PHD

Genomic signatures of neurodegeneration and the evolution of mammalian brain.

Castillo Morales, Atahualpa

Award date:
2015

Awarding institution:
University of Bath

Genomic signatures of neurodegeneration and the evolution of mammalian brain.

Atahualpa Castillo Morales

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology

June 2015

Table of Contents

Acknowledgments VI

Contributions VIII

Abstract IX

List of Figures X

List of Tables XI

List of Supplementary Tables XII

Abbreviations XIII

1 Introduction 1

1.1 The genomic basis of complex phenotypes 1
1.2 Trade-offs in the evolution of large brains 3
1.4 Genetic signatures of brain evolution 8
1.5 Neocorticalization, differential scaling of large brains 12
1.6 Neurodegeneration and cellular longevity 14
1.7 Objectives & approach of this thesis 15

2. Increased brain size in mammals is associated with size variations in gene families with cell signalling, chemotaxis and immune-related functions 21

2.1 Abstract 22
2.2 Introduction 23
2.3 Methods 24
2.3.1 Gene family annotations 24
2.3.2 Encephalization index 25
2.3.3 Correlation coefficients of gene family size and encephalization index 25
2.3.4 Gene ontology terms analysis 25
2.3.5 Maximum lifespan and partial correlation coefficients 26
2.3.6 Phylogenetic relatedness test 26
2.3.7 Gene expression in human brain 27
2.3.8 Co-expression network analysis 27

2.4 Results 28

2.4.1 Gene family size variations in line with encephalization are over-represented in mammals 28

2.4.2 Association between gene family size and encephalization is not explained by lifespan variations 31

2.4.3 Phylogenetic relatedness does not explain the observed bias in the distribution of gene families associated with encephalization 33

2.4.4 Gene families with size increases in line with encephalization show expression signatures consistent with brain functions 33

2.5 Discussion 39

2.6 Conclusion 41

3. Neocortex expansion in mammalian lineages explains gene family size variations associated with larger brain size 44

3.1 Abstract 45

3.2 Introduction 46

3.3 Methods 49

3.3.1 Gene Family annotations 49

3.3.2 Encephalization index, Neocortex ratio and Maximum Lifespan 49

3.3.3 Correlation coefficients of gene family size and phenotypes 50

3.3.4 Gene Ontology terms enrichment 50

3.3.5 Phylogenetically controlled regression 51

3.3.6 Gene expression prior and after peak in neocortex thickness 51

3.4 Results 52

3.5 Discussion 58

4. Post-mitotic cell longevity-associated genes: a transcriptional signature of post-mitotic maintenance in neural and non-neural tissues 62

4.1 Abstract 63

4.2 Background 64

4.3 Results 66

4.4 Discussion 81

4.5 Conclusions 88

4.6 Materials and Methods 89

4.6.1 Cellular longevity estimates. 89

4.6.2 Human tissues gene expression data 89

4.6.3 Co-expression analyses.. 91
4.6.4 Enrichment of disease down regulated genes... 91
4.6.5 Functional enrichment analysis. 91
4.6.7 Statistical analysis... 92
5. General Discussion 96

5.1 Variations in gene family size and encephalization 97
5.2 Changes in gene family size and neocorticalization. 98
5.3. Molecular basis of neuronal post mitotic maintenance.................................... 100
6. References.. 104

7. Appendices... 129

Acknowledgments

Firstly, I would like to thank my supervisors Dr Araxi Urrutia and Dr Humberto Gutierrez for their support, guidance and patience. Being under their mentorship has allowed me to grow both as a scientist as well as an individual. Araxi really helped me to rediscover my lost love with genomics and computational biology. The endless discussions with them leading to this theses taught me not only how to look at problems from different perspectives, but also how to keep my temper when it comes to my scientific endeavour, without losing my passion for it. The fact that they pushed me to reach my limits helped me acquire experience in different research areas, as well as scientific management, which I'm sure have proven key to my career development and are the reason I managed to advance as much as I have in my career. I would also like to thank Jimena Monzon-Sandoval, my partner in crime both in the lab and outside of it. Without her contributions to this work, none of these projects would have been realized and without her patience and understanding I wouldn't have been able to endure a PhD in what at first was a strange and stressing foreign land.

I would also like to thank my assessors Prof Ed Feil and Dr Paula Kover for their comments and helpful discussions which have certainly helped in polishing my project and develop my career.

Next I would like to thank all past and present members of Araxi Urrutia's group, Dr Jaime Tovar, Dr Stephen Bush, Dr Lu Chen, Dr Wang Wei, Dr Nina Ockendon, Alin Acuña, Abduljalil Al-Zadjali, Kathryn Maher, Bing Xia for allowing me to take part on their research and/or helping me with mine either directly or with comments and discussions, and more importantly for their friendship and comradery, as well as for sharing these 4 years with me and making this group such an interesting place to work. In the same way I would like to thank all my collaborators without whose support and data many of my publications would not be possible.

I would also like to thank James Wallis, Saurav Sinha, Agnieszka Ziolek, Francesca Brown, Jessica Steven, Mayra Ruiz, Smruti Deoghare and all of the Masters and undergrad final year students who allowed me to dispose of their time in order to practice my teaching and supervision skills, and I hope I didn't traumatize you into leaving science.

I’m very grateful with Dr Avazeh Ghanbarian, Dr Jennifer McDowall, Dafina Angelova Dr Sergio Ancona, Cristina Carmona and all my friends who I met at the University. Without you pushing me to be happy, taking me out of isolation and inviting me have a merry pint, a soothing cup of coffee or a nice calm chat once in a while, I wouldn’t have enjoyed my stay in this University as much as I lately have.

From the bottom of my heart I also want to thank my mother and my sister, the main reason I strive to be a good scientist and a good human being, as well as my uncles, aunts, grandparents and all of my family who through their love and support have laid the foundation of everything that I am and everything I have accomplished so far.

Last but not least, none of this work would have been possible without the financial support from the Science and Technology National Council (CONACyT, Mexico).

Contributions

I confirm that the findings presented in this thesis are the result of my own work carried out with the advice and support of my supervisors Dr Araxi Urrutia and Dr Humberto Gutierrez, with the following exceptions

In chapter 2 and 3 gene expression analysis where performed by Jimena Monzon-Sandoval

In chapter 4 co-expression analysis where performed by Jimena Monzon-Sandoval

In chapter 3 data collection was carried out by Dr Alexandra de Souza

All the aforementioned also contributed valuable edits and comments both to the experimental design of this studies and to the wording of the manuscripts here presented.

Results presented in chapter 2 have been published in the journal Proceedings of the Royal Society B and may be cited as:

Castillo-Morales, A., et al. (2014), 'Increased brain size in mammals is associated with size variations in gene families with cell signalling, chemotaxis and immune-related functions', Proceedings of the Royal Society B-Biological Sciences, 281 (1775).

Results presented in chapters 3 and 4 are being prepared for submission.

I have made the following contributions to additional published papers presented in the Appendix section:

In Zhang et al. 2013, I analysed rates of gene expression across human tissues.

In Bush et al. 2014, I carried out functional characterization of genes displaying presence-absence variation

In Chen et al. 2014, I carried out phylogenetic regression analysis between organismal complexity and its predictors.

Abstract

Due to the complex adaptive costs and benefits of large brains and large neocortical volume, mammalian species exhibit huge variation in brain size. However, the precise nature of the genomic changes accounting for these variations remains poorly understood. Using genome-wide comparative analysis of gene family size of more than 39 fully sequenced mammalian species, I studied whether changes in the number of copies of genes involved in distinct cellular and developmental functions has contributed to shaping the morphological, physiological and metabolic machinery supporting brain evolution in mammalians. My results reveal an overrepresentation of gene families displaying a positive association between GFS and level of encephalization. This bias occurs most prominently in families associated with specific biological functions, such as cell-cell signalling, chemotaxis and immune system. Moreover, I find that most gene family size variations associated with increased brain size are mostly explained by the link between neocortex ratio and gene family size variations. The results in this study suggest that variations in gene family size underlie morphological adaptations during brain evolution in mammalian lineages.

Lastly, using comparative transcriptomics analysis across different human tissue types with cellular longevities ranging from 120 days to over 70 years, I set out to identify the molecular signature of long term post-mitotic cell maintenance. I found that genes down regulated in Alzheimer's and Parkinson's disease are significantly enriched in genes whose expression levels are associated with increased post-mitotic cellular longevity (PMCL). This holds true also for genes down regulated in Hutchinson-Gilford progeria-derived fibroblasts. The work here presented suggest that PMCL-associated genes are part of a generalized machinery of post-mitotic maintenance and functional stability in both neural and non-neural that becomes compromised in two specific neurodegenerative conditions and supports the notion of a common molecular repertoire for cell maintenance differentially engaged in different cell types with differing survival requirements.

List of Figures

Figure 2.1.Enrichment of gene family size variations in line with increased encephalization in mammals. 29

Figure 2.2. Relationship between Ei-associated GFS variations and patterns of gene expression in the human nervous system. 35

Figure 2.3. Temporal patterns of cortical expression of Ei-associated gene families. .. 38

Figure 3.1. Enrichment of gene family size variations in line with increased encephalization index and neocortex ratio in mammals. 53

Figure 3.2. Gene Ontology enrichment analysis of families with GFS variations in line with Encephalization index and Neocortex ratio. 55

Figure 3.3. Comparison between the cortical expression levels prior and after maximal neocortex thickness is reached of genes within the families more significantly associated with neocortex ratio after correcting for con founding variables. 58

Figure 4.1. Genes down regulated in brain tissue of Alzheimer’s and Parkinson’s disease patients normally display increased level of expression in longer living tissues. 68

Figure 4.2. Detection of a transcriptional signature of long term post-mitotic maintenance. 71

Figure 4.3. PMCL-associated genes are down regulated in Alzheimer’s and Parkinson’s disease. 75

Figure 4.4. PMCL-associated genes are down regulated in Hutchinson-Gilford Progeria syndrome. 77

Figure 4.5. PMCL-associated genes display a reduced level of co-expression in in Alzheimer’s and Parkinson’s disease 78

Figure 4.6. Down regulation of PMCL-associated genes in degenerative conditions is specifically linked to their association with cellular longevity. 80

List of Tables

Table 4.1. Sources of gene expression data. 69

Table 4.2. PMCL-associated genes 72

Table 4.3. GO slim terms and transcription factor target significantly enriched among PMCL-associated genes (Adjusted p-value < 0.05) 95

List of Supplementary Tables

Supplementary Table S1.1. Genes with evolutionary patterns associated with the nervous system evolution 16

Supplementary Table S2.1 42

Supplementary Table S3.1 61

Supplementary Table S4.1 94

Abbreviations

Ei	Encephalization Index
Nr	Neocortex Ratio
BMR	Basal Metabolic Rate
GFS	Gene Family Size
MLSP	Maximum Lifespan
PMCL	Post-mitotic Cell Longevity
GO	Gene Ontology
GWAS	Genome-Wide Association Study
SNP	Single Nucleotide Polymorphism
MYA	Million Years Ago
^{14}C	carbon-14
CR	caloric restriction
AD	Alzheimer’s disease
PD	Parkinson’s disease
HGPS	Hutchinson-Gilford syndrome progeria
EC	entorhinal cortex
PC	posterior Cingulate cortex
MTG	medial temporal gyrus
HIP	hippocampus
SFG	superior frontal gyrus
VCX	visual cortex

SEM	standard error of the mean
GEO	Gene Expression Omnibus
TF	transcription factors
UPR	unfolded protein response
UPS	ubiquitin/proteasome system
ER	endoplasmic reticulum.

1 Introduction

1.1 The genomic basis of complex phenotypes

What at the genomic level underlies the evolution of complex phenotypes and the changes that lead to disease states are key questions in genomics science. Early studies in genotype-phenotype led to the discovery of genes with large, dramatic effects on traits, which led to "a gene for" and “genetic blueprint” paradigms that are still pervasive across certain circles in biology, and limited the success of this endeavour to relevant but simple phenotypic traits, such as regulatory switches. It is now recognised however that complex phenotypes are driven and modulated by tens and sometimes hundreds of genes acting in concert.

Phenotype evolution can result from various types of mutations including single nucleotide substitutions, insertions and deletions which can affect coding regions and non-coding regulatory elements. Classically, quantitative genetics methods have been used for the study of the genetic basis of phenotypes and their evolution (Falconer and Mackay 1996; Hill 2010; Lynch and Walsh 1998). The development of genomics techniques to associate regions of the genome with the variation of traits has had a huge impact on this approaches by increasing the resolution and sensitivity of analysis; permitting to integrate expression profiling, marker-based fingerprinting, chromatin, and methyl-DNA immunoprecipitation among other sources of high-throughput data; and exploiting and developing many statistical tools used in the analysis of quantitative trait loci.(Jansen and Nap 2001; Perez-Enciso et al. 2007; Prins et al. 2012). One of the most widespread applications of this methods are genome-Wide Association Studies (GWAS), which, since the first GWAS in 2002 (Ozaki et al. 2002) followed by its popularization with a 2005 study on age-related macular degeneration (Klein et al. 2005), have identified thousands of genes and genetic variants (mainly SNPs) that contribute to phenotypic variations in many traits in different systems and species and have broadened our understanding of many diseases and phenotypes.

Nevertheless, association-based approaches have several important limitations. To begin, only a small fraction of SNPs/genes and their functional mechanisms have been functionally characterized. In most cases, loci found by GWAS have a very weak, additive predictive power which only explains a small fraction of the phenotypic variance (Kraft and Hunter 2009; Marjoram et al. 2014; Visscher et al. 2012; Ward and Kellis 2012), and inferring causal polymorphisms that account for such a tiny fraction of the variance requires a huge sample size (Long and Langley 1999). The low variance explain by predicted loci resulting from GWAS suggests that rare genetic variants or structural variants poorly captured by current technology account for most of the heritability of traits; or that these kind of analysis suffers from low power to detect epistatic interaction (Manolio et al. 2009), crucial in the case of complex traits where SNPs are unlikely to act alone. Moreover, many of the variants fall in noncoding regions of the genome, and due to linkage disequilibrium, can encompass many variants; and as such their functional effect is not immediately discernible (Ward and Kellis 2012). Furthermore, GWAS are marred with polymorphisms falsely identified as associated to a trait, in many cases due to confounding variables, such as environmental factors (e.g., geographic origin in a structured population) (Platt et al. 2010).

With the sequencing and transcriptomic profiling of an ever growing list of non-model species, comparative genomics approaches have increasingly proven invaluable in the efforts to characterize the conservation and variation in genomic features such as genomic sequence, genes, gene order, amino acid usage, protein rates of evolution, gene family size, regulatory motifs, in order to trace their origin and changes across evolution, the mechanisms and evolutionary forces shaping them, and their relationships to functional phenotypes. For example, the comparison of the human genome with that of our closest extant relative, the chimp, revealed how two non-synonymous changes in a single gene, the Fork head box protein P2 (FOXP2), are partially responsible for the control of orofacial movement that allowed humans to develop a spoken language (Enard et al. 2002). Studies using comparative genomics have even allowed us to understand more of the function of the non-coding regions of the genome; many cis and trans acting regulatory elements have been identified and characterized based on the evolutionary conservation of their sequence or transcription level across species (Boyle et al. 2014; Consortium et al. 2007; Gerstein et al. 2014; Ho et al. 2014; Marques and Ponting 2014; Prabhakar et al. 2008; Visel et al. 2007).

Comparative genomic studies on longer evolutionary ranges have permitted us to detect patterns of changes in the genome, such as chromosomal rearrangements, and granted us a view into how these rearrangements characterize differentiation and speciation, likely through their effect on recombination and gene flow (Faria and Navarro 2010; Feder et al. 2014; Kirkpatrick and Barton 2006; Yeaman 2013). Similarly, scans for regions on the genome that display signatures of positive selection can highlight the genes underlying phenotypic adaptations (Bustamante et al. 2005; Capra et al. 2013; Gaya-Vidal and Alba 2014; Hubisz and Pollard 2014; Kelley and Swanson 2008; Nielsen et al. 2007; Pritchard et al. 2010).

By extending our understanding of the genomic basis of phenotypes and their evolution over time, comparative studies have also advanced our understanding of disease states. A good example of the power of comparative genomics approaches can be found on a study comparing the human genome to those of model plant *Arabidopsis*, and the ciliated single cell organism *Chlamydomonas* (Li et al. 2004). This study allowed researchers to identify key genes involved in the formation of cilia in several cell types. Further characterisation of these genes led to the discovery of a cilia disease causing gene in humans.

In this thesis I use comparative approaches and transcriptome analyses to shine a light on the genomic basis of brain evolution and the mechanisms enabling the long term post-mitotic maintenance of neurons and other long lived cell types in the human body. Together my findings revealed that genes involved in a distinct set of cellular processes are likely to underlie the adaptations of the human brain which enabled our species to evolve a high cognitive ability.

1.2 Trade-offs in the evolution of large brains

One of the most distinctive features of hominid species is the increase in brain size with modern humans possessing one of the largest brains compared to body size. When compared to other vertebrates, mammalian species in general tend to have larger brain to body size ratios and this relationship is particularly pronounced in some primate and cetacean species (Roth and Dicke 2005). Larger brains have long been associated with higher cognitive capabilities, higher social interaction complexity and better ability to cope with unstable environmental conditions (Dunbar 1992; Jerison 1973, 1985; R. D. Martin 1983).

The organisms who possess large brains, undoubtedly enjoy great cognitive advantages which should translate in a reduction in extrinsic mortality and favour a longer reproductive life, thereby compensating, at least partially, for the underlying fitness costs (Allman et al. 1993; Gonzalez-Lagos et al. 2010; Isler and van Schaik 2009; Sol 2009b). Whether it is by an expansion of the neocortex, relative brain size, encephalization, number and size of neurons, number of synaptic connections, or any of the other brain related traits the scale with an increase in brain size, a larger brain comes with increased cognitive abilities, which can confer a higher behavioural plasticity to a species and facilitate the construction of novel or altered behavioural patterns, which in turn could provide a buffer against socioecological challenges (Allman et al. 1993; Barrickman et al. 2008; Deaner et al. 2003; Ricklefs 2004; Sol 2009b). This cognitive buffer would enable an animal to do such things as track variations in the availability of resources, use tools in order to include hard-to-eat foods in their diets, colonize new ecological niches, deal with environmental complexity, avoid unfamiliar predators and collect information from conspecifics (Dukas 2004; Sol 2009a). Evidence on this regard comes from the observed association between innovation rate and relatively larger brains in both birds (Louis Lefebvre et al. 1997; L. Lefebvre et al. 2004) and primates (L. Lefebvre et al. 2004; S. M. Reader and Laland 2002), together with the fact that species with larger brains show lower adult mortality rates in the wild when compared with the species with smaller brains (Sol et al. 2007) and, importantly, that large-brained species are more successful than small-brained species when introduced by humans to novel environments (Sol et al. 2002; Sol et al. 2005; Sol et al. 2008). Furthermore, successful invaders display a high behavioural innovation rate in their native ranges and are less likely to experience population decline due to alterations of their habitat (Shultz et al. 2005) and are more tolerant to climatic variability (Schuck-Paim et al. 2008; van Woerden et al. 2012). These observations have led to the suggestion that the relatively large brain of the *Neornithes* compared to other archaic birds and pterosaurs might be partially responsible for their survival during the mass extinction event at the Cretaceous–Tertiary boundary, when all other flying *Ornithodira* groups became extinct (Milner and Walsh 2009).

