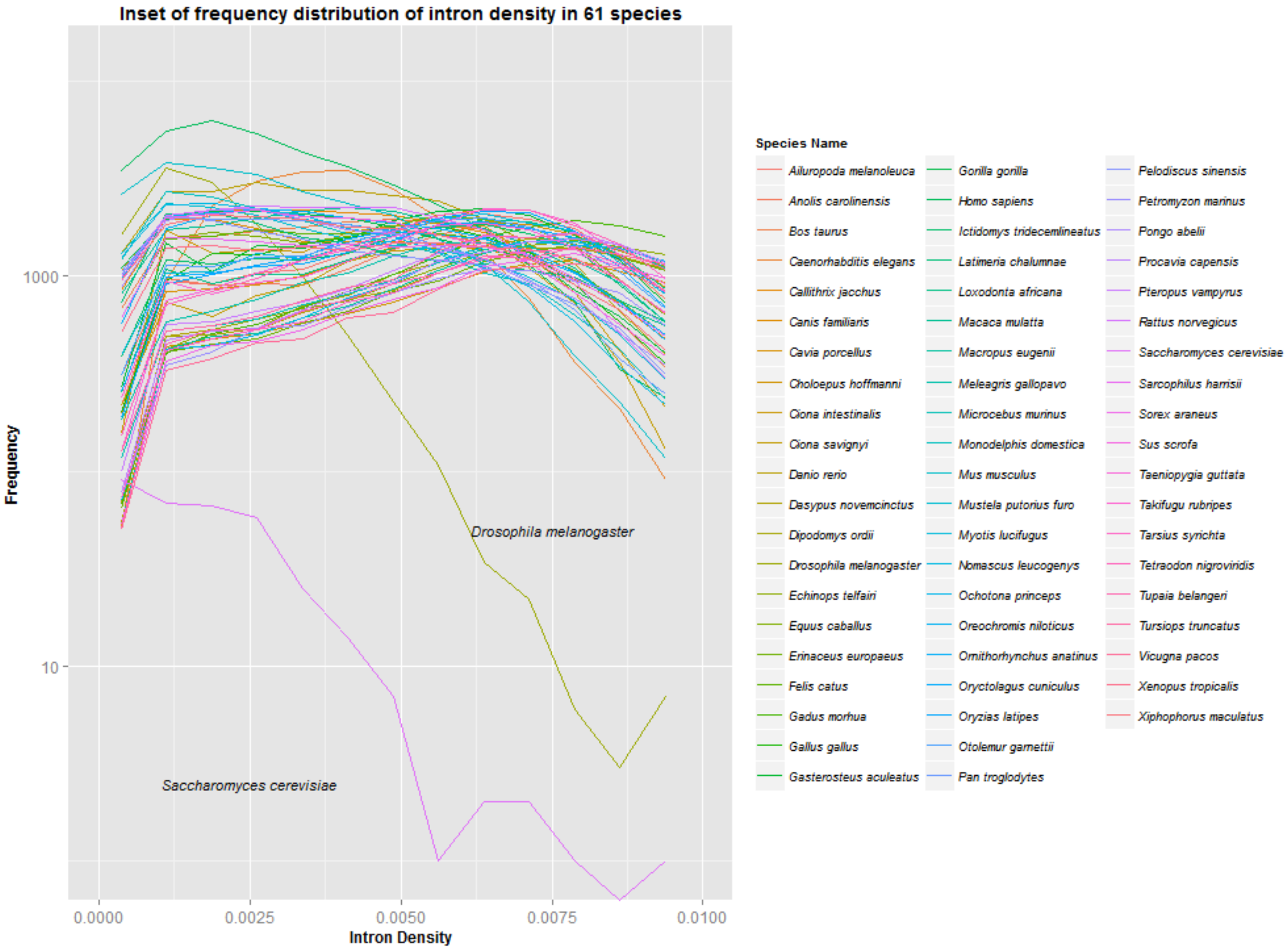
Supplementary Figure 5.2 - A boxplot displaying the relationship between gene family size and intron count for the pooled intron and gene family data of all 61 species used in this study.