Large brains are also associated with qualitative differences in mating system across birds and mammals, with species that live in pair-bonded social systems having the largest brains (Shultz and Dunbar 2007). In anthropoid primates the relationship between social system and brain size, particularly neocortex size, become quantitative with species with larger social groups having larger brains (Dunbar 1992; Dunbar and Shultz 2007a; Dunbar 2009). This has led to the hypothesis that bigger groups exert a selective pressure towards large brains and larger neocortices, through the demands of behavioural coordination, in order to manage complex social relationships (Byrne and Whiten 1988; Dunbar 1992; Dunbar and Shultz 2007a; Dunbar 2009; Shultz and Dunbar 2007). This social complexity might increase fitness by allowing groups to better defend against predation (Dunbar 1992; Dunbar and Shultz 2007a; Dunbar 2009; Shultz et al. 2004) and by permitting the social transmission of survival, reproduction and foraging skills (S. M. Reader and Laland 2002; van Schaik and Burkart 2011). In species with large group sizes, cooperative breeding might also alleviate part of the high maternal investment cost associated with a larger brain in the offspring (Burkart et al. 2009; Isler and van Schaik 2009; Isler 2011; Navarrete et al. 2011).

As previously mentioned, all these cognitive advantages would ultimately have an adaptive impact by incrementing survival rate and allowing selection to favour individuals with a longer lifespan (Allman et al. 1993; Gonzalez-Lagos et al. 2010), but the selective forces acting on a greater longevity can also, in turn, exert pressure towards a larger brain. For instance, long lived species are more likely to be exposed to environmental changes throughout their life, and as such would be benefited more from information acquisition and flexible behavioural flexibility than short lived species (Deaner et al. 2003; Sol 2009a, 2009b). Moreover, a longer life may favour a delay in reproductive cycles, allowing progenitors to invest more resources in their offspring (Covas and Griesser 2007). If individuals living in stable social groups face higher cognitive demands than that individuals living alone, this might lead to an increase in brain size (Charvet and Finlay 2012; Connor 2007; Dunbar and Shultz 2007a; Gonzalez-Lagos et al. 2010; Shultz and Dunbar 2007). Large brains, however, represent an evolutionarily costly adaptation, which require organisms to undergo various trade-offs to maintain. To begin, the brain is a metabolically expensive tissue. In vertebrates, between 2% and 8% of the basal metabolic rate (BMR) is used by the central nervous system. In primates, this load can rise to over 10% of the BMR and, in the case of human, 20%, escalating to 80% during development

(Mink et al. 1981). Certainly, increments in brain size result in a higher metabolic load on the organisms, which poses an important constraint on brain expansion (R. D. Martin 1981). Nevertheless, variations in BMR in mammals can only explain a small proportion of the variance in the relative brain size in these species (between 13.3% in mammals and 20% in non-human primates (Isler and van Schaik 2006a)).

A reduction in the relative size of one or more metabolically active organs, such as the liver, kidney, testes, heart, or intestine could partially alleviate the metabolic cost imposed by increments in the relative brain size (Aiello and Wheeler 1995; Barrickman and Lin 2010; Fish and Lockwood 2003; Kozlovsky et al. 2014; Tsuboi et al. 2015). In this respect, it has been shown that there is a negative correlation between the brain size and the gut size in anthropoid primate species (Aiello and Wheeler 1995; W. R. Leonard et al. 2003). Similar patterns have been observed in fish (Kaufman et al. 2003; Tsuboi et al. 2015), but not in birds (Isler and van Schaik 2006b) or bats (K. E. Jones and MacLarnon 2004). Regardless, in this last clade, there is an inverse relationship between brain size and testis size (Lemaitre et al. 2009; Pitnick et al. 2006). Additionally, humans, the mammalian species with the largest relative brain size, have substantially lower levels of muscle mass when compared with other primates (W. R. Leonard et al. 2003; W. R. Leonard et al. 2007). Furthermore, the whole of the *Primata* order, which generally has undergone several events of brain expansion along its evolutionary history, has relatively less muscle mass when compared to other mammals (W. R. Leonard et al. 2007; Snodgrass et al. 2009). When taken together, these findings suggest the possibility that different trade-offs occur in different groups (Barton 2006).

Energetic trade-offs between the brain and other abundant, yet less expensive tissues may also account for part of the variations in brain size. For instance, Dror and Hopp propose a metabolic trade-off between hair and brain in human evolution, based on the fact that these are two of the three organs with the greatest essential amino acid requirement from a whole body perspective, and hair contains large amounts of several amino acids that are essential for brain development and function (methionine, cysteine, tyrosine, phenyl alanine and arginine), but are in limited quantity in food (Dror and Hopp 2014). Another example of a possible energetic trade-off between the brain and other profuse tissue is posed by adipose depots, which account for a considerable proportion of the body mass of some mammalian species (Pond 1998), and imposes an energetic and adaptive cost due to the fact that it needs to be carried around and may increase extrinsic mortality by predation (Navarrete et al. 2011). Indeed, Navarrete et al. showed that there is a negative correlation between relative brain size and the size of adipose depots in a sample of 100 mammals, although this relationship is not significant in primates (Navarrete et al. 2011). Conversely, fat depots could act as a buffer against starvation and enable to stabilize the energy supply available for a larger brain (Kuzawa 1998; W. R. Leonard et al. 2003; Wells 2006), which would also face important seasonality-induced energetic constraints (Jiang et al. 2015; van Woerden et al. 2010). Furthermore, adiposity might alleviate more than only the energetic requirements of the brain. For instance, cognitive ability of a child has been shown to be associated with a lower waist to hip ratio in the mother, suggesting that gluteo-femoral adiposity might provide neurodevelopmental resources in the shape of fatty acids needed for foetal brain development (Lassek and Gaulin 2008).

Another non-excluding strategy which can lead to an increase in the net energy available for a larger brain consists on the adoption of a higher quality diet, energy-dense and high in structural carbohydrates and proteins. (Aiello and Wheeler 1995; Fish and Lockwood 2003; W. R. Leonard et al. 2003; W. R. Leonard et al. 2007; Verginelli et al. 2009). A shift in diet towards greater meat consumption, would also provide for increased levels of fatty acids necessary for the evolution of a large brain (Cordain et al. 2001; Crawford et al. 1999; W. Leonard et al. 2011). In particular during the evolution of the *Homo* genus, tool-assisted processing of food, followed by development of cooking also helped to increase diet quality and promote brain evolution (W. Leonard et al. 2011; Plummer 2004; Wrangham 2009).

Other trade-offs are present in the shape of an impact of the brain size on life history traits. Due to the costly, complicated and long developmental processes of large brains, both neonatal and adult brain size has been found to be associated with gestation length in mammals, even when controlling for the overall size of the body (Finarelli 2010; Isler and van Schaik 2006b; Isler and van Schaik 2009; R. Martin 1996; Sacher and Staffeldt 1974). Furthermore, Barton and Capellini found that evolutionary changes in pre- and postnatal brain growth correlate with duration of both gestation and lactation phases in placental mammals, strongly suggesting the large maternal investment required for a larger brain size, which inevitably leads to a reduction on the annual fertility rate of species with large brains due to longer inter-birth periods (Barton and Capellini 2011). This long developmental periods result in increased offspring mortality risk (Barrickman et al. 2008;

Deaner et al. 2003; Sacher and Staffeldt 1974) and delayed age of first reproduction (Barrickman et al. 2008; Deaner et al. 2003). The large maternal investment requirements imposed by large brains can also be offset by a reduction in litter size (Finarelli 2010; Isler and van Schaik 2009) or by an increase in the BMR of the mother during pregnancy (R. Martin 1996). All of these trade-offs impose a fitness cost which should be compensated for by an increased reproductive lifespan if a species is to maintain demographic viability (Isler and van Schaik 2009).

In any case, the plethora of cost and adaptive impacts and trade-offs of larger brains has resulted in a high variation in brain size even across closely related species (e.g. (Barrickman et al. ; Harvey et al. 1980; Huber et al. 1997; Kotrschal et al. 1998; Sol and Price)), yet the exact nature of the molecular changes accounting for variations in encephalization across mammalian species is at present poorly understood.

1.4 Genetic signatures of brain evolution

One of the goals of evolutionary neurobiology is to underpin the molecular changes accounting for the extraordinary expansion in brain size observed across mammalian evolution. Changes in brain size can be associated with changes at any molecular level, from a single nucleotide in a particular gene or regulatory elements, to evolutionary patterns on protein domains, whole proteins, gene families or pathways.

One approach used to discover these changes consist on measuring the strength of selective pressure acting on a gene or set of genes, in taxa that display the phenotype of interest and compare it against those who do not. Most studies in this regard take a candidate based approach, with a focus on genes involved in neural proliferation, cell death, energy metabolism or any other process closely associated with the development and function of the brain, as well as genes involved in neurological disorders. One of the most prominent examples of this comes from studies on the rates of evolution of genes involved in primary microcephaly, a developmental disease which results in a small but, otherwise generally normal brain (Woods et al. 2005). There are twelve known genes that cause this neurodevelopmental disorder (reviewed in (Faheem et al. 2015)), and five of these (ASPM, CDK5RAP2, CENPJ, SHH and MCPH1) have been found to display signatures of positive selection in the primates relative to other mammals, and this acceleration seems to be particularly prominent in the primate lineage leading to humans (reviewed in (S. L. Gilbert et al. 2005)). Furthermore, there is a significant correlation between the rates of protein evolution and neonatal brain size for ASPM and CDK5RAP2 (Montgomery et al. 2011; Montgomery and Mundy 2012b, 2014). Protein rates of evolution of the gene NIN, a centrosome maturation factor integral for neurogenic division of radial glial cells (X. Wang et al. 2009), also display an association with brain mass across anthropoids (Montgomery and Mundy 2012a), suggesting that all of these genes evolved adaptively during anthropoid evolution and may have a role in the evolution of brain size in this taxa.

Other genes that show accelerated rates of evolution that may underlie the evolution of brain size, through their involvement in neurogenesis include ADCYAP1, which encodes a secreted protein that regulates transcriptional activation of key mediators of neuroendocrine stress responses and cortical neurogenesis and underwent an accelerated evolution in the human lineage since the divergence from chimpanzee (Y. Q. Wang et al. 2005). The glutamate receptors GRIN3A and GRIN3B, which regulate excitatory synaptic transmission in the brain; and the Fork head box protein P2 (FOXP2), a transcription factor with mutations associated to severe speech and language disorder (Lai et al. 2001) also displays signatures of positive selection in the primate lineage leading to human (Enard et al. 2002; Goto et al. 2009), an in which two amino acid substitution that occurred after the split between human and chimp fixated, conferring this gene the ability to regulate new targets involved in guiding neuronal morphology, dendritic length and plasticity in cortico-basal ganglia, suggesting that this mutations contributed to increased fine-tuning of motor control and vocal learning during human evolution (Enard et al. 2009; Enard 2011; Konopka et al. 2009). AHI1 (Abelson helper integration site 1), one of the genes associated with a rare brain malformation called Joubert syndrome and involved in directing axons from the brain to the spinal cord, also shows an accelerated rate of evolution along the human lineage since its split from the chimp (Ferland et al. 2004).

A complementary approach utilized to understand the genetic components of brain evolution consists in identifying non-coding regions conserved across primates or mammals with an accelerated substitution rate in a particular lineage. (Bush and Lahn 2005;

Pollard 2006) One of the most prominent loci derived from this kind of analysis is the human accelerated region 1 (HAR1), the conserved non-coding region with the higher evolutionary rate along the human lineage (Pollard 2006). HAR1, is part of an RNA gene (HAR1F) that is expressed highly and specifically in Cajal–Retzius neurons in the human neocortex early in human embryonic development, particularly during stages characterized by cortical neuron specification and migration (Pollard 2006). HAR1F is co-expressed with reelin, a product of Cajal–Retzius neurons that is of fundamental importance in specifying the six-layer structure of the human cortex. Furthermore, HAR1F has been shown to be downregulated in neurodegeneration in Huntington's disease (Johnson et al. 2010). Changes in non-coding regions by mutations in regulatory elements could also translate in variations in gene expression and these have been suggested as an important contributor to the evolution of uniquely human biological traits, such as our oversized brain (Rakic 2009; Sholtis and Noonan 2010).

Changes in gene expression resulting in phenotypic changes can also be derived from posttranslational histone modifications (Allfrey et al. 1964). Genome-wide profiling of these modifications has been used to compare regulatory element activities across species to identify promoters and enhancers that have gained activity in humans. Using this approach, Reilly et al. found that such gains are significantly enriched in modules of coexpressed genes in the cortex that function in neuronal proliferation, migration, and cortical-map organization, suggesting that these regulatory changes might be an important driving force of human cortical evolution (Reilly et al. 2015).

Among the mechanisms by which our DNA may alter the evolution of the brain, one of the most relevant is through duplication of protein coding genes. These may arose from segmental duplications and potentially leaded the origin of new functions. Even if the number of gene duplications that are retained during evolution is small given the large frequency with which these events occurs through evolutionary time, the strong selection acting on those that remain may have a considerable effect in driving the differences we observe between species (Gokcumen et al. 2011; Lynch and Conery 2000), and of particular interest those reflecting brain size (Supplementary Table 1.1). ARHGAP11B encodes a Rho GTPase and is a clear example of human specific partial duplication that has been recently described to promote the expansion of the neocortex through increasing basal progenitor cell population (Florio et al. 2015). ARHGAP11B role in brain evolution was identified by its particular expression pattern at basal progenitor cells and the lack of orthologous counterparts in mice and chimp (Florio et al. 2015). Another human specific lineage duplication possibly associated with the evolution of the brain occurred at the SLC6A13 gene, a transporter of the main neurotransmitter, GABA (Fortna et al. 2004). Changes in the number of copies of this gene have been linked to higher cognitive functions and its loss has been associated to anxiety disorders (Saus et al. 2010). Among the most prominent variations in number that has occurred after the chimp split is the expansion of DUF1220 gene family, also known as the neuroblastoma breakpoint family (NBPF). The ancestral DUF1220 domain can be found in the centrosomal protein myomegalin, and both the number of copies of the NBPF family members and DUF1220 domains on them are highly correlated to brain size between primate species and within human populations (Dumas 2012; Keeney et al. 2014a; Keeney et al. 2014b). It is worth mentioning that, as the copy number increases in autistic persons, the individuals show a more extreme severity of impaired social reciprocity, communicative ability and increased repetitive behaviours.

While total gene number has remained relatively constant throughout the past 800 million years of metazoan evolution (Ponting 2008), there exist large variations among organisms in the number of copies of genes involved in a variety of biological functions, and changes in this genotypic trait have shown to occur frequently (Demuth et al. 2006; Fortna et al. 2004; Hahn et al. 2007; Hughes and Friedman 2004; Rubin et al. 2000). As mentioned above, changes in gene family size are of summary importance as a driver of phenotypic evolution, since gene duplications offer material for the origin of new functions and expression patterns, whilst gene loss acts as response to selective pressures (Krylov et al. 2003; Lynch and Conery 2000). In spite of this, whether changes in the size of gene families involved in distinct processes has contributed to shaping the machinery driving brain evolution in mammalian lineages remains an interrogate. While candidate gene studies have undeniably contributed to our understanding of the molecular mechanisms underlying the complex process that is brain evolution, generally these studies have been marred by an anthropocentric view of this process, and as such focus only in a few primate species at most. In Chapter 2, I examined the association between changes in gene family size (GFS) and degree of encephalization in 39 fully sequenced mammalian species using a genome-wide comparative approach with aims to further our understanding of the genomic correlates of encephalization across the whole Mammalia class.

1.5 Neocorticalization, differential scaling of large brains

While encephalization is a significantly variable trait across mammalian species (Barrickman et al. 2008; Harvey et al. 1980; Kaas 2006), not all brain structures have scaled up proportionally along the evolutionary history of this taxa. In this taxa, most variations in relative brain size can be explained by changes in the size of the neocortex (Jerison 1973, 1990; Kaas 2006).

The neocortex is a structure uniquely present in mammals that surrounds the cerebral hemispheres. This structure is the newest part of the mammalian cerebral cortex, and has an origin approximately 220 million years ago with the origin of this lineage (Kaas 2011; Meredith et al. 2011; Northcutt and Kaas 1995; M. A. O'Leary et al. 2013). The neocortex is likely derived from the thin dorsal ridge of reptiles (Kaas 2011; Northcutt and Kaas 1995; M. A. O'Leary et al. 2013), and has been greatly modified into six layers of cells, which differ in density and size of neural cell bodies and axons, and contains around 24 billion neurons and 33 billion glial cells on average in humans (Pelvig et al. 2008). While evidence points towards little change in the size of the neocortex of mammals until around 60 million years ago, the major radiations of marsupials and placental mammals after the Cretaceous-Paleogene (K-Pg) mass extinction brought with many independent neocorticalization events across different mammalian groups (Northcutt and Kaas 1995).

Traditionally, the neocortex has been regarded as the seat for the neurobiological mechanisms of higher cognitive abilities, such as self-awareness, consciousness, abstract reasoning and planning, in mammals (Crick and Koch 1990; Eccles 1994; P. Gilbert et al. 1995; Grober et al. 1992; Steven M. Platek et al. 2004; S. M. Platek et al. 2008; Sugiura et al. 2005). Specific areas in the human neocortex have been found to drive the understanding and production of language (Aiello and Dunbar 1993; Letinic et al. 2002). Other highly specialised areas, such as those for recognizing faces (Allison et al. 1994; Nestor et al. 2011), or places (V. M. Miller and Best 1980; Poucet et al. 2003), have also been identified in the neocortex. Neocortex to brain size ratio is correlated with social group size (R. I. M. Dunbar 1992), and it has been theorized that the quantity of neocortical neurons is a constraining factor in determining the number of social relationships which an animal can manage (R. I. M. Dunbar 1992), and in the hominid lineage, the expansion of the neocortex is thought to have played a key role for the evolution of modern humans (DeFelipe 2011). Furthermore, mirror-neurons which have been implicated in social learning (they respond both to doing an action or seeing it be done by other individuals, and seem to play a part in action understanding, speech perception, emotion recognition and imitation) (Adolphs et al. 1994; Enticott et al. 2008; Iacoboni 2005; Schulte-Ruther et al. 2007; Spaulding 2013; van der Gaag et al. 2007; Wicker et al. 2003), have been identified in the neocortex (Gallese et al. 2002; Molenberghs et al. 2009; Rizzolatti et al. 1996).

In spite of the highly significant part that the evolution of the neocortex has played in mammalian evolution, the genomic features underlying their evolution remain poorly understood (de Sousa and Proulx 2014; Hawrylycz et al. 2012). Thus far, there have been few studies attempting to identify signatures of the impacts of brain evolution on the genome. While a previous effort to detect a genomic signature of the bran evolution reported that genes involved in different aspects of nervous system biology displayed accelerated sequence evolution on the lineage leading from ancestral primates to humans (Dorus et al. 2004), this claim was heavily contested soon after (Kosiol et al. 2008; Shi et al. 2006). A recent study showed that the degree of encephalization is significantly associated with overall protein amino acid composition., perhaps mirroring the selective demands imposed by a larger brain, by conducting a genome-wide analysis of amino acid composition across 37 fully sequenced mammalian genomes (Gutierrez et al. 2011).

Changes in gene family size are one of the main forces driving many evolutionary changes. Duplication events provide source of material for the origin of novel gene functions and expression patterns, whereas gene loss is suggested to act as response to selective pressures (Krylov et al. 2003; Lynch and Conery 2000). Marked differences in gene family size have been identified in drosophila and vertebrates with families involved in particular functions being enriched in those experiencing the largest changes, suggesting that lineage-specific changes in gene family size play a large role in adaptation (Demuth et al. 2006; Demuth and Hahn 2009; Hahn et al. 2005; Hahn et al. 2007; Han et al. 2009). A recent study found that encephalization in mammalian lineages is associated with significant variations in gene family size (see chapter 2 and (Castillo-Morales et al. 2014)) with the most positively associated gene families significantly enriched in several functional categories including immune system response, chemotaxis and cell-cell signalling, however part of these observations could be a by-product of the high degree of association between encephalization and neocorticalization.

In chapter 3 we investigate if the variations in neocortex to brain ratio in mammalian lineages are associated with changes in gene family size. By tracing back events of gene gain and loss per gene family we are able to distinguish gene family expansion and gene family contraction in association with neocorticalization. We further explore the extent to which any changes in gene family size associated with neocortex explain previously reported variations in gene family size and encephalization.

1.6 Neurodegeneration and cellular longevity

Intelligence has evolved independently many times among vertebrates (Emery and Clayton 2004; Simon M. Reader et al. 2011; Roth and Dicke 2005), yet excluding humans, few other vertebrates suffers from an age-related neurodegenerative syndromes, such as Alzheimer’s and Parkinson’s disease (Heuer et al. 2012; Vite and Head 2014). This suggest that susceptibility to Alzheimer’s evolved recently in human evolution, likely coinciding with the rapid expansion of brain and neocortex size occurred in this lineage, and possibly as a by-product or trade-off of this evolutionary process. A larger and more complex brain, with its attached high metabolic costs, could become less efficient with age as a result of changes in gene expression that affect normal neural functions such as synaptic transmission, axonal integrity and myelination (Bishop et al. 2010; Loerch et al. 2008). In this respect, a direct comparison between the aging brain transcriptomes of mouse, macaque and human revealed a major evolutionary divergence in dysregulation of many neural related genes with aging, with genes involved in functions such as regulation of axonogenesis, neurogenesis and GABA signalling showing a marked downregulation unique to old humans (Loerch et al. 2008). These changes could potentially reveal the genetic basis of vulnerability to neurodegeneration.

Although the need for long term survival is common to many cell types, nowhere is cell maintenance more critical than in neurons as mature post-mitotic neurons need to survive and preserve their functional complexity during the entire lifetime of an individual

(Magrassi et al. 2013), with failure at any level in the underlying supporting mechanisms resulting in a wide range of neurodegenerative conditions (Drachman 1997; Fishel et al. 2007; Mattson and Magnus 2006).

Post-mitotic neuron maintenance, as well as that of other cell types, must be the result of an interplay of a wide network of interacting molecular mechanisms that act at several levels of the cell's physiology to ensure its functional and structural stability (Lanni et al. 2010; Mattson and Magnus 2006). Identifying these networks should allow us to understand both cell survival as well as degeneration.

Until recently, our knowledge regarding post-mitotic cell longevity in human tissues has been limited due to the lack of means to accurately measure cell turnover in human subjects. However, recent attempts at estimating cell turnover rate based on 14C-based retrospective birth dating have proven more successful (Bhardwaj et al. 2006; K. L. Spalding et al. 2005). Using a comparative transcriptomic approach we attempt to identify the molecular signature of long term post-mitotic maintenance in 7 tissues with measured post-mitotic cellular longevity (PMCL) ranging from 120 day to over 70 years, by measuring the degree of association of expression patterns with this measurement of PMCL. In chapter 4, we identify a set of PMCL associated genes whose expression levels consistently mirror the differences in cell longevity across 7 different tissues. Furthermore, we show that these genes display concerted expression patterns in nerve cells and other long living tissues suggesting a functional association between these genes. We also found that PMCL-associated genes are down regulated in the cerebral cortex and substantia nigra of Alzheimer's and Parkinson's disease patients respectively, as well as Hutchinson-Gilford progeria-derived fibroblasts, further cementing their possible involvement on regulating cell survival. Finally, we found that sexual dimorphism in the patterns of gene expression of PMCL-associated genes in the brain reflects known differences between sexes in lifespan of humans and macaques.

1.7 Objectives & approach of this thesis

The overall objective of this thesis is to attempt to find the molecular basis underlying two complex phenotypes; the evolution of the mammalian brain, and the differences in long term post-mitotic cell maintenance across different cell types, using a comparative genomics and comparative transcriptomics approach.

Supplementary Table S1.1.

Genes with evolutionary patterns associated with the evolution of the central nervous system.

Gene/Element Name	**Mechanism of Change**	**Proposed Phenotype**	**Possible Gene-Associated Disease**	**Refs.**
Fork head box P2 (FOXP2)	Amino acid change/Positive Selection	Language/speech development and increased length of dendrite spines	Speech-language disorder-1	(Enard 2002)
Glutamate receptor, ionotropic, N-methyl-D-aspartate 3A (GRIN3A)	Amino acid change/Positive Selection	Learning and memory	Unknown	(Goto et al. 2009)
Glutamate receptor, ionotropic, N-methyl-D-aspartate 3B (GRIN3B)	Amino acid change/Positive Selection	Higher brain function	Unknown	(Goto et al. 2009)
Cholinergic receptor, nicotinic alpha 7 and FAM7A fusion (CHRFAM7A)	Copy number increase	Higher brain function	P50 sensory gating deficit	(Fortna et al. 2004)
Dopamine receptor D5 (DRD5)	Copy number increase	Regulation of mood, memory, learning, attention, movement	DRD5 deficiency, ADHD, primary cervical dystonia	(Fortna et al. 2004)
p21 protein (Cdc42/Rac)-activated kinase 2 (PAK2)	Copy number increase	Neuronal differentiation	3q29 microdeletion syndrome	(Fortna et al. 2004)
Peripheral myelin protein 2 (PMP2)	Copy number increase	Myelin stabilization/Protection from demyelination	Charcot-Marie-Tooth peroneal muscular atrophy	(Fortna et al. 2004)
Phosphodiesterase 4D interacting protein (PDE4DIP)	Copy number increase	Higher brain function	Myeloproliferative disorder associated with eosinophilia	(Fortna et al. 2004)

Solute carrier family 6 (facilitated glucose transporter) member 13 (SLC6A13)	Copy number increase	Higher brain function	Schizophrenia	(Fortna et al. 2004)
SLIT-ROBO Rho GTPase activating protein 2 (SRGAP2)	Copy number increase	Increased neuronal branching	Early infantile epileptic encephalopathy	(Fortna et al. 2004)
Rho GTPase Activating Protein 11B (ARHGAP11B)	Copy number increase	Neocortical expansion/ promotes basal progenitor generation	Unknown	(Florio et al. 2015)
Glutamate dehydrogenase 2 (GLUD2)	Copy number increase/ positive selection	Metabolic changes in brain	Parkinson's disease	(Burki and Kaessmann 2004)
MAS-related gene (MRG) family	Copy number increase/ positive selection	sensitivity and/or selectivity of nociceptive neurons to aversive stimuli	Unknown	(Choi and Lahn 2003)
Protocadherin 11 X Y linked (PCDH11XY)	Copy number increase/Expression change	Cerebral asymmetry/Language development, neuroendocrine transdifferentiation	Klinefelter's syndrome, Alzheimer's disease, prostate cancer	(Fortna et al. 2004)
Growth arrest and DNA-damage-inducible, gamma (GADD45G)	Deletion of regulatory DNA/Expression change	Expansion of human forebrain	Thyroid carcinoma	(McLean 2011)
Transforming Growth Factor, Beta Receptor III (TGFβR3)	Epigenetic gains	Neocortical expansion	Familial cerebral saccular aneurysm	(Reilly et al. 2015)
Collagen, type XIII, alpha-1 (COL13A1)	Epigenetic gains	Neocortical expansion	Embryo lethal mutant	(Reilly et al. 2015)
Ephrin receptor EphA2 (EPHA2)	Epigenetic gains	Neocortical expansion	Cataract	(Reilly et al. 2015)

LIM Homeobox transcription factor 1, beta (LMX1B)	Epigenetic gains	Neocortical expansion	Nail-patella syndrome, genitopatellar syndrome (microcephaly)	(Reilly et al. 2015)
Creatine kinase brain (CKB)	Expression change	Metabolic changes in brain	Multiple sclerosis	(Pfefferle 2011)
Solute carrier family 2 (facilitated glucose transporter) member 1 (SLC2A1)	Expression change	Metabolic changes in brain and skeletal muscle/Brain size	GLUT1 deficiency syndrome 1 and 2, susceptibility to HTLV infection	(Fedrigo et al. 2011)
Solute carrier family 2 (facilitated glucose transporter) member 4 (SLC2A4)	Expression change	Metabolic changes in brain and skeletal muscle/Brain size	Noninsulin-dependent diabetes mellitus	(Fedrigo et al. 2011)
Thrombospondin 4 (THBS4)	Expression change	Synaptic organization and plasticity	Familial premature coronary heart disease	(Caceres et al. 2007)
Prodynorphin (PDYN)	Expression change	Metabolic changes in brain	Spinocerebellar ataxia 23, dissociative amnesia	(Rockman et al. 2005)
Sialic acid-binding Ig superfamily lectin 11 (SIGLEC11)	Gene conversion/Expression change	Alleviation of neurotoxicity from activated microglia. Potential neurotrophic effects.	Unknown	(X. Wang 2011)
Fork head box D4 (FOXD4)	Novel Gene Variant	Nervous system development	Dilated cardiomyopathy, suicidality, OCD	(Cooper and Kehrer-Sawatzki 2011)
Survival of motor neurone2, centromeric (SMN2)	Novel Gene Variant	Motor neuron maintenance, neuronal growth	Spinal muscular atrophy severity	(Fortna et al. 2004)

Asp (abnormal spindle) homolog, microcephaly associated (ASPM)	Positive selection	Brain expansion	Microcephaly	(P. D. Evans 2004)
CDK5 regulatory subunit associated protein 2 (CDK5RAP2)	Positive selection	Brain expansion	Microcephaly	(Bond et al. 2005)
Human accelerated region 1 forward (HAR1f)	Positive Selection	Development of neocortex	Huntington's disease	(Pollard 2006)
Microcephalin 1 (MCPH1)	Positive Selection	Brain expansion	Microcephaly	(Rimol 2010)
Abelson helper integration site 1 (AHI1)	Positive selection	Higher motor function	Joubert syndrome	(Ferland et al. 2004)
Centromere protein J (CENP-J)	Positive selection	Brain expansion	Microcephaly	(Bond et al. 2005)
Sonic hedgehog (SHH)	Positive selection	Brain expansion	Microcephaly. Holoprosencephaly, other developmental disorders	(Dorus et al. 2006)
Ninein (NIN)	Positive selection	Increase in neuron number	Microcephaly	(Montgomery and Mundy 2012a)
Adenylate Cyclase Activating Polypeptide 1 (ADCYAP1)	Positive selection	Brain expansion	Post-traumatic stress disorder	(X. Wang 2011)
Cernunnos-XLF	Positive selection	Brain expansion	Microcephaly	(Pavlicek and Jurka 2006)

DUF1220/Neuroblastoma breakpoint factor (NBPF) family	Protein domain copy number increase (hyperamplification)	Brain expansion	Microcephaly, macrocephaly	(Fortna et al. 2004)
Myosin, heavy chain 16 (MYH16)	Pseudogeneization	Under-developed masticatory system releasing cranium from geometric constraint	Unknown	(Stedman et al. 2004)

2. Increased brain size in mammals is associated with size variations in gene families with cell signalling, chemotaxis and immune-related functions

Atahualpa Castillo-Morales[1], Jimena Monzón-Sandoval[1], Araxi O. Urrutia[1*], Humberto Gutiérrez[2*]

1 Department of Biology and Biochemistry, University of Bath, Bath, BA2 7AY, UK

2 School of Life Sciences, University of Lincoln LN6 7TS, UK

ACM: Acm39@bath.ac.uk

JMS: Jms52@bath.ac.uk

AOU: a.urrutia@bath.ac.uk

HG: hgutierrez@lincoln.ac.uk

* To whom correspondence should be addressed.

Citation: Castillo-Morales, A., et al. (2014), 'Increased brain size in mammals is associated with size variations in gene families with cell signalling, chemotaxis and immune-related functions', *Proceedings of the Royal Society B-Biological Sciences,* 281 (1775).

2.1 Abstract

Genomic determinants underlying increased encephalization across mammalian lineages are unknown. Whole genome comparisons have revealed large and frequent changes in the size of gene families, and it has been proposed that these variations could play a major role in shaping morphological and physiological differences among species. Using a genome-wide comparative approach, we examined changes in gene family size (GFS) and degree of encephalization in 39 fully sequenced mammalian species and found a significant over-representation of GFS variations in line with increased encephalization in mammals. We found that this relationship is not accounted for by known correlates of brain size such as maximum lifespan or body size and is not explained by phylogenetic relatedness. Genes involved in chemotaxis, immune regulation and cell signalling-related functions are significantly over-represented among those gene families most highly correlated with encephalization. Genes within these families are prominently expressed in the human brain, particularly the cortex, and organized in co-expression modules that display distinct temporal patterns of expression in the developing cortex. Our results suggest that changes in GFS associated with encephalization represent an evolutionary response to the specific functional requirements underlying increased brain size in mammals.

2.2 Introduction

Mammalian species in general tend to have larger brain to body size ratios compared with other vertebrates and in some primate and cetacean species this relationship is particularly pronounced (Roth and Dicke 2005). Large brains represent an evolutionarily costly adaptation as they are metabolically expensive, demand higher parental investment than in species with smaller brains and impose a substantial delay in reproductive age (Gonzalez-Lagos et al. 2010; Isler and van Schaik 2006b; W. R. Leonard et al. 2003; Roth and Dicke 2005; Weisbecker and Goswami 2010). In spite of the cost and adaptive impact of larger brains, the precise nature of genomic changes accounting for variations in encephalization across mammalian species is at present poorly understood (Dorus et al. 2004; Shi et al. 2006).

Whole-genome sequencing efforts have made it possible to study not just individual variations in specific sequences, but also large-scale differences in gene complements between species. Although overall gene number has changed little over the past 800 million years of metazoan evolution, comparative genomic studies have found large disparities among organisms in the number of copies of genes involved in a variety of cellular and developmental processes, and analyses of gene family evolution have shown that instances of gene family expansion and contraction are frequent (Demuth et al. 2006; Fortna et al. 2004; Hahn et al. 2007; Hughes and Friedman 2004; Rubin et al. 2000). In a recent analysis of Drosophila species, for instance, large numbers of gains and losses have been described, with over 40% of all gene families differing in size among the analysed species. Importantly, the fact that, in these species, rapid gene family size (GFS) evolution is accentuated in some functional categories strongly suggests that changes in gene number within gene families may reflect evolutionary responses to specific adaptive demands (Hahn et al. 2007). In this regard, gene duplication events specifically linked to distinct aspects of vertebrate evolution have been described. Examples include the expansion, during early evolution of the vertebrate lineage, of HOX and PAX gene families which are widely believed to have played a key part in the evolution of many known vertebrate innovations (Holland and Short 2008; Soshnikova et al. 2013).

A major goal in evolutionary neurobiology is to understand the molecular changes underlying the extraordinary expansion in brain size observed in mammalian evolution. Whether changes in the number of copies of genes involved in distinct cellular and developmental functions has contributed to shaping the morphological, physiological and metabolic machinery supporting brain evolution in mammalian lineages is not known.

By conducting a genome-wide analysis of 39 fully sequenced mammalian species, we set out to establish whether changes in GFS can be linked to increased encephalization. Our results reveal a proportion of gene families displaying a positive association between GFS and level of encephalization significantly larger than expected by chance. This bias occurs most prominently in families associated with specific biological functions. By examining expression data in human tissues, we further found that gene families displaying the highest association between encephalization and GFS are also statistically enriched in genes that are prominently expressed in the brain, with maximal expression in the cortex and displaying an expression signature distinctly associated with cortical development.

2.3 Methods

2.3.1 Gene family annotations

Annotated gene families encompassing 39 fully sequenced mammalian genomes were obtained from ENSEMBL (Flicek et al. 2012). In the context of this annotation, a given gene family constitutes a group of related genes that include both paralogues within the same species and orthologues and paralogues from other species. Any given gene can only be assigned to a single gene family. GFS represents the total number of genes per gene family. In order to maximize the number of families covered in this study (more than 10 000), we included all gene families with members present in no less than six of the 39 mammalian species.

2.3.2 Encephalization index

Because larger species have larger brains, it is necessary to estimate brain mass controlling for the allometric effect of body size. We therefore adopted residuals of a log–log least-squares linear regression of brain mass against body mass as this is the most widely accepted index of encephalization (Ei; supplementary material, table S2.1) (Herculano-Houzel et al. 2007; Herculano-Houzel 2011). While direct estimates of the ratio of brain mass to body mass have also been used as an alternative encephalization index (Deaner et al. 2000; Gonzalez-Lagos et al. 2010), this measure is known to be poorly related to brain complexity across taxa (Herculano-Houzel et al. 2007; Herculano-Houzel 2011). Accurate estimates of brain residuals based on a sample of 493 mammalian species were kindly provided by Gonzalez-Lagos et al. (Gonzalez-Lagos et al. 2010).

2.3.3 Correlation coefficients of gene family size and encephalization index

Simple Pearson correlations between Ei and GFS as well as multiple regressions (where maximum lifespan (MLSP) was included as covariate, see below) were carried out using R-based statistical functions. Numerical randomizations to determine statistical significance were conducted using specially written R-based scripts.

2.3.4 Gene ontology terms analysis

Gene ontology (GO) annotations were obtained from the Gene Ontology database (www.geneontology.org). In this study, a particular GO term was associated with a family whenever that term was linked to any of its members in any species. Only terms found to be linked with more than 50 families were examined.

For each GO category, the average Pearson correlation coefficient was calculated. Statistical significance and expected average Pearson correlation per GO was measured using at least 10 000 equally sized random samples taken from the whole gene family population to directly determine the corresponding p-values. Bonferroni correction was used in all analyses to correct for multiple tests.

Enrichment analysis of GO categories was carried out by counting the number of families assigned to each GO term within the analysed set of gene families. However, any bias in family counts per GO within a set of families could be owing to a bias in the overall density of GO annotation events within that sample. In order to adjust for differences in the density of GO annotations between the test and background samples, we divided the family counts per GO from each sample, by the samples' average number of GO annotations per family. Statistical significance was numerically assessed by obtaining the expected (adjusted) number of families per GO in 10 000 equally sized random samples derived from the overall population of gene families.

2.3.5 Maximum lifespan and partial correlation coefficients

MLSP recorded for each species was obtained from the animal ageing and longevity database (AnAge) (Tacutu et al. 2013). To correct for the potential contribution of MLSP to the association between GFS and Ei, partial correlation coefficients were calculated for each gene family, including MLSP as covariate. The resulting partial coefficient represents the contribution of Ei to the variance in GFS which is not explained by variations in MLSP. Only those gene families displaying a significant partial correlation coefficient ($p < 0.01$) between GFS and Ei were considered further.

2.3.6 Phylogenetic relatedness test

Phylogenetic generalized least-square approach (PGLS) and maximum-likelihood estimation of λ-values were carried out using the CAPER module in R. Because the parameter λ measures the degree to which the phylogeny predicts the pattern of covariance of a given trait across species (where λ-values close to 0 represent no phylogenetic autocorrelation while values close to 1 represent full phylogenetic autocorrelation) (Freckleton et al. 2002; Garland et al. 2005; Pagel 1999), this approach allows us to obtain a single accurate measure of phylogenetic autocorrelation for each individual gene family. In order to remove the effect of phylogenetic relationships from our analysis, we determined the parameter λ for each of the 713 gene families with significant partial correlation coefficients for Ei and GFS (correcting for MLSP) and eliminated all gene families with a significant phylogenetic interdependence ($p < 0.05$ of $\lambda = 0$, and $p > 0.05$ of $\lambda = 1$). This filtering resulted in 501 gene families on which GO enrichment analyses were subsequently carried out as described above.

2.3.7 Gene expression in human brain

RNA-seq data were obtained for 18 052 genes in a total of 16 human tissues, including brain, derived from the Illumina human body map dataset (ENSEMBL v. 62) (Flicek et al. 2012). Individual genes were categorized as prominently expressed in the brain if their expression level in this tissue was the highest or second highest among all 16 tissues included (top 12.5th percentile). Over-representation was assessed by counting the number of these genes within a given sample. Statistical significance was assessed by comparing this count with those observed in 10 000 equally sized random samples drawn from the wider pool of gene families.