Appendix 5.4 – Inset of supplementary figure 5.1 showing frequency distribution of intron count in 61 species



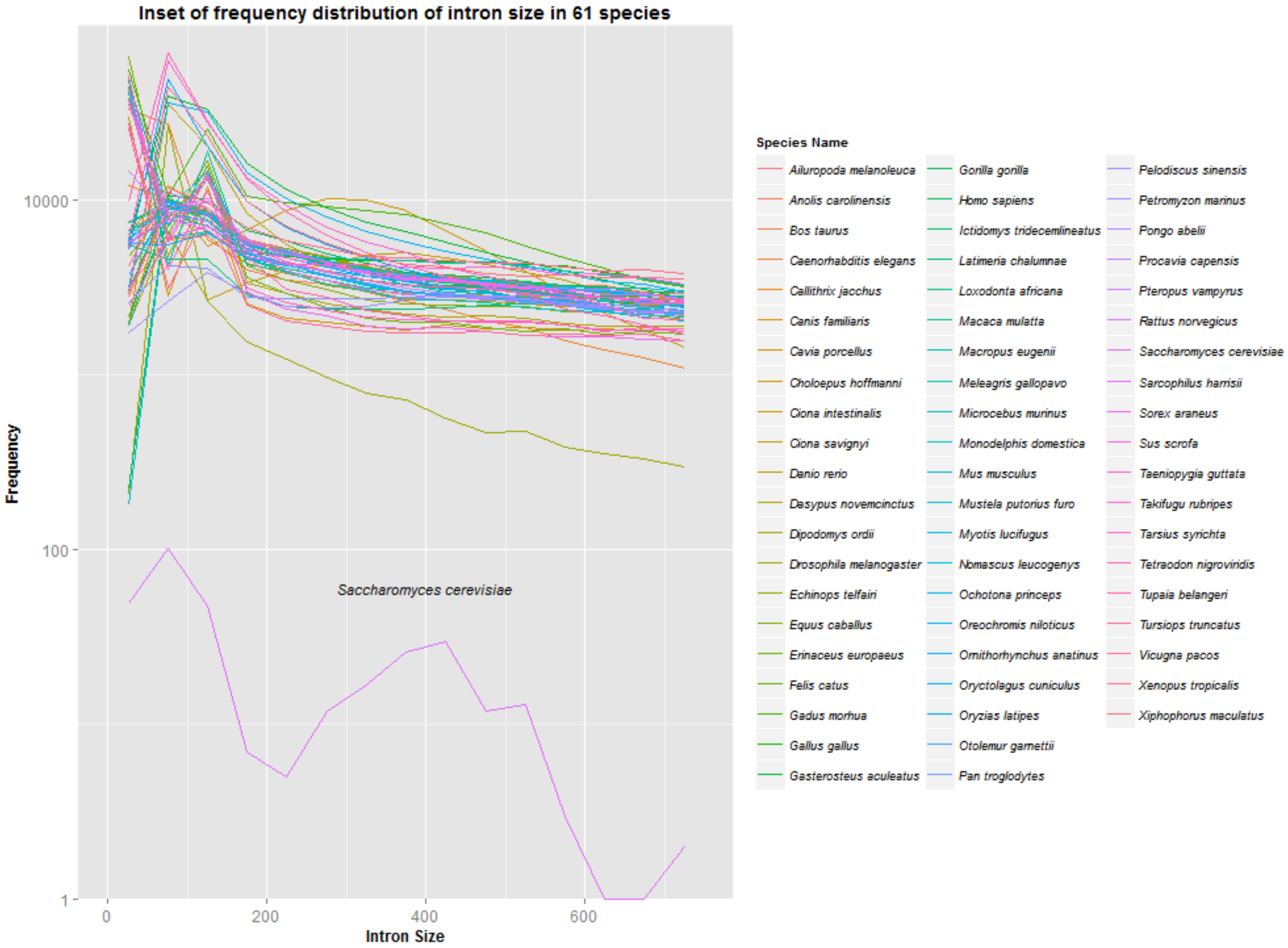
Supplementary Figure 5.3 – Inset of frequency distribution of intron count in all 61 species in Ensembl release 70. Cut-off at 12.5 to emphasize mode count in the distribution.

Appendix 5.5 – Inset of figure 5.3 showing frequency distribution of intron density in 61 species



Supplementary Figure 5.4 – Inset of frequency distribution of intron density in all 61 species in Ensembl release 70. Cut-off at 0.01 to emphasize mode density in the distribution.

Appendix 5.6 – Inset of figure 5.4 showing frequency distribution of intron size in 61 species



Supplementary Figure 5.5 – Inset of frequency distribution of intron size in all 61 species in Ensembl release 70. Cut-off at 750 bp to emphasize mode size in the distribution.