2.3.8 Co-expression network analysis

Weighted gene co-expression network analysis was carried out based on pairwise Pearson correlations between the expression profiles obtained from the BrainSpan database (http://www.brainspan.org) for over 21 000 genes. Unsupervised hierarchical clustering was used to detect groups, or modules, of highly co-expressed genes following the method described by Zhang & Horvath (B. Zhang and Horvath 2005).

2.4 Results

2.4.1 Gene family size variations in line with encephalization are over-represented in mammals

In order to assess the relationship between encephalization and GFS variations in mammalian taxa, gene family annotations for 39 fully sequenced mammalian genomes were obtained from ENSEMBL (Flicek et al. 2012). We included in this study all families with members present in no less than six of the 39 mammalian species (see Methods). This resulted in a total of 12 373 non-overlapping gene families encompassing 595 535 genes, with a mean number of 48.13, and a number of copies per gene family per species ranging from 0 to 448.

Ei for each species was defined as the residual of a log–log least-squares linear regression of brain mass against body mass (see Methods). We obtained correlation coefficients for GFS and Ei for each gene family and the resulting distribution of correlation coefficients showed a distinct shift towards positive values (figure 2.1a). A Monte Carlo simulation of the expected distribution based on random permutations of GFS values across species revealed that the observed bias is highly significant (Embedded Image). In total, we found 8789 families with $r > 0$, representing a shift of 2602 gene families from the negative to the positive tail of the distribution relative to the expected equal number of positively and negatively correlated families ($\chi2 = 2189.608$, $p \approx 0$; figure 2.1a, inset). This result demonstrates a highly pronounced over-representation of gene families displaying a positive association between GFS and Ei. This observation is not explained by an overall expansion in gene number across species in line with Ei ($r = 0.251$, $p = 0.127$), but rather by an over-representation of small gene families among those highly associated with encephalization, combined with few larger gene families displaying decreases in size.

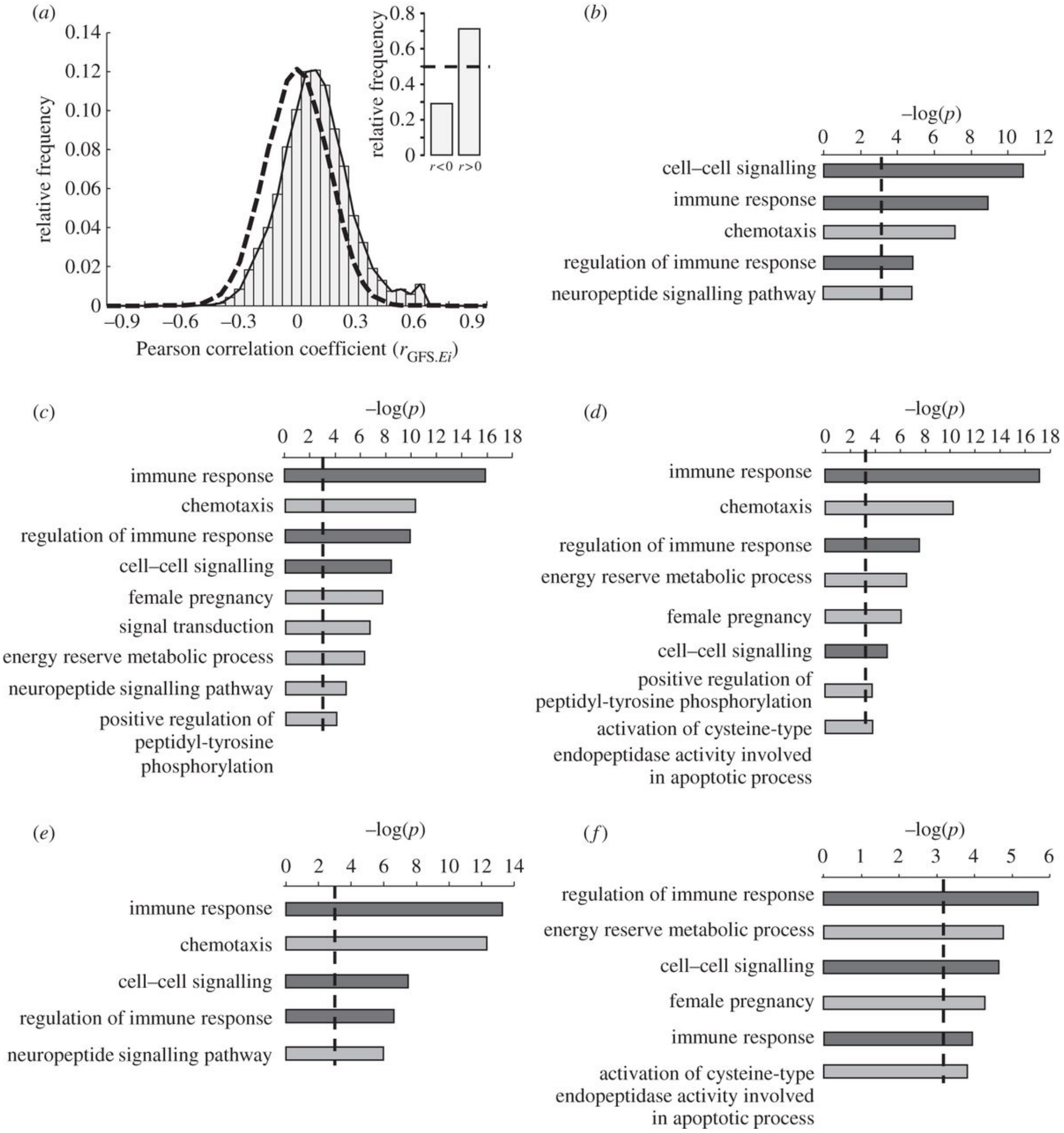

Figure 2.1.Enrichment of gene family size variations in line with increased encephalization in mammals.

(a) Histogram showing the distribution of correlation coefficients for GFS and Ei in 12 373 gene families encompassing 39 mammalian genomes. A randomization-based estimation of the expected distribution is represented by the dashed line. Inset: distribution of positive and negative correlations relative to the expected distribution (dashed line). (b) Deviations from random expectations in the mean correlation coefficient of gene families associated with individual GO terms (expressed as −log(p-value)). Only GO categories with a significant bias are shown. (c) Over-representation of GO terms among gene families most significantly associated with encephalization ($p < 0.05$, $n = 1292$). (d) GO enrichment analysis among the families displaying the most significant correlation with encephalization after removing all families with a stronger association with MLSP than with Ei (n = 927). (e) GO-terms enrichment analysis among gene families with the most significant positive partial correlation coefficients for Ei after controlling for the contribution of MLSP in a multiple regression analysis (n = 713). (f) GO-terms enrichment analysis among gene families with the most significant positive partial correlation coefficients for Ei with no significant phylogenetic interdependence (n = 501). Bonferroni-corrected significance thresholds are indicated with a dashed line. Dark bars indicate common GO terms across all five analyses.

We next asked whether the observed enrichment of Ei-related GFS variations was unspecific in terms of the gene populations involved or, alternatively, if this enrichment occurred in gene families specifically associated with certain biological functions. To this end, we used functional GO annotations for ‘biological processes’ and carried out two complementary tests to assess deviations (from random expectations) in the distribution of GO terms associated with gene families displaying a high correlation between GFS and Ei. First, we examined whether there were any significant deviations in the mean correlation coefficient of gene families associated with individual GO terms (see Methods). Out of all 260 functional categories included, only gene families associated with cell–cell signalling, immune response, chemotaxis, neuropeptide signalling pathways and regulation of immune response displayed a significantly higher than expected average correlation values, between GFS and Ei, after Bonferroni correction (figure 2.1b). By contrast, no significant bias was observed in functional categories containing families with negative average correlations (not shown).

Second, we measured over-representation of GO terms among the gene families whose GFS variations were most significantly associated with Ei ($r > 0$, $p < 0.05$, $n = 1292$). Among these families, we found that GO terms for immune response, chemotaxis, regulation of immune response, female pregnancy, cell–cell signalling, signal transduction, energy reserve metabolic processes, positive regulation of peptidyl-tyrosine phosphorylation and neuropeptide signalling pathways were significantly over-represented after Bonferroni correction (figure 2.1c). No GO terms were found to be significantly over-represented among gene families with the highest negative covariance between GFS and Ei (not shown). Taken together, these results show that the observed collective variation in GFS in line with encephalization is not randomly distributed across functional categories but is significantly pronounced in families associated with specific biological functions.

2.4.2 Association between gene family size and encephalization is not explained by lifespan variations

A number of studies on brain evolution have uncovered a robust relationship between relative brain size and lifespan (Allen et al. 2005; Barrickman et al. 2008; Gonzalez-Lagos et al. 2010). In agreement with this, we found a strong association between MLSP and Ei among the species included in this study ($r = 0.7912$, $p < 10−8$). Thus, the observed associations between Ei and GFS could be secondary to an underlying association between MLSP and GFS. Of the 1292 most significantly correlated families ($r > 0$, $p < 0.05$), 927 displayed a stronger association with Ei than with MLSP (r(Ei, GFS) > r(MLSP, GFS)), thereby suggesting a preferential contribution of Ei to the observed bias in the correlation distribution ($\chi 2 = 858.74$, $p = 3.3572e−187$, relative to a random equal distribution of stronger associations). GO enrichment analysis was then repeated including only these 927 families revealing a significant over-representation of gene families associated with immune response, chemotaxis, regulation of immune response, energy reserve metabolic processes, female pregnancy, cell–cell signalling, positive regulation of peptidyl-tyrosine phosphorylation and activation of cysteine-type endopeptidase activity involved in apoptotic processes (figure 2.1d). It is worth noting that the complementary GO enrichment analysis carried out on gene families with both the most significant association between MLSP and GFS ($r > 0$, $p < 0.05$) and a stronger association with MLSP than Ei (r(MLSP, GFS) > r(Ei, GFS), $n = 1321$), resulted in no significant enrichment of any GO category. These results shows that enrichment of specific GO terms occurred only among gene families preferentially associated with degree of encephalization, whereas GFS variations potentially associated with increased MLSP showed no significant association with any particular functional category.

Because MLSP may still partly explain the covariance between GFS and Ei even if the correlation coefficient of GFS with Ei is higher than with MLSP, we used multiple regression analysis to obtain partial correlation coefficients between GFS and Ei after controlling for the contribution of MLSP (see Methods). GO terms enrichment analysis was then carried out only among those gene families with the most significant positive partial correlation coefficients (partial $r > 0$, $p < 0.05$, $n = 713$). This analysis revealed a significant enrichment of families functionally associated with regulation of immune response, chemotaxis, cell–cell signalling and neuropeptide signalling pathways (figure 2.1e). These results show that variations in GFS specifically associated with encephalization (i.e. not accounted for by variations in MLSP) are also specifically associated with distinct biological functions.

2.4.3 Phylogenetic relatedness does not explain the observed bias in the distribution of gene families associated with encephalization

For a given gene family, any association between Ei and GFS could be the secondary to existing phylogenetic relationships among the species analysed, as in the absence of any selective forces, closely related species will tend to have both similar degrees of Ei and similar GFS (Freckleton et al. 2002; Pagel 1999). In order to determine the degree to which phylogenetic effects contribute to the observed shift in the correlation distribution, we used a PGLS approach (see Methods) (Freckleton et al. 2002; Pagel 1999). Out of 713 gene families with the most significant positive partial correlation coefficients between Ei and GFS (after correcting for MLSP, see previous analysis), we found a total of 501 gene families for which phylogenetic relationships among species could not account for the covariance between GFS and Ei. Among these families, we observed a significant over-representation of gene families associated with regulation of immune response, cell–cell signalling, energy reserve metabolic processes, female pregnancy and activation of endopeptidase activity involved in apoptosis (figure 2.1f). These findings demonstrate that the over-representation of specific biological functions among those gene families most strongly associated with higher Ei is neither explained by the known association between MLSP and Ei nor by existing phylogenetic relationships among the species analysed.

2.4.4 Gene families with size increases in line with encephalization show expression signatures consistent with brain functions

To assess whether gene family variations in line with encephalization were directly associated with brain function, we characterized the potential relationship between Ei-associated GFS variations and patterns of gene expression in the human nervous system. For this analysis, we selected the top 501 Ei-associated gene families with both the most significant partial correlation coefficient between Ei and GFS and no significant phylogenetic effects (figure 2.1f). Using available expression data from the Illumina human body map (see Methods), we looked at the possible over-representation of genes highly expressed in the human brain within the selected 501 gene families. Individual genes were categorized as prominently expressed in the brain if their expression level in this tissue was the highest or second highest among all 16 tissues included (top 12.5th percentile). Statistical significance was assessed by comparing with equally sized random samples drawn from the wider pool of gene families (see Methods). This analysis revealed a significant enrichment, within these gene families, of genes prominently expressed in the brain (figure 2.2a). By contrast, no significant enrichment of genes prominently expressed in the brain was detected among those gene families with the strongest association with MLSP and no significant phylogenetic effects (figure 2.2a).

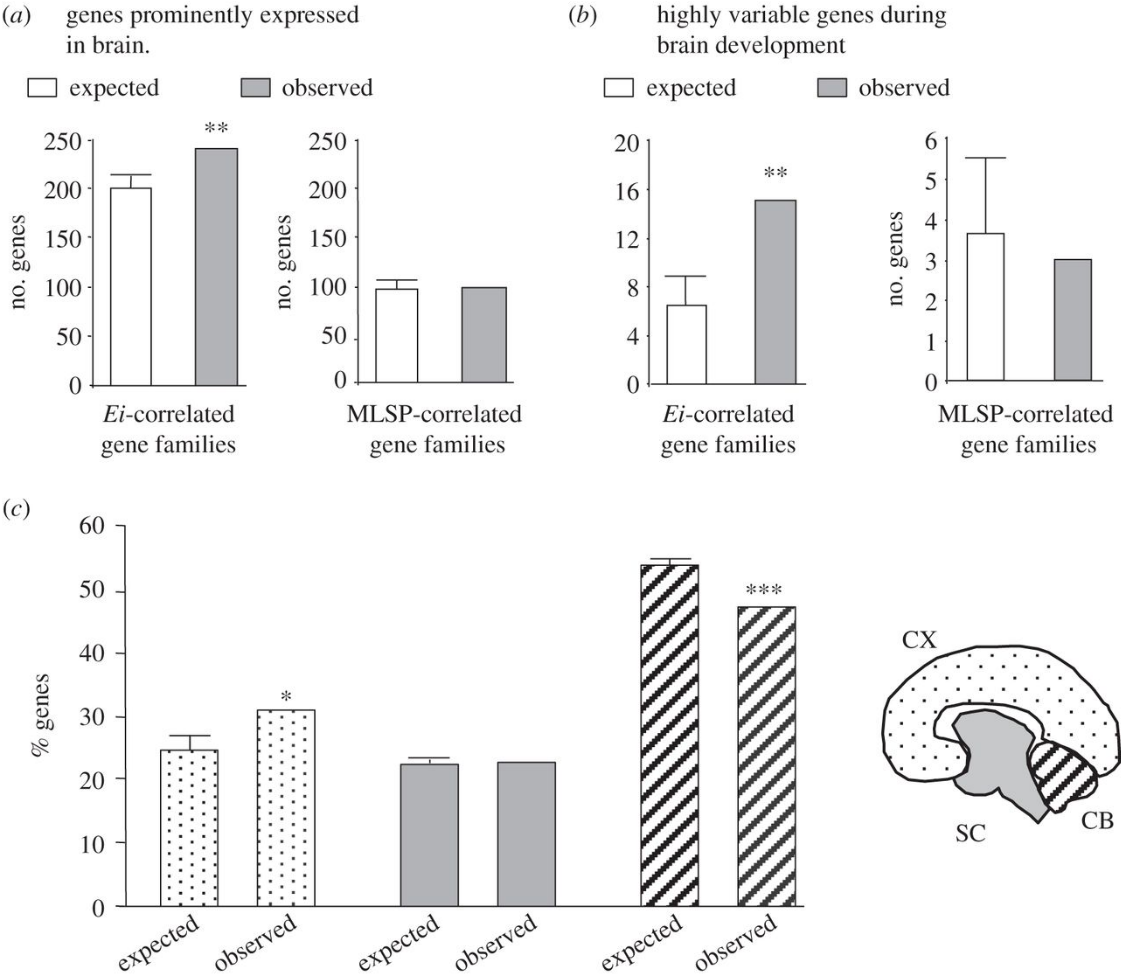

Figure 2.2. Relationship between Ei-associated GFS variations and patterns of gene expression in the human nervous system.

(a) Over-representation of genes prominently expressed in the human brain (top 12.5th percentile) among the top 501 Ei -associated or the top MLSP-associated gene families compared to random expectations. (b) Over-representation of genes displaying the highest expression variance during human cortical development relative to adulthood among the top Ei-associated or the top MLSP-associated gene families. (c) Percentage of genes maximally expressed in the cortical (CX), subcortical (SC) and cerebellar (CB) regions respectively. Expected values (mean ± s.e.m.) were numerically determined using sized-matched random samples of genes drawn from the wider pool of gene families. *p < 0.01; **p < 0.001; ***p < 0.0001.

Genes involved in cortical development have been shown to display higher variance in expression level during the developmental period of the cerebral cortex compared with adulthood (Sterner et al. 2012). We therefore looked at the possible representation of genes displaying the highest expression variance during human cortical development relative to adulthood, as defined by Sterner et al. (Sterner et al. 2012), within the same 501 gene families and found a significant enrichment of genes displaying this pattern of expression (figure 2.2b). By contrast, no significant enrichment of these same genes was observed among the top MLSP-associated gene families (figure 2.2b).

We next asked whether there was any statistical bias in the relative expression of Ei-associated gene families across different brain regions. Using human brain RNA-seq data from the BrainSpan dataset (see Methods), we obtained the average expression for each gene in the cortex, subcortical regions or cerebellum and split them into three categories according to the region where the highest average expression was found. This analysis revealed a statistically significant enrichment, among those genes contained within the top 501 Ei-correlated gene families, of genes maximally expressed in the cortex (figure 2.2c). No significant enrichment of genes maximally expressed in subcortical regions was observed among these families. By contrast, genes maximally expressed in the cerebellum were found to be significantly under-represented among the top Ei-correlated gene families. Taken together, these results reveal that gene families displaying the highest association between Ei and GFS are enriched in genes that are prominently expressed in the brain, with maximal expression in the cortex and display an expression signature distinctly associated with cortical development.

In order to characterize further the cortical expression profile of Ei-associated gene families, we used a weighted gene co-expression network analysis approach to identify modules of co-expression among genes contained within the top 501 Ei-correlated gene families. Using human developmental expression data derived from the BrainSpan dataset, we identified 18 modules (figure 2.3a) associated with distinct temporal patterns of expression. Figure 3b shows the time course of expression of six of these modules summarized by the eigengene associated with each module's co-expression matrix. Some of these modules showed the highest expression levels during the early or late foetal period followed by a progressive decline in expression levels with age. This trend may reverse in some instances in late-adult stages (black module, figure 3b) or show a progressive increase throughout development as illustrated by the yellow module.

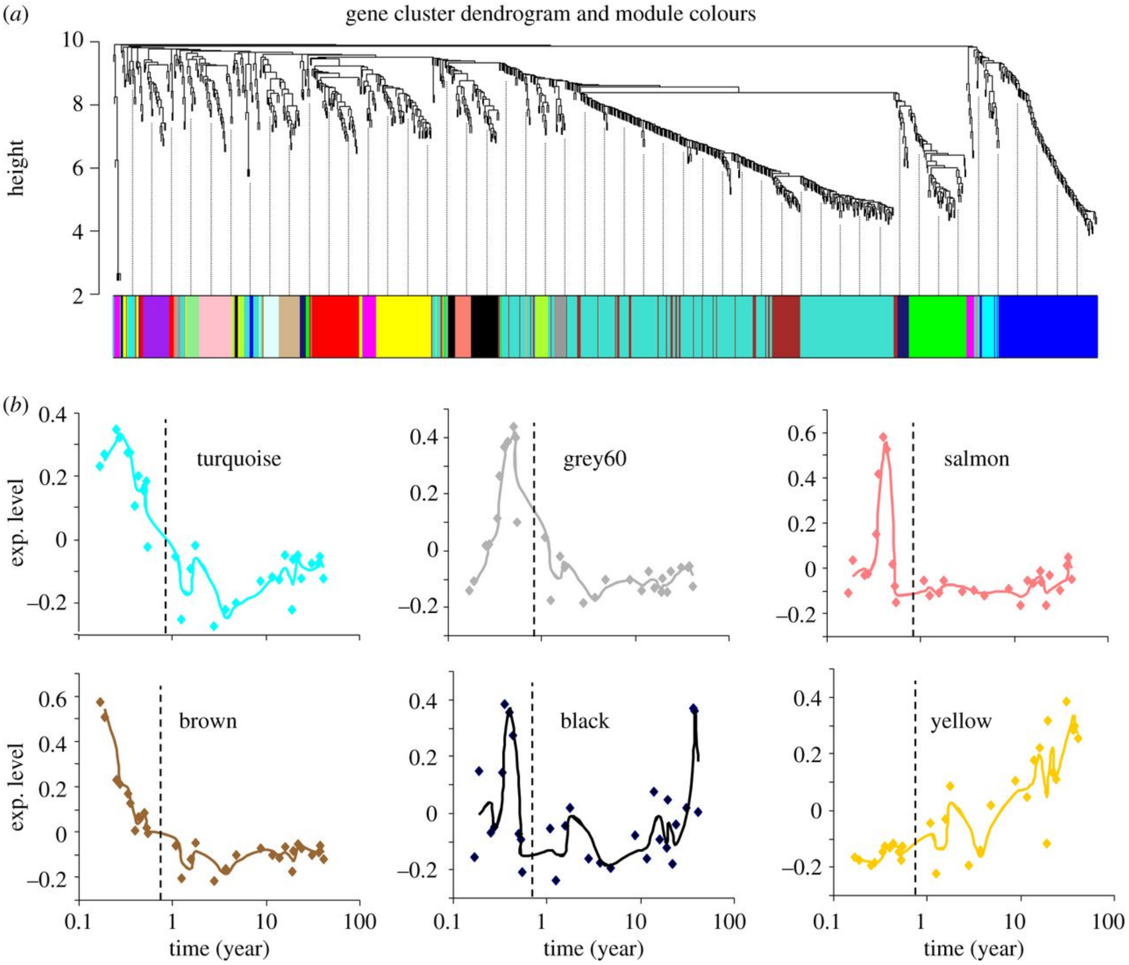

Figure 2.3. Temporal patterns of cortical expression of Ei-associated gene families.

(a) Weighted gene co-expression network analysis was used to detect co-expression modules among genes contained within the top 501 Ei-associated gene families using human brain temporal expression data, revealing 18 co-expression modules (coloured). (b) Developmental time course of expression of six representative modules summarized by the level of expression of the eigengene associated with each module's co-expression matrix. Birth point is indicated with a dashed line.

2.5 Discussion

Our results reveal a highly significant over-representation of gene families displaying a positive association between GFS and level of encephalization. This bias occurs most prominently in families associated with specific biological functions. The most robust and consistent bias was observed in gene families associated with cell signalling, immune regulation and chemotaxis.

While chemotaxis and cell signalling functions are known to play central roles in the nervous system, the significance of the observed enrichment of immune system-associated functions among gene families displaying the highest association between GFS and Ei is less clear. In recent years, however, signalling and regulatory mechanisms originally described in the immune system have increasingly been found implicated in key neural-specific roles both in the developing and adult nervous system (Crampton et al. 2012; Gavalda et al. 2009; Gutierrez et al. 2005; Gutierrez et al. 2008; McKelvey et al. 2012; Nolan et al. 2011; O'Keeffe et al. 2008). In addition, in the human cerebral cortex, immune system-related functions have been found to be significantly over-represented among genes displaying higher expression variability in the developing cerebral cortex than in the adult (Sterner et al. 2012), suggesting a substantial involvement of immune-related signals during cortical development.

Our results, showing a significant over-representation of immune-related functions among Ei-associated gene families, support the notion of an underlying and substantial overlap in the regulatory and signalling machinery shared by both the immune and nervous system and in particular during development of the latter.

One possible interpretation is that the observed enrichment of immune-related functions among Ei-associated gene families reflects an underlying expansion of immune surveillance in mammals that could be in some way permissive to increased encephalization. While we cannot rule out this possibility, at present, there is little evidence in support of any systematically pronounced and sustained expansion of immune functionalities in mammalian lineages (Boehm 2012). An alternative interpretation is that signalling and regulatory molecular components that were originally involved in immune-specific functions became gradually recruited by the nervous system in response to the developmental and functional demands of increasingly more complex brains.

The observed association between degree of encephalization and variations of GFS in a large number of gene families is further supported by our finding that Ei-associated gene families display a transcriptional signature consistent with brain-specific functions. Indeed, among the gene families most highly correlated with encephalization with no significant phylogenetic effects, we found a statistically significant enrichment of genes prominently expressed in the brain, strongly indicating that these genes are under comparably higher demand in the nervous system relative to other tissues. When restricting the analysis to the relative expression levels within central nervous system regions, we found that these families are enriched in genes prominently expressed in the cortex, suggesting that Ei-correlated changes in GFS may have played a substantial role supporting key aspects of cortical evolution. In this regard, it is worth noting that brain evolution in mammalian lineages is characterized by a disproportional expansion of the brain cortex (Kaas 2013; Nomura et al. 2013). Analysis of the developmental pattern of expression of these families in the human cortex showed that these genes are organized in co-expression clusters or modules with distinct temporal profiles suggesting a substantial involvement of these families in the developmental organization of the brain.

Genes with the highest degree of connectivity within a module are termed hub genes and are expected to be functionally important within the module. By way of illustration, we examined the turquoise module (figure 2.3b) and identified a member of a zinc finger gene family (gene family ID: ENSFM00620000999432) as its main hub gene. Interestingly, all but two of the 20 members of this gene family in humans are contained within the same co-expression module. Because genes contained within a co-expression module are thought to be functionally related (Lee et al. 2004; B. Zhang and Horvath 2005), the fact that most members of this zinc finger family are found within the same co-expression module strongly suggests that these genes are functionally related during brain development. We reconstructed the phylogenetic tree of this family and found that the observed pattern is the result of a combination of events of gene loss and gene gain from an original set of four ancestral proteins at the base of the mammalian evolution, overall resulting in a steady increase in the number of gene family members in line with increased level of encephalization ($r = 0.7547$, $p = 2.86 \times 10{-}8$).

2.6 Conclusion

In this study, we have found a significant over-representation of GFS variations in line with increased encephalization in mammals. Importantly, this relationship is not accounted for by known correlates of brain size and is not explained by phylogenetic relatedness. The observed bias occurs most prominently in families preferentially expressed in the brain, in particular the cortex, and significantly associated with distinct biological functions.

Based on our results, we propose that variations in GFS associated with encephalization provided an evolutionary support for the specific cellular, physiological and developmental demands associated with increased brain size in mammals.

Acknowledgements.

We thank Leticia Ramirez-Lugo from the National University of Mexico for her technical support and helpful comments from two anonymous reviewers.

Funding statement.

This work was supported by CONACyT PhD scholarships to A.C.M. and J.M.S., Dorothy Hodgkin Research Fellowship and Royal Society Research Grant to A.U.O. and University of Lincoln support to H.G.

Supplementary Table S2.1

Species	MLSP[1]	Brain Mass[2-13]	Body Mass[2-13]	*Ei*[14]
Ailuropoda melanoleuca	36.8	235.1	117920	-2.01376
Bos Taurus	20	456	520000	-2.30092
Callithrix jacchus	16.5	7.24	280	-1.62664
Canis familiaris	24	100.9	19240	-1.69931
Cavia porcellus	12	4.28	971	-2.94818
Choloepus hoffmani	37	28.5	4000	-1.95829
Dasypus novemcinctus	22.3	12	3700	-2.77339
Dipodomys ordii	9.9	1.97	54	-1.87492
Echinops telfairi	19	0.52	60	-3.27431
Equus caballus	57	650.03	441175	-1.84119
Erinaceus europaeus	11.7	3.77	697	-2.86287
Felis catus	30	28.4	2500	-1.661
Gorilla gorilla	55.4	438.18	122500	-1.41552
Homo sapiens	122.5	1300	65000	0.151656
Loxodonta africana	65	4480	2750000	-1.08197
Macaca mulatta	40	97.45	8250	-1.19216
Macropus eugenii	15.1	23.7	4425	-2.20734
Microcebus murinus	18.2	1.68	50	-1.9849
Monodelphis domestica	5.1	0.95	100	-2.9986
Mus musculus	4	0.45	24	-2.83246
Myotis lucifugus	34	0.175	8	-3.07381
Nomascus leucogenys	44.1	7000	119.4	-0.88387
Ochotona princeps	7	2.39	169	-2.41184
Ornithorhynchus anatinus	22.6	9.22	1030.3	-2.21869
Oryctolagus cuniculus	9	9.14	1411.8	-2.42902
Otolemur garnettii	18.3	10.45	946.7	-2.03931
Pan troglodytes	59.4	371.05	45500	-0.94796
Pongo pygmaeus	59	343	36900	-0.89249
Procavia capensis	14.8	20.5	3800	-2.25494
Pteropus vampyrus	20.9	9.53	1060	-2.20381
Rattus norvegicus	5	2.38	339	-2.86154
Sarcophilus harrisii	13	16.24	6126.8	-2.79237
Sorex araneus	3.2	0.23	8.4	-2.83174
Spermophilus tridecemlineatus	7.9	3.2	175	-2.14231
Sus scrofa	27	180.2	158320	-2.46825
Tarsius syrichta	16	3.5	117	-1.79503
Tupaia belangeri	11.1	3.1	150	-2.0754
Tursiops truncatus	51.6	1679.6	180910	-0.32137
Vicugna pacos	25.8	188	50000	-1.68822

1) AnAge database (http://genomics.senescence.info/species/). **2)** Domaradzka-Pytel et al., 2007. **3)** Dunbar and Shultz, 2007b. **4)**Garwicz et al., 2009. **5)** Gilmore et al., 2000. **6)** Gittleman, 1986. **7)** Herculano-Houzel et al., 2007. **8)** Leonard et al., 2007. **9)** McNab and Eisenberg, 1989. **10)** Sacher and Staffeldt, 1974. **11)** Stephan et al., 1981. **12)** Wang et al., 2008. **13)** Weisbecker and Goswami, 2010. **14)** Gonzalez-Lagos et al., 2010

3. Neocortex expansion in mammalian lineages explains gene family size variations associated with larger brain size

Atahualpa Castillo-Morales[1, 2], Alexandra de Sousa[3], Jimena Monzon-Sandoval[1, 2], Jessica Stevens[1], Araxi O. Urrutia[1*] and Humberto Gutierrez[2*]

1 Department of Biology and Biochemistry, University of Bath, Bath, BA2 7AY, UK

2 School of Life Sciences, University of Lincoln LN6 7TS, UK

3 Department of Psychology, University of Bath, Bath, BA2 7AY, UK

ACM: Acm39@bath.ac.uk

AS: aads20@bath.ac.uk

JMS: Jms52@bath.ac.uk

AOU: a.urrutia@bath.ac.uk

HG: hgutierrez@lincoln.ac.uk

* To whom correspondence should be addressed.

3.1 Abstract

Enlarged brain size in humans and other mammals is a key trait and is related to behavioural complexity. Most of the observed variations in brain size are the result of neocorticalization: the expansion, relative to the rest of the brain, of the neocortex, a key mammalian specific brain structure which has been associated with higher cognitive processes. What at the genomic level underlies the morphological evolution of the neocortex remains poorly understood. By comparing the genomes of 28 mammalian species, we show that neocortical expansion is associated with variations in gene family size which are significantly enriched in cell-cell signalling and immune response functional annotations. Moreover, we find that previously reported gene family size variations associated with increased brain size are largely accounted for by the link between neocortex ratio and gene family size variations. These results suggest that variations in gene family size underlie morphological adaptations during brain evolution in mammalian lineages.

3.2 Introduction

Increased brain size represent a key innovation of mammals and is thought to have played an important role in the expansion of these clade. Brain size is a key zoological trait and has been related to increased behavioural complexity and the ability to cope with changing environment (Deaner et al. 2007; S. M. Reader and Laland 2002) . Larger brains however, are associated with high metabolic cost (Aiello and Wheeler 1995; Fish and Lockwood 2003; Isler and van Schaik 2006a; Navarrete et al. 2011) and are often associated with longer times to reach maturity and higher parental investment (Barrickman et al. 2008; Barton and Capellini 2011; Deaner et al. 2003; Finarelli 2010; Isler and van Schaik 2009).

Brain size is a highly variable trait both among mammalian and non-mammalian species with marked differences observed even between relatively close species (e.g. (Aristide et al. ; Harvey et al. 1980; Huber et al. 1997; Kotrschal et al. 1998; Sol and Price)). Brain size closely scales with variations in body mass across species (R. D. Martin 1990). Calculating an encephalization index -correcting brain size by body mass in order to express the increase in brain size beyond that expected due to the brain-body allometric relationship-ranks humans, apes and some cetaceans at the top of the list as the most enchephalised mammalian species, aligning more closely with behavioural capacity (Jerison 1985; R. D. Martin 1983).

Encephalization is not associated with proportional expansion of all brain structures. In mammals, most variations in encephalization indexes are largely explained by changes in the size of the neocortex (Jerison 1973, 1990) .

The neocortex is a brain structure unique to the mammalian brain that envelops the cerebrum and plays a key role in higher cognitive functions. Also known as the isocortex, the neocortex is the newest part of the cerebral cortex, and shares a common origin with the dorsal cortex in reptiles (Kaas 2011; Medina and Reiner 2000; Molnár et al. 2007; Northcutt and Kaas 1995), perhaps originating between 220 and 280 mya, after the split between synapsids and sauropsids, and before this latter group gave rise to the mammalian lineage (Kaas 2011; Northcutt and Kaas 1995; M. A. O'Leary et al. 2013). It is composed of six layers of cells, mainly excitatory pyramidal neurons, inhibitory interneurons and glial cells, which differ in density and size of neural cell bodies and axons, and in humans contains on average around 24 billion neurons and 33 billion glial cells (Pelvig et al. 2008). The neocortex encompasses most primary motor and sensory areas, as well as association areas which together process and regulate sensory perception and motor commands.

The increase in the size of the neocortex relative to the rest of the brain, called neocorticalization, a process that appears to have taken strength noticeably around 60 million years ago, when the major radiations of marsupial and placental mammals began (Northcutt and Kaas 1995), has long been a leading consideration when evaluating the evolution of mammalian brain form (Anthony 1938; Sawaguchi and Kudo 1990; Wirz 1950). These changes have been associated with more complex cortical processing thus allowing the emergence of new behaviours (Kaas 1989). Neocortex to brain size ratio is correlated with social group size (Dunbar 1992), and it has been speculated that the number of neocortical neurons is a limiting factor in determining the number of social relationships which an animal can monitor(Dunbar 1992) .

Neocortex size has been associated with the diversification of highly specialised association areas (Changizi 2001). In the hominid lineage, the expansion of the neocortex is thought to have played a key role for the evolution of modern humans (DeFelipe 2011): The neocortex constitutes 90% of the cerebral cortex in humans (Noback et al. 2005) and has been the focus for studies investigating the neurocorrelates of human-specific behaviour (Aiello and Dunbar 1993; Barker 1995; Barton 1996; Dunbar 1993; Passingham and Wise 2012). Classically, the neocortex has been regarded at least partially, as the seat for the neurobiological mechanisms of so called higher cognitive abilities, such as self-awareness, consciousness, abstract reasoning and planning, in mammals (Crick and Koch 1990; Eccles 1994; P. Gilbert et al. 1995; Grober et al. 1992; Steven M. Platek et al. 2004; S. M. Platek et al. 2008; Sugiura et al. 2005). Particular areas in the human neocortex have been found to drive the understanding and production of language (Aiello and Dunbar 1993; Letinic et al. 2002). Other highly specialised areas, such as those for identifying faces (Allison et al. 1994; Nestor et al. 2011), or places (V. M. Miller and Best 1980; Poucet et al. 2003), have also been identified in the neocortex. Furthermore, mirror-neurons which have been implicated in social learning (they respond both to doing an action or seeing it be done by other individuals, and seem to play a part in action understanding, speech perception, emotion recognition and imitation) (Adolphs et al. 1994; Enticott et al. 2008; Iacoboni

2005; Schulte-Ruther et al. 2007; Spaulding 2013; van der Gaag et al. 2007; Wicker et al. 2003), have been identified in the neocortex (Gallese et al. 2002; Molenberghs et al. 2009; Rizzolatti et al. 1996).

Despite the importance of the neocortex in mammalian evolution, molecular mechanisms controlling the development of brain structures and the genomic features underlying their evolution remain poorly understood (de Sousa and Proulx 2014; Hawrylycz et al. 2012). So far, there have been few efforts to identify features reflecting the genomic impact of brain evolution. While a previous attempt to detect a genomic signature of the evolution of brain reported a widespread accelerated sequence evolution of genes functioning in the nervous system during human origins (Dorus et al. 2004), this claim was heavily contested soon after (Kosiol et al. 2008; Shi et al. 2006). By conducting a genome-wide analysis of amino acid composition across 37 fully sequenced mammalian genomes, Gutierrez et al. showed that encephalization is significantly correlated with overall protein amino acid composition, possibly reflecting the selective demands imposed by a larger brain (Gutierrez et al. 2011).

Changes in gene family size can reflect changes in the relative relevance of specific functions in an organism. Duplication events have been proposed to be a vital driving force for many evolutionary changes by providing source of material for the origin of novel gene functions and expression patterns, whilst gene loss is suggested to act as response to selection (Krylov et al. 2003; Lynch and Conery 2000). Marked differences in gene family size have been identified in drosophila and vertebrates with families experiencing the largest changes being enriched in specific functions (Demuth et al. 2006; Hahn et al. 2005; Hahn et al. 2007). Among mammals, marked variations in the number of olphactory receptors likely reflecting variations in the reliance of different lineages in their sense of smell to find food and or avoid predators (Hoover 2013; Kajiya et al. 2001; Malnic et al. 1999; Niimura et al. 2014). A recent study found that encephalisation in mammalian lineages is associated with significant variations in gene family size (Castillo-Morales et al. 2014) with the most positively associated gene families were significantly enriched in several functional categories including immune system response, chemotaxis and cell-cell signalling. Here we investigate if the variations in neocortex to brain ratio in mammalian lineages are associated with changes in gene family size. By tracing back events of gene gain and loss per gene family we are able to distinguish gene family expansion and gene family contraction in association with neocortification. We further explore the extent to which any changes in gene family size associated with neocortex explain previously reported variations in gene family size and encephalisation. .

3.3 Methods

3.3.1 Gene Family annotations

Annotated gene families encompassing 28 fully sequenced mammalian genomes were obtained from Ensembl release 76 (Cunningham et al. 2015; http://www.ensembl.org). In the context of this annotation, Ensembl families are defined by clustering all Ensembl proteins along with metazoan sequences from UniProtKB. Any given gene family constitutes a group of related genes that includes both paralogs within the same species and orthologues and paralogs from other species. Any given gene can only be assigned to a single gene family. Gene family size (GFS) represents the total number of genes per gene family. In order to maximize the number of families covered in this study we included all gene families with members present in no less than six of the 28 mammalian species (n=11943). We excluded any family with no variance in GFS across this species.

3.3.2 Encephalization index, Neocortex ratio and Maximum Lifespan

Because larger species have larger brains, it is necessary to estimate brain mass controlling for the allometric effect of body size. Size-corrected values of brain mass (Ei) were computed as log [brain mass/body mass^{b}]. The slope (*b*) was estimated as 0.64 by Gonzalez-Lagos et al. (Gonzalez-Lagos et al. 2010) based on a log–log least squares linear regression of brain mass against body mass over 493 mammalian. Neocortex volumes, taken from the literature, included the grey and white matter of the cerebral cortex; the grey matter of the paleocortex (entorhinal cortex, schizocortex, hippocampus and amygdala) was segmented out. Neocortex ratio (Nr) was defined as (neocortex volume in cm^{3})/(brain volume in cm^{3}

- neocortex volume in cm^3) after Dunbar (Dunbar 1992) . Maximum lifespan (MLSP) for each species was obtained from the animal ageing and longevity database (Tacutu et al. 2013). Raw values of brain and body mass, as well as neocortex and non-cortical brain volume and maximum life span (MLSP) for the analysed species with their corresponding references are presented in Supplementary Table 3.1.

3.3.3 Correlation coefficients of gene family size and phenotypes

Simple Pearson correlations between GFS and Ei, Nr or MLSP were carried out using R - based statistical functions. Ten thousand (10 000) Monte-Carlo randomizations of the phenotypes to determine statistical significance of the distribution of Pearson correlation coefficients were conducted for each phenotypic variable.

In order to control the potential contribution of each of the other confounding variables on the relationship with each phenotype and GFS, partial correlation coefficients were computed for each gene family including the other two as co-variates. The resulting partial coefficient represents the contribution of each phenotype to the variance in gene family size which is not explained by variations in the other two phenotypes.

3.3.4 Gene Ontology terms enrichment

Biological Process Gene Ontology annotations for each species were obtained from Ensembl’s Biomart release 76 (Cunningham et al. 2015). In the present study, a particular GO term was associated to a family whenever that term was linked to any of its members in any species. To minimise the effect of very small functional categories, only terms linked with more than 200 families were examined (n=116). Gene families annotated to any GO term with less than 200 families were assigned to a “Small biological process GO terms” category. Gene families not annotated to any GO term in any species were grouped into a “Not annotated” category. Enrichment analysis of these GO terms was carried out as described in (Castillo-Morales et al. 2014). In brief, over-representation of genes associated to specific GO terms was assessed by counting the number of gene families assigned to each GO term within the analysed set of gene families. Statistical significance was numerically assessed by obtaining the expected number of families per GO in 1000 equally sized random samples derived from the overall population of gene families. Because genes vary in the number of GO terms associated to them, we adjusted for differences in the density of GO annotations between the test and background samples, by dividing the family counts per GO from each sample, by the samples' average number of GO annotations per family.

3.3.5 Phylogenetically controlled regression

Further to unpicking the contribution of each phenotype to GFS, in order to account for the phylogenetic non-independence of taxa on the relationships of morphological traits with size, we used phylogenetic independent contrasts (PIC) analysis (Felsenstein 1985). First, for each phenotype we obtained a residual by regressing it against the other two in a multivariate lineal model. The same regression was performed on GFS of each family for each variable, again, using the other two as predictors.

PIC for both the partial correlation coefficients for both the phenotype and GFS of each family were computed using the ape package in R. Pearson correlation coefficients were then assessed between each of these partial correlation coefficients for phenotypes and phenotypes and GFS for each family. This correlation coefficient reflects the degree of association between each phenotype and GFS when both confounding variables and phylogenetic non-independence are controlled.

Ultrametric phylogeny of the 28 analysed mammalian species obtained from TimeTree website (http://www.timetree.org/ ; S. Kumar and Hedges 2011)

3.3.6 Gene expression prior and after peak in neocortex thickness.

RNA-seq RPKM normalized expression data summarized to genes was obtained from NIMH Transcriptional Atlas of Human Brain Development (http://brainspan.org ; J. A. Miller et al. 2014). A selection of 143 samples corresponding to 11 cortical regions and 13 different ages were chosen. The cortical regions include primary auditory cortex (core) (A1C), Dorsolateral prefrontal cortex (DFC), Posteroinferior (ventral) parietal cortex (IPC), inferolateral temporal cortex (area TEv, area 20) (ITC), primary motor cortex (area M1, area 4) (M1C), Anterior (rostral) cingulate (medial prefrontal) cortex (MFC), Orbital frontal cortex (OFC), Primary somatosensory cortex (area S1, areas 3,1,2) (S1C), Posterior (caudal) superior temporal cortex (area TAc) (STC), Primary visual cortex (striate cortex, area V1/17) (V1C) and Ventrolateral prefrontal cortex (VFC). The sample covered developmental stages 16, 24, 37 post conception weeks, 4 months after birth and 1, 3, 8, 13, 19, 21, 30, 36 and 37 years old. Gene expression data was further normalized against the total expression per sample, and divided in two groups, corresponding to the periods prior and after the peak in cortical thickness occurs (around 13 years) (Shaw et al. 2008). For each gene, we average the expression across stages and structures of the same group and compare the differences between matched samples via a Wilcoxon test. P values were adjusted for multiple testing using a Bonferroni correction

3.4 Results

In order to assess the association between gene family size and neocortex expansion, we compiled data on neocortex to brain volume ratio (Nr) from the literature for 28 mammalian species with fully sequenced genomes (Table 1). Gene family size (GFS) was assessed for a total of 11943 non-overlapping families. Correlation coefficients between GFS and Nr were then calculated for each gene family. We found an excess of positive associations between GFS and Nr (Figure 3.1) (χ^2 = 2973.263083, p < $1x10^{-20}$). A Monte Carlo simulation showed that the shift towards positive values is significant (Z_{Nr} = 2.225819868, p = 0.013). This result shows a high over-representation of gene families displaying a positive association between GFS and Nr. This observation could result from combined duplications of gene families in lineages undergoing neocorticalization, as well as gene losses in species with a low neocortex ratio.

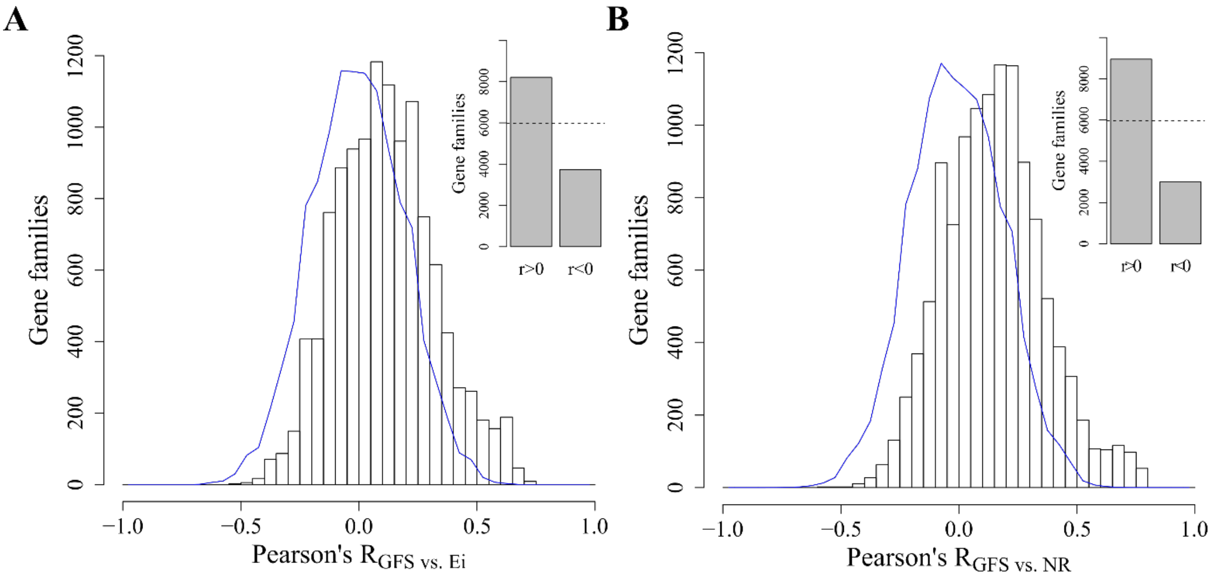

Figure 3.1. Enrichment of gene family size variations in line with increased encephalization index and neocortex ratio in mammals. (a) Histogram showing the distribution of correlation coefficients for GFS and Ei in 11943 gene families encompassing 28 mammalian genomes. (b) Histogram showing the distribution of correlation coefficients for GFS and Ei in 11943 gene families encompassing 28 mammalian genomes. In each figure, an estimation of the expected distribution derived from 10000 Monte Carlo simulations is represented by the blue line. Inset: distribution of positive and negative correlations relative to the expected distribution (dashed line).

In order to assess whether this shift in the distribution involved unspecific gene populations or instead involves genes associated with specific functional categories we examined functional annotations for the 1607 gene families whose sizes are most strongly associated with Nr ($r > 0$ and $p < 0.05$). For this, we obtained Gene Ontology (GO) functional annotations (Cunningham et al. 2015) and assessed the over-representation of individual GO terms among gene families with the strongest associations to Nr. A total of 17 GO functional categories were found to be significantly enriched (after correction for multiple tests (Benjamini-Hochberg correction)) among the gene families, with the strongest size associations with Nr being immune response, chemotaxis and cell-cell signalling among the top most overrepresented functional categories (Figure 3.2). Notably, genes with no functional annotations showed the highest over-representation.

As the Nr is known to be highly correlated to overall brain size, it is possible that the variations in gene family size with this variable are explained by the previously reported association between GFS and encephalization quotient (*Ei*), an index of brain size corrected by body mass (Castillo-Morales et al. 2014). We calculated correlation coefficients between GFS and Ei for each gene family in the same set of 28 species (Figure 3.1).We found a similar shift in the distribution favouring positive associations using the set of 28 species for which Nr could be calculated as that previously reported using a larger set of 39 species (Castillo-Morales et al. 2014). When contrasting Nr to *Ei*, we observed that the shift in the distribution of values is stronger for Nr (Significance of the deviation showed in line with encephalization is lower, Z_{Ei} = 1.70943, p = 0.044). When examining the functional associations for the set of gene families most significantly associated with encephalization quotient, 15 GO functional categories were found to be significantly enriched (Figure 3.2). As could be expected from the strong relatedness between Nr and Ei, we observed a high overlap in the sets of gene families most significantly associated with Nr and Ei with 75% of the gene families most significantly associated with Nr also being found among those most significantly associated with Ei. Of the 17 GO categories significantly enriched among families most associated with Nr, 14 were also found to be overrepresented among the gene families most significantly associated with Ei.

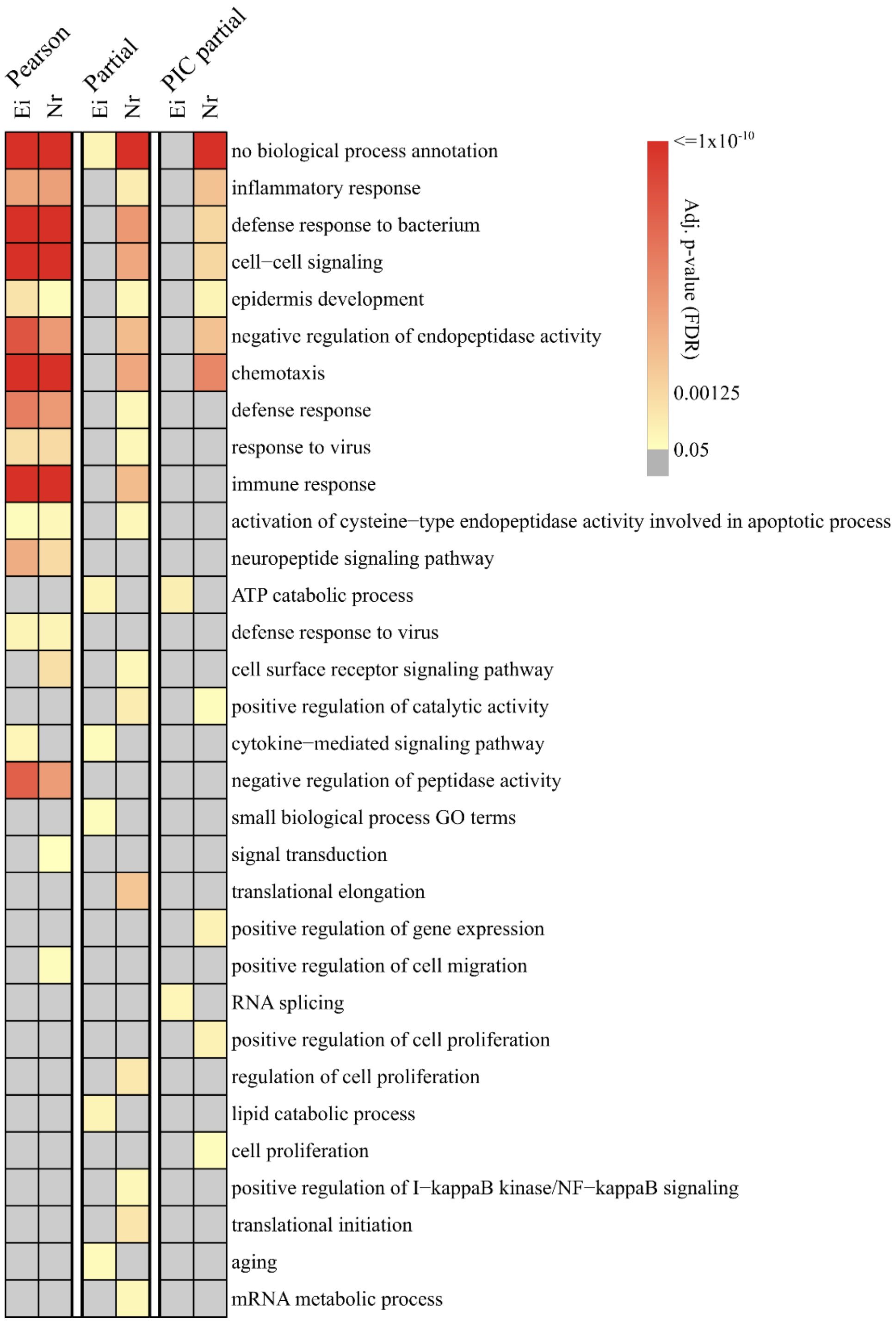

Figure 3.2. Gene Ontology enrichment analysis of families with GFS variations in line with Encephalization index and Neocortex ratio. Heatmap of the significance of the overrepresentation of GO terms (expressed as Benjamini-Hochberg-corrected p-value) among gene families most significantly associated with encephalization index and neocortex ratio. First two columns correspond to gene families with the most significant association between GFS and Ei or Nr respectively ($r_{Ei\ GFS} > 0$, $p < 0.05$, $n = 1323$ and $r_{Nr\ GFS} > 0$, $p < 0.05$, $n = 1607$). Third and fourth columns represent GO terms enriched among

gene families whose GFS variations display the most significant association with one of the brain phenotypes after accounting for the shared variance with the other neural phenotype using partial correlation and MLSP ($r_{\text{Ei GFS.Nr MLSP}}$ >0, $p < 0.05$, n = 132 and $r_{\text{Nr GFS.Ei MLSP}}$ >0, $p < 0.05$, n = 502). Fifth and sixth columns show enrichments after accounting for confounding variables, as well as the phylogenetic relationship of the analysed species using independent contrast analysis ($r_{\text{PIC(Ei Nr MLSP) PIC(GFS.Nr MLSP)}}$ >0, $p < 0.05$, n = 251 and $r_{\text{PIC(Nr Ei MLSP) PIC(GFS.Ei MLSP)}}$ >0, $p < 0.05$, n = 1144 respectively). Only GO terms significantly enriched after B-H multiple testing correction are shown in the figures.

In order to discern the associations of Nr and Ei with GFS, we carried out a partial correlation analysis to correct for the association between Nr and Ei as well as the known dependence of both variables with maximum lifespan (MLSP) (Dunbar and Shultz 2007b; Gonzalez-Lagos et al. 2010). The number of gene families with strong associations was reduced for both Nr and Ei ($r > 0$ and $p < 0.05$; n= 132 and 502, respectively). However, the strongest reduction in the set of gene families was observed for *Ei* suggesting that a significant proportion of the association between Ei and GFS is explained by the association between Nr and GFS but not the other way around. Moreover, while most (n = 13) enriched functional GO terms for Nr remain enriched when correcting for the variance explained by Ei; this is not the case for the functional categories enriched when examining Ei. In fact, after accounting for the variance explained by Nr, only six GO terms are found to be significantly enriched among the gene families with the strongest associations with Ei. Moreover, the strength of the enrichment is lower and involves a different set of GO terms to those found enriched before correcting for Nr (Figure 3.2). This is consistent with Nr having a stronger association with GFS compared to Ei with most observed covariance with GFS being explained as a by-product of the by the correlation between Nr and both Ei and GFS.

Both morphological traits as well as gene family size have a phylogenetic component with most closely related species being more likely to have a higher similarity in their morphological traits as well as in their gene sets. Thus, in order to eliminate the effect of phylogenetic relatedness from the associations between Nr and GFS we carried out a correction of independent contrasts on both the gene family sizes and the morphological traits examined. Accounting for the phylogenetic signal uncovered a stronger association between Nr and GFS: We found that while the number of gene families with significant associations with Ei increases just slightly (n = 251), there was a huge increase in the number of gene families with a strong association with Nr (n = 1144). Most GO term enrichment related to Ei is lost after the removal of the effect of Nr and MLSP and the phylogenetic signal, with only ATP catabolic process and RNA splicing remaining significantly enriched (Figure 3.2).

In contrast, a total of 11 GO terms were found to be enriched among the gene families whose size was most significantly correlated with Nr after correcting for Ei and MLSP as well as removing the phylogenetic signal, with inflammatory response, chemotaxis and cell-cell signalling being among those categories consistently found to be enriched. Interestingly, a number of developmental and cell proliferation GO terms were also found to be significantly enriched in this set of gene families (Figure 3.2).

Although the structure of the layers in the neocortex is stablished during early development (Greig et al. 2013), the neocortex show differing levels of complexity and keeps experiencing growth in childhood and adolescence, reaching a peak in thickness on average around 13 years of age (Shaw et al. 2008). If the association between gene family size and neocorticalization respond to functional demands imposed by the development of a large neocortex, we should expect the genes that compose them to have a particularly high level of activity before this cortical thickness peak is reached. To assess this we made use of neocortex derived expression data BrainSpan Atlas of the Developing Human Brain (J. A. Miller et al. 2014)(see Methods). We found that gene members of this set of families showed higher expression levels during human development prior to the neocortex reaching maximum thickness (before 13 years old) compared to later stages, reflecting a potential involvement of some of these genes in the development of the neocortex (Figure 3.3).

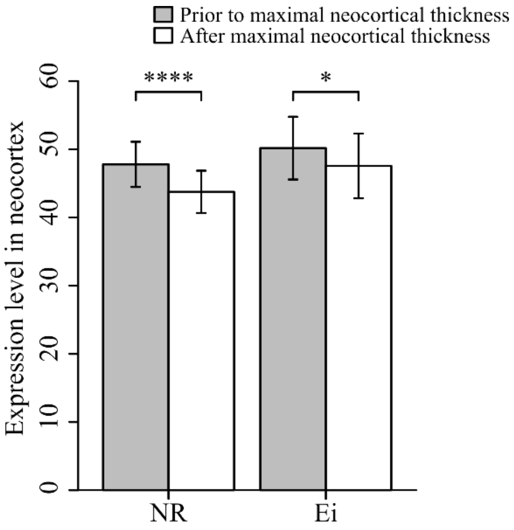

Figure 3.3. Comparison between the cortical expression levels prior and after maximal neocortex thickness is reached of genes within the families more significantly associated with neocortex ratio after correcting for con founding variables. Bars indicate the mean expression for the Neocortex and Encephalization-associated gene families before and after the peak cortical thickness occur, while error bars denote standard error. Wilcoxon signed rank test p values were Bonferroni adjusted for multiple testing comparison (p-value$_{Nr}$ = 9.508343x10^{-9}, p-value$_{Ei}$ = 0.02667157).

3.5 Discussion

The finding that GFS has a strong association with Nr highlights the importance of neocorticalization in the already uncovered relationship between Ei and GFS. The association between Ei and GFS is largely explained by the association between Nr and GFS. The relationship between Nr and GFS is probably due to a shared evolutionary response to specific functional requirements which are also responsible for Ei.

The neocortex is held up as the seat of the highest brain functions, including self-control, consciousness, and thinking. A key aspect of the neocortex is that it is divided into functionally discrete regions. The structure, size and occurrence of these regions is not universal across mammalian species, with variation being related to species' specific functional demands (Krubitzer and Huffman 2000). The process by which the neocortex is subdivided into functional fields is called arealization, and has its basis in a combination of epigenetic and genetic mechanisms (Alfano and Studer 2013; Dehay et al. 1996; D. D. O'Leary et al. 2007). It has further been suggested that in large-brained species "new" areas arise to take on new functions; thus the demands for the molecular mechanisms of arealization might increase as brains get larger.

Gene family size in particular should be related to gene duplication events which result in new genes to perform more diverse functions. Likewise, increase in neocortex size is linked to a proliferation of new cortical areas to perform new functions (Changizi 2001; Kaas et al. 2013). (The duplication of association areas is probably achieved pretty easily, as even within humans duplications arise (Sereno and Huang 2006)). How might GFS and NR then be related? One hints at the molecular basis of neocorticalization comes from the signalling molecule FGF8, whose ectopic expression can lead to duplication of the primary somatosensory area (S1) to create of a new neocortical area (Fukuchi-Shimogori and Grove 2001).

The analysis of GO category function annotation unveiled an association between Nr and gene families. In line with increased Nr, there is an excess of gene families which are significantly enriched in cell-cell signalling, chemotaxis and immune response functional annotations. Both cell-cell signalling and chemotaxis are categories of processes which have important functions throughout the nervous system, and thus the association seems to reflect the increasing demands on this system due to neocorticalization. For example, cell-cell signalling encompasses functions including synaptic signalling and neurotransmission. The enlarged association areas, in the human in particular, may be enriched in the number and distribution of synapses, as indicated by increased density of dendritic spines and more elaborate dendritic branching patterns; for example neocortically-enlarged humans have evolved specific paralogs of the spine maturation promotor SRGAP2 (Charrier et al. 2012). Also, cell-cell signalling by agents such as the chief mammalian inhibitory neurotransmitter GABA. GABAergic neurons comprise one of the two major classes of neurons in the mammalian neocortex, have a role in cortical plasticity, and increase in number and complexity in human evolution (E. G. Jones 1993; Letinic et al. 2002). As we have previously suggested, the enrichment of immune response functional annotations may be due to the gradual integration of an immune signalling system into the mammalian nervous system, to meet the demands of an increasingly large brain (Castillo-Morales et al. 2014). Also, because Nr is positively related to group size, perhaps it is increased exposure to pathogens coincident with increased exposure to conspecifics which has increased demands on the immune system (Pasquaretta et al. 2014).

As seen in Figure 3.2 there is an excess of gene families with no functional category annotations. This may reflect the fact that gene families which are specifically involved in driving the evolution of a larger brain and/or neocortex will be missed from characterisations in rodent models. By using a comparative approach it is possible to uncover sets of genes which may play an important role in the development of key structures such as the neocortex.

ACKNOWLEDGEMENTS

This study was supported by a PhD CONACYT scholarship for ACM and JMS, a Royal Society Dorothy Hodgkin Research Fellowship (DH071902), Royal Society research grant (RG0870644) and a Royal Society research grant for fellows (RG080272) to AUO and a University of Lincoln PP grant to HG.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and design of the study. ACM and JMS carried out analyses presented. ACM, AdS, AOU and HG wrote the manuscript with contributions from all authors.

Supplementary Table S3.1

Species	Common Name	Ei	Non neocortex brain Volume cm3	Neocortex Volume cm3	Nr	MLSP
Ailuropoda melanoleuca	Giant Panda	-2.014	211.80935	136.43571	1.81	36.8
Callithrix jacchus	Marmoset	-1.627	7.241	4.371	1.52	16.5
Canis familiaris	Dog (poodle)	-1.699	458.273	177.753	0.63	24
Cavia porcellus	Guinea Pig	-2.948	4.671815	1.5798	0.51	12
Echinops telfairi	Lesser Hedgehog Tenrec	-3.274	0.566	0.0515	0.1	19
Erinaceus europaeus	Hedgehog	-2.863	3.05	0.522	0.21	11.7
Gorilla gorilla	Gorilla	-1.415	470.359	341.444	2.65	55.4
Homo sapiens	Human	0.152	1251.847	1006.525	4.1	122.5
Loxodonta africana	Elephant	-1.082	3886.7	2460.1	1.72	65
Macaca mulatta	Macaque	-1.192	87.896	63.482	2.6	40
Macropus eugenii	Wallaby	-2.207	11.6637	4.3987	0.61	15.1
Microcebus murinus	Mouse Lemur	-1.985	1.68	0.74	0.79	18.2
Mus musculus	Mouse (C57BL/6J)	-2.832	0.48	0.12	0.32	4
Mustela putorius furo	European Polecat	-2.548	8.8996	4.147	0.87	11.1
Ornithorhynchus anatinus	Platypus	-2.219	8.57145	4.09928	0.92	22.6
Ovis aries	Sheep	-1.961	100.332	53.793	1.16	22.8
Pan troglodytes	Chimpanzee	-0.948	382.103	291.592	3.22	59.4
Papio anubis	Olive baboon	-1.178	190.957	140.142	2.76	37.5
Pongo abelii	Orangutan	-0.892	304.2	219.8	2.6	59
Procavia capensis	Hyrax	-2.255	12.68	5.54	0.78	14.8
Pteropus vampyrus	Megabat	-2.204	8.89	3.61	0.68	20.9
Rattus norvegicus	Rat	-2.861	1.69	0.58	0.52	5
Sarcophilus harrisii	Tasmanian devil	-2.792	15.1517	3.7334	0.33	13
Sorex araneus	Shrew	-2.832	0.188	0.0264	0.16	3.2
Sus scrofa	Pig	-2.468	106.660	54.3913	1.04	27
Tarsius syrichta	Tarsier	-1.795	3.393	1.768	1.09	16
Tursiops truncatus	Dolphin	-0.321	1376.976	1088.615	3.78	51.6
Vicugna pacos	Alpaca	-1.688	181.467	101.81	1.28	25.8

4. Post-mitotic cell longevity-associated genes: a transcriptional signature of post-mitotic maintenance in neural and non-neural tissues

Atahualpa Castillo-Morales[1,2], Jimena Monzón-Sandoval[1,2], Araxi O. Urrutia[2*] and Humberto Gutiérrez[1*].

[1] School of Life Sciences, University of Lincoln, Lincoln LN6 7TS, UK

[2] Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

ACM: Acm39@bath.ac.uk

JMS: Jms52@bath.ac.uk

AOU: a.urrutia@bath.ac.uk

HG: hgutierrez@lincoln.ac.uk

* To whom correspondence should be addressed.

4.1 Abstract

Different cell types have different post-mitotic maintenance requirements. Nerve cells, however, are unique in this respect as they need to survive and preserve their functional complexity for the entire lifetime of the organism. These differences across different tissues could in principle arise from the differential engagement of a general molecular repertoire involved in general maintenance mechanisms. However, whether the onset of certain neurodegenerative conditions is associated to an overall failure at any level of these general supporting mechanisms is not known. By comparing whole genome transcriptome data derived from Alzheimer's and Parkinson's disease patients we found that genes abnormally down regulated in these two conditions are significantly enriched in genes whose expression levels are closely associated with increased post-mitotic cellular longevity (PMCL) across a variety of human tissues ranging in longevity from 120 days to over 70 years. PMCL-associated genes are enriched in specific biological processes and transcription factors targets compared to randomly selected gene samples, and, in addition to being down regulated in the cerebral cortex and substantia nigra of Alzheimer's and Parkinson's disease patients, these genes are also down regulated in Hutchinson-Gilford progeria-derived fibroblasts. We demonstrate that the observed down regulation of PMCL-associated genes in these degenerative conditions is specifically linked to their underlying association with cellular longevity. Moreover, we found that sexually dimorphic brain expression of PMCL-associated genes reflects sexual differences in lifespan in humans and macaques, indicating a link between differential demands in neuronal maintenance between males and females and level of engagement of PMCL-associated genes. Taken together our results suggest that PMCL-associated genes are part of a generalized machinery of post-mitotic maintenance and functional stability in both neural and non-neural cells, that becomes compromised in two specific neurodegenerative conditions and support the notion of a common molecular repertoire differentially engaged in different cell types with different survival requirements.

Keywords: post-mitotic cell maintenance, cell longevity, neuronal survival, neurodegeneration, transcriptomics.

4.2 Background

In multicellular organisms most cells have a shorter lifespan than the organism and are continuously replaced. However not all cell types are replaced at similar rates. Differential demands in turnover across different cell types are necessarily matched by corresponding differences in post-mitotic maintenance. When measured in terms of post-mitotic rate of survival, these differences in requirements range in humans from a few days in skin cells and gut epithelium, to several months or years in the case of bones and muscles (K. L. Spalding et al. 2005).

Although the need for long term survival is common to many cell types, nowhere is post-mitotic cell maintenance more critical than in neurons as mature post-mitotic neurons need to survive and preserve their functional complexity during the entire lifetime of an individual (Magrassi et al. 2013). More importantly, failure at any level in the underlying supporting mechanisms is likely to play a central role in the onset of a wide range of neurodegenerative conditions (Drachman 1997; Fishel et al. 2007; Mattson and Magnus 2006).

Cellular maintenance in neurons and other cell types is likely to be the result of a wide network of interacting molecular mechanisms that act at several levels of the cell's physiology to ensure its structural and functional stability (Lanni et al. 2010; Mattson and Magnus 2006). Identifying these molecular networks is critically important in order to understand both cell survival and its pathological counterpart, cell degeneration.

Current research on neuronal long term survival and maintenance mainly focuses in the study of the signalling events that regulate programmed neuronal death during development or the abnormal reduction in cellular support leading to cell death in models of injury or neurotoxicity (Harrington and Ginty 2013; Jaiswal et al. 2012; Lanni et al. 2010). During development, neurons freely activate cell death pathways to fine-tune the number of neurons that are needed during the precise formation of neural networks. These cell death pathways are remarkably active during early development, and although they become highly restricted as neurons mature, negative regulation of cell death alone is unlikely to account for the characteristic long-term survival potential of nerve cells (Kole et al. 2013).

Specific regulatory events directing post-mitotic survival are known to vary across cell types. Thus, for instance, in developing neurons post-mitotic survival is mostly regulated by neurotrophins and their associated receptors and signalling networks (Cole and Frautschy 2007; Harrington and Ginty 2013; Lanni et al. 2010; Mattson and Magnus 2006). In other cell types, such as plasma B cell, post-mitotic survival is known to respond to the regulatory control of a different array of extrinsic signals including member of the TNF superfamily of ligands, interleukin 4,5 and 6, CXCL12 and others (Benson et al. 2008; Cassese et al. 2003; Mattson and Magnus 2006; O'Connor et al. 2004).

Regardless of the specific regulatory mechanisms engaged by different terminally differentiated post-mitotic cell types, the basic molecular events ensuring appropriate levels of DNA repair, protein stability, protein turnover capacity and organelle integrity are likely to recruit a common repertoire of molecular mechanisms with the only difference being the level of activation of these mechanisms in response to the survival requirements of different cell types and tissues.

Because little is known of the molecular determinants specifically accounting for long term neuronal maintenance, whether the unique long term demands of functional stability in neurons result from the activation of distinct neural-specific maintenance mechanisms or the enhanced activation of an otherwise common molecular repertoire across cell types is not known. Here we asked whether transcriptional alterations in two neurodegenerative conditions, Alzheimer’s and Parkinson’s disease, are associated with genes that display increasing levels of expression in line with variations of post-mitotic survival or cellular longevity (PMCL) demands in other tissues.

In human tissues, our knowledge regarding post-mitotic cell longevity and turnover has been scarce in the past due to the lack of means to accurately measure cell turnover in human subjects. In recent years, however, ^{14}C-based retrospective birth dating has been successfully used to estimate the rate of cell turnover in several human tissues (Bhardwaj et al. 2006; K. L. Spalding et al. 2005). Taking advantage of the availability of these estimates for seven human tissues ranging in longevity from 120 day to over 70 years, here we set out to identify the molecular signature of long term post-mitotic maintenance. To this end, we conducted genome-wide comparisons of human transcriptome data derived from these tissues and screened for genes whose expression patterns are closely associated with changes in post-mitotic cell longevity.

We identified a set of post-mitotic cell longevity (PMCL)- associated genes whose expression levels are robustly and consistently associated with increased cell longevity. Using expression data from six independent sources (Table 4.1), we further found that: 1) genes abnormally down regulated in the cerebral cortex and substantia nigra of Alzheimer's and Parkinson's disease patients respectively, are significantly enriched in genes whose expression levels are closely associated with increased post-mitotic cellular longevity (PMCL) across a variety of human tissues ranging in longevity from 120 days to over 70 years. 2) conversely, genes robustly associated with PMCL across different tissues are also down regulated in brain tissue of Alzheimer's and Parkinson's disease patients respectively; 3) down regulation of PMCL-associated genes in Alzheimer's and Parkinson's disease is specifically linked to their underlying association with cellular longevity; 4) these genes also significantly enriched in specific biological processes and transcription factors targets further supporting the notion that these genes share related biological functions in addition to common regulatory pathways; 5) sexually dimorphic brain expression of PMCL-associated genes reflects sexual differences in lifespan in humans and macaques. Our results demonstrate that two neurodegenerative conditions, AD and PD, are associated to the abnormal expression and dysregulation of PMCL-associated genes and provide the first evidence of a generalised cell longevity pathways in human tissues differentially engaged in different cell types with different survival requirements.

4.3 Results

As the vast majority of genes expressed in the nervous system are also expressed in most tissues, we started by asking whether genes differentially expressed in AD and PD have a particular tendency to show increased levels of expression in line with variations of post-mitotic survival or cellular longevity (PMCL) demands in across different neural and non-neural tissues. To this end, we first identified differentially expressed genes in AD and PD using available microarray expression data derived from 87 samples from Alzheimer's disease patients comprising five different cortical regions and hippocampus and 24 biological samples of substantia nigra from Parkinson's disease patients, with corresponding controls for each condition (n = 74 and 11 respectively, Dataset 1, Table 4.1). Using linear models of microarray analysis (LIMMA) we identified 2935 and 1019 genes displaying significant down regulation in AD and PD respectively, relative to the corresponding control microarrays.

For each gene in the lists of down-regulated genes in each condition, we measured the degree of association between their level of expression across a number of reference tissues and PMCL estimates in these same tissues.

To this end we used available accurate cell longevity estimates obtained from ^{14}C-based retrospective birth dating for cerebellum, cardiac myocyte, pancreatic islet, small intestine (parenchyma), skeletal muscle, adipocyte and leukocytes, ranging in longevity from 120 days to over 70 years (Supplementary table 4.1) (Bergmann et al. 2009; Perl et al. 2010; K. L. Spalding et al. 2005; Whitehouse et al. 1982). Expression data for these tissues was obtained from the Affymetrix GeneChip HG-133U part of the Human U133A/GNF1H Gene Atlas data set, which compiles microarray gene expression data for 79 human tissue samples and cell lines (Dataset 2, Table 4.1; see methods). To obtain an unbiased estimator of the degree of association between the expression level across the above seven tissues and the cellular longevity estimates for the same tissues, we computed a jack-knife correlation for each gene. That is, a sequence of seven pseudovalues is calculated by obtaining the Pearson correlation coefficient while dropping in turn each of the tissues from the analysis. The jack-knife correlation is then defined as the mean of these pseudo-values. This process was repeated for each of the 11 449 genes for which expression data were available in all seven tissues (Dataset 2, table 4.1). We then compared the average jack-knife correlation of either AD or PD-associated genes with the distribution of average jack-knife correlations derived from 1000 000 equally sized random samples of background genes. As shown in figure 4.1, genes found down-regulated in the brain of AD and PD patients show a statistically significant increase in their average correlation with PMCL estimates across all seven reference tissues. In other words, AD and PD-downregulated genes, display a significant tendency towards increased levels of expression in progressively longer living tissues. These results strongly suggest that a substantial proportion of genes displaying abnormal down regulation in two neurodegenerative conditions could also be part of a wider cell-maintenance machinery normally present in human tissues and differentially engaged in different cell types with different survival requirements.

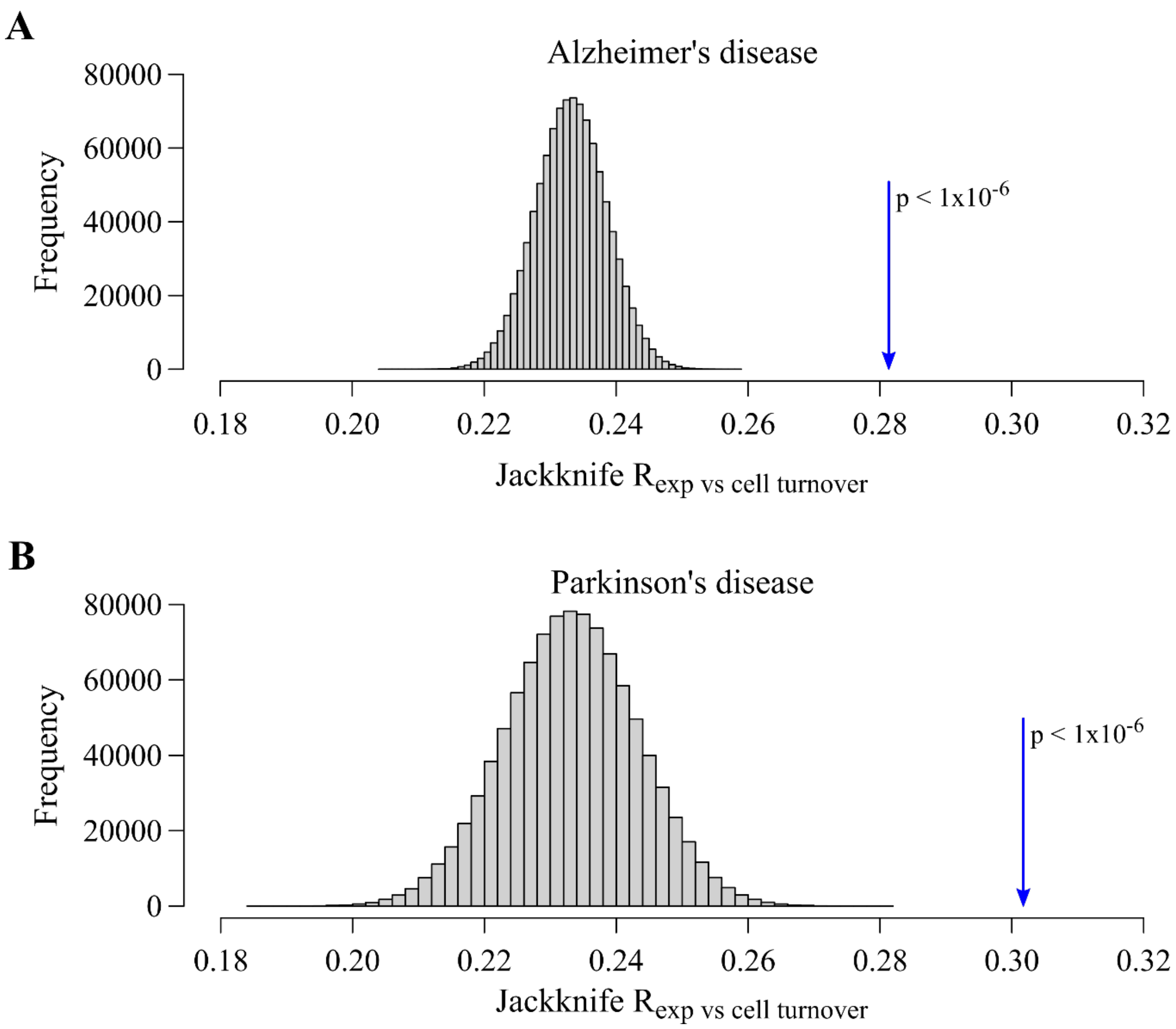

Figure 4.1. Genes down regulated in brain tissue of Alzheimer's and Parkinson's disease patients normally display increased level of expression in longer living tissues. Using available microarray expression data derived from 87 samples from Alzheimer's disease patients comprising five different cortical regions and hippocampus, and 24 biological samples of substantia nigra from Parkinson's disease patients, with corresponding controls for each condition (n = 74 and 11 respectively, Dataset 1, Table 4.1), we identified 2935 and 1019 genes displaying significant down regulation in AD and PD respectively. To obtain an unbiased estimator of the level of association between expression of these genes and post-mitotic cell maintenance in different tissues, we computed the average jack-knife correlation between accurate estimates of post-mitotic cellular longevity in seven reference tissues (ranging in longevity from 120 days to over 70 years) and the normal level of expression of AD and PD-downregulated genes in these same reference tissues. A) Histogram showing the distribution of mean jack-knife correlations of one million independent samples of 2935 random background genes. Blue

arrow indicates the average jack-knife correlation of the actual 2935 AD-associated genes. B) Histogram of the mean Jack-knife correlation of one million independent samples of 1019 random background genes, with blue arrow indicating the mean Jack-knife correlation of the actual PD associated genes. Numerically estimated p values are indicated.

Table 4.1. Sources of gene expression data.

Dataset	Source	Platform	Reference
1	GSE5281(cortex/hippocampus) GSE8397-GPL96 (Substantia nigra)	RNA Microarray RNA Microarray	(Liang et al. 2008) (Moran et al. 2006)
2	BioGPS	RNA Microarray	(Su et al. 2004)
3	GSE13162 (Normal frontal brain) GSE11681–GPL96(Control muscle) GSE42114 (Normal skin)	RNA Microarray RNA Microarray RNA Microarray	(Chen-Plotkin et al. 2008) (Saenz et al. 2008) (Gulati et al. 2013)
4	GSE24487 (Fibroblasts)	RNA Microarray	(Liu et al. 2011)
5	Brawand et al. (supplementary material)	RNA-seq	(Brawand et al. 2011)
6	Brainspan	RNA-seq	(http://brainspan.org ; J. A. Miller et al. 2014Miller)
7	GSE11291	RNA Microarray	(Barger et al. 2008)
8	GSE38012	RNA Microarray	(Mercken et al. 2013)

In order to directly test this hypothesis, we conducted the following reverse analysis: First we independently looked for genes whose pattern of expression were closely and robustly associated with changes in post-mitotic cell longevity across different tissues and then looked at whether these same genes show abnormal expression in the brain of AD and PD patients.

In order to specifically identify gene expression correlates of enhanced cellular maintenance across different tissues, we screened for genes that met two independent criteria: First using our previous unbiased estimates of degree of association between expression and PMCL we only selected genes with an absolute Jack-knife correlation coefficient value greater than 0.8. Second, because complex phenotypes are usually the result of an assembly of molecular and genetic components acting in concert (Hartwell et al. 1999) and genes involved in related biological pathways display correlated expression patterns reflecting their functional association (Eisen et al. 1998; Homouz and Kudlicki 2013), we selected, among genes meeting the first criterion, those that displayed a consistent association with each other across a wider range of tissues and cell types.

By screening for genes meeting the first criterion we initially identified a set of 98 genes with Jack-knife values ranging from 0.801 to 0.972. Interestingly, no genes were identified in the negative tail of the resulting distribution.

To meet the second criterion, from this original set of 98 candidate genes we removed those that failed to display a strong correlation with at least other two genes from the same set when examining their collective pattern of co-expression across a wider set of tissues. To this end, we took advantage of the fact that Dataset 2 also compiles expression data for a total of 28 separate tissues, and extracted the Pearson correlation values of all possible pairs of these genes across all 28 samples and looked for a single connected component or network linked by strong correlations (R > 0.8). This analysis revealed a single cluster of 81 genes leaving 17 isolated genes. Figure 4.2A shows the average expression of our selected set of 81 PMCL-associated genes across the initial seven tissues used in their selection (Table 4.2 and Supplementary table 4.1).

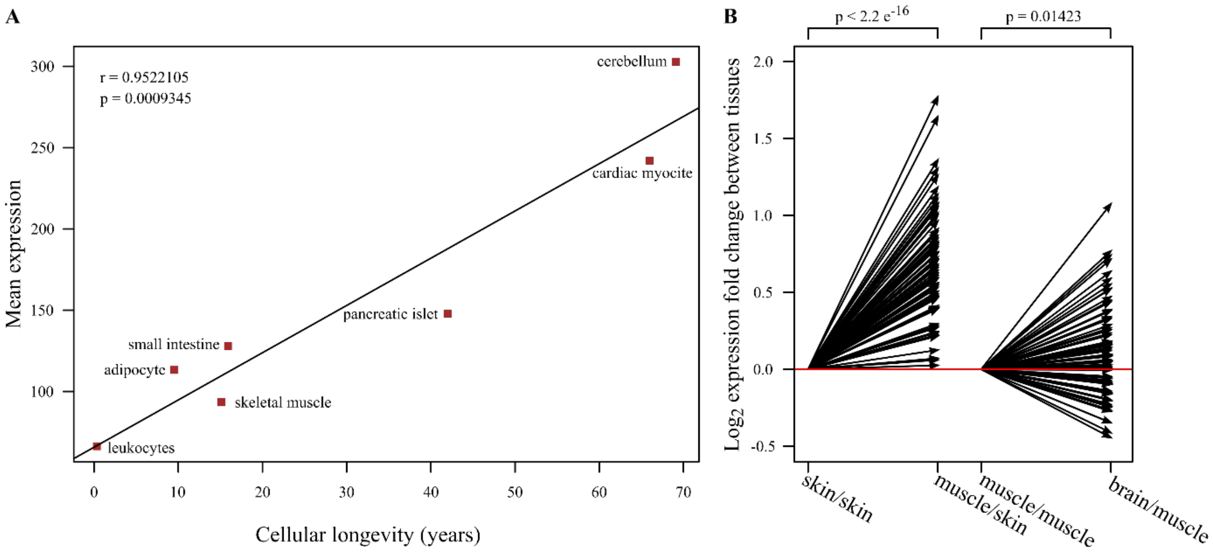

Figure 4.2. Detection of a transcriptional signature of long term post-mitotic maintenance. In order to independently identify a transcriptional signature of post-mitotic cellular longevity (PMCL), we compared whole genome transcriptome data from seven reference tissues and found a set of 81 genes whose expression levels are robustly associated with increased cell longevity (Jack-knife R > 0.8) and highly correlated with each other (see methods). A) Regression plot showing the average normalized expression of 81 PMCL-associated genes as a function of cell longevity in seven separate reference tissues for which carbon dating estimations are available. Expression data used corresponds to dataset 1 and each individual data point represents the average normalized expression of all 81 genes. Pearson correlation coefficient and associated p-value are indicated. B) Fold-change in gene expression for each PMCL-associated gene comparing skin vs muscle expression and muscle vs brain expression respectively. Each arrow represents the direction in $-\log_2$ (fold change) for each particular PMCL-associated gene in each indicated pair of tissues. P-values for the observed average differences in expression of PMCL-associated genes based on paired t-test per comparisons.

Table 4.2. PMCL-associated genes

SYMBOL	GENE NAME	JACKKNIFE R	SYMBOL	GENE NAME	JACKKNIFE R
CCT7	T-complex protein 1 subunit eta	0.968	**MEA1**	male-enhanced antigen 1	0.802
UCHL1	ubiquitin carboxyl-terminal hydrolase isozyme L1	0.934	**PAPSS1**	3'-phosphoadenosine 5'-phosphosulfate synthase 1	0.944
PSMC4	26S protease regulatory subunit 6B	0.933	**TPI1**	triosephosphate isomerase 1	0.811
HSP90AB1	heat shock protein HSP 90-beta	0.933	**COL13A1**	collagen, type XIII, alpha 1	0.855
COPZ1	coatomer subunit zeta-1	0.901	**NES**	nestin	0.825
PFDN2	prefoldin subunit 2	0.875	**MYH10**	myosin-10	0.821
COPS6	COP9 signalosome subunit 6	0.834	**ITM2C**	integral membrane protein 2C	0.883
USP14	ubiquitin specific peptidase 14	0.83	**ATXN2**	ataxin 2	0.881
CDC37	cell division cycle 37	0.813	**ATXN10**	ataxin 10	0.865
TUBB4B	tubulin beta-4B chain	0.818	**TMEM132A**	Heat Shock 70kDa Protein 5 Binding Protein 1	0.836
MZT2B	mitotic spindle organizing protein 2B	0.972	**XRCC6**	X-ray repair cross-complementing protein 6	0.832
TUBGCP2	gamma-tubulin complex component 2	0.879	**TEK**	angiopoietin-1 receptor	0.821
FAM96B	family with sequence similarity 96, member B	0.861	**TRIM28**	tripartite motif containing 28	0.817
CKAP5	cytoskeleton associated protein 5	0.861	**SLC7A5**	large neutral amino acids transporter small subunit 1	0.801
MAPK6	mitogen-activated protein kinase 6	0.857	**ARL3**	ADP-ribosylation factor-like 3	0.916
DCTN3	dynactin 3 (p22)	0.84	**UNC5B**	netrin receptor UNC5B	0.91
NUDC	nuclear migration protein nudC	0.828	**GPI**	glucose-6-phosphate isomerase	0.867
ACTR1A	alpha-centractin	0.815	**SMARCA4**	transcription activator BRG1	0.931
PPP1R7	protein phosphatase 1 regulatory subunit 7	0.804	**SSRP1**	structure specific recognition protein 1	0.84
YWHAE	14-3-3 protein epsilon	0.803	**STIP1**	stress-induced-phosphoprotein 1	0.866
EID1	EP300 interacting inhibitor of differentiation 1	0.818	**SLC3A2**	4F2 cell-surface antigen heavy chain	0.889
PPM1G	protein phosphatase 1G	0.808	**EPM2AIP1**	EPM2A-interacting protein 1	0.833
PAPD7	PAP associated domain containing 7	0.836	**PTS**	6-pyruvoyltetrahydropterin synthase	0.95
CBX5	chromobox homolog 5	0.964	**PFKP**	6-phosphofructokinase type C	0.863
ATP13A2	ATPase type 13A2	0.895	**COX8A**	cytochrome c oxidase subunit 8A	0.834
PNMA2	paraneoplastic antigen Ma2	0.891	**ATP6V1H**	V-type proton ATPase subunit H	0.801
RRAGA	Ras-related GTP binding A	0.814	**CHCHD2**	coiled-coil-helix-coiled-coil-helix domain containing 2	0.809
UBE2Z	ubiquitin-conjugating enzyme E2Z	0.832	**NHP2L1**	NHP2-like protein 1	0.959

YARS	tyrosyl-tRNA synthetase	0.853	**EXOSC10**	exosome component 10	0.819
MAGED1	melanoma antigen family D, 1	0.946	**NHP2**	NHP2 ribonucleoprotein	0.901
NPDC1	neural proliferation, differentiation and control, 1	0.901	**GTF3C4**	general transcription factor 3C polypeptide 4	0.816
AKT3	RAC-gamma serine/threonine-protein kinase	0.869	**VDAC2**	voltage-dependent anion channel 2	0.837
NR2F1	COUP transcription factor 1	0.868	**EEF1E1**	eukaryotic translation elongation factor 1 epsilon 1	0.875
HDGFRP3	Hepatoma-derived growth factor-related protein 3	0.866	**HARS**	histidyl-tRNA synthetase	0.824
RBFOX2	RNA binding protein fox-1 homolog 2	0.837	**SARS**	seryl-tRNA synthetase	0.812
PDXK	pyridoxal kinase	0.825	**NUCKS1**	nuclear casein kinase and cyclin-dependent kinase substrate 1	0.891
FLNB	filamin-B	0.814	**BRD9**	bromodomain containing 9	0.842
FEV	protein FEV	0.813	**GPKOW**	G patch domain and KOW motifs	0.825
IFT46	intraflagellar transport protein 46 homolog	0.81	**SETD5**	SET domain containing 5	0.816
PFN2	profilin 2	0.807	**FAM171A1**	protein FAM171A1	0.809
LARP1	la-related protein 1	0.803			

It is to be noted that, in identifying these 81 PMCL-associated genes, multiple testing corrections where not carried out due to the low statistical power derived from using only seven tissues. However, if our candidate set of genes was the random outcome of a multiple testing artefact we would expect these genes not to display a consistent pattern of association with cellular longevity when examining independent expression data. Accordingly, in a first test of consistency, we used data from occipital lobe and skin which were also present in dataset 2 but not originally used to identify our set of PMCL-associated genes. While we lacked radio carbon-based estimates of skin cell turn-over, other methods place this value between the 39-61 days range in humans (Bergstresser and Taylor 1977; Iizuka 1994; Weinstein et al. 1984). On the other hand, since little or no neuronal turnover has been observed in human brain cortex and up to 50% of cortical cells are neurons (Azevedo et al. 2009; Bhardwaj et al. 2006; K. L. Spalding et al. 2005), we should expect the expression of PMCL-associated genes to reflect that of a long and short living tissue for occipital cortex and skin respectively. We found that 61 out of 81 of our PMCL-associated candidates displayed higher expression in the occipital cortex than in the skin. As all expression data was originally normalized to mean expression levels, this expression bias was significantly stronger than expected by chance (X^2 =

20.75, p = 5.22x10^{-6}), demonstrating that PMCL-associated genes collectively display a higher level of expression in the long living occipital cortex than in the skin.

Using a separate approach, we used additional expression data from a separate microarray dataset containing at least 8 biological replicates of gene expression measures derived from normal human brain, skeletal muscle and skin (Dataset 3, Table 4.1), and compared the differences in average expression level of PMCL-associated genes across these tissues. As shown in Figure 4.2B, the expression of PMCL-associated genes was systematically higher in the brain relative to skeletal muscle (paired t test = 2.23, p = 0.014), while expression in the latter was also systematically higher relative to skin (paired t test = 18.02, p = 2.2x10^{-16}), further supporting a robust association between the level of expression of this set of genes and post-mitotic cellular longevity.

PMCL-associated genes are down regulated in Alzheimer’s and Parkinson’s disease.

Having independently identified a robust set of PMCL-associated genes, we went on to determine if these genes are collectively down-regulated in the brain Alzheimer’s and Parkinson’s disease patients. Using microarray data from Dataset 1, a paired t-test comparison of PMCL-associated genes between each condition and their corresponding controls revealed a statistically significant decrease in the average expression of PMCL-associated genes in each of these conditions relative to their healthy counterparts (Figure 4.3A).

Using a complementary approach, we used the list of previously identified down-regulated genes in each condition to conduct an enrichment analysis aimed at detecting overrepresentation of disease-related down regulated genes among PMCL-associated genes. As shown in Figure 4.3B, the observed proportion of PMCL-associated genes that are also down regulated in the cerebral cortex of Alzheimer’s disease patients is significantly higher than expected by chance (p < 1x10 $^{-6}$). Because Dataset 1, contains at least nine biological replicas for each of six separate cortical regions (entorhinal cortex, superior frontal gyrus, posterior cingulate cortex, visual cortex and medial temporal gyrus) as well as hippocampus plus corresponding healthy controls, we were able to assess down regulation of PMCL associated genes in each region separately, and found that the proportion of PMCL-associated genes down-regulated in AD was significantly higher than random expectations in all regions except the primary visual cortex and superior frontal gyrus. Likewise, the proportion of PMCL-associated genes that were also down-regulated in the substantia nigra of Parkinson’s disease patients was significantly higher than expected by chance ($p = 3x10^{-6}$, Figure 4.3B).

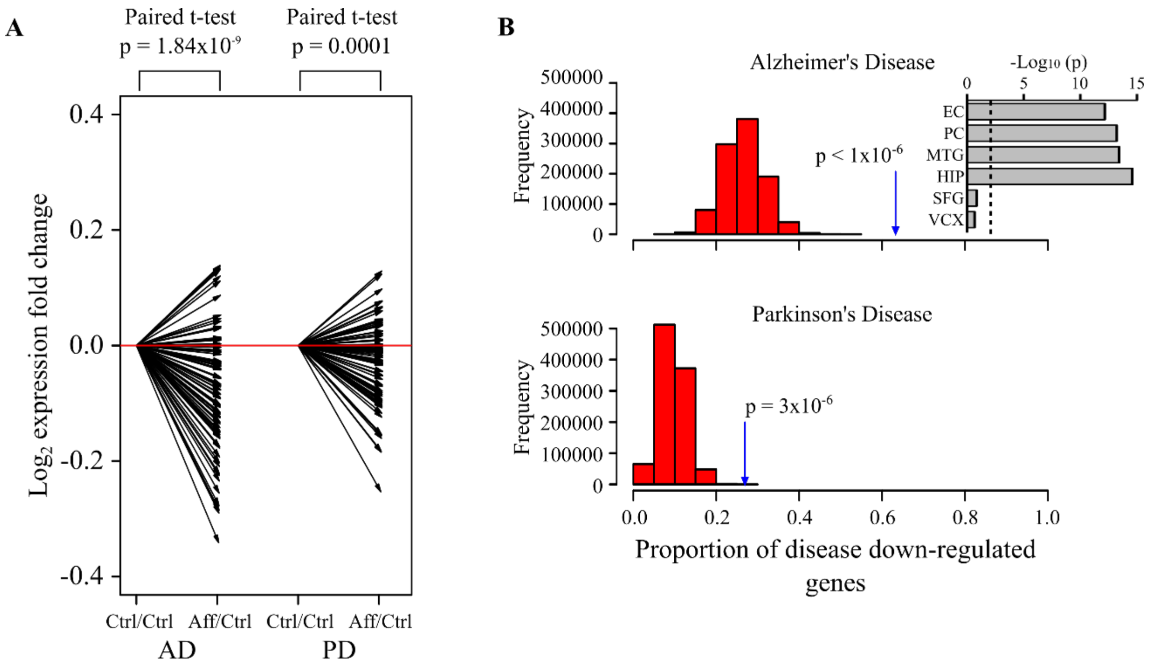

Figure 4.3. PMCL-associated genes are down regulated in Alzheimer’s and Parkinson’s disease. A) fold-change in gene expression for each PMCL-associated gene relative to their control counterpart for each indicated condition. Each arrow represents the direction in –log_2 (fold change) for each particular PMCL-associated gene in each condition. P-values for the observed average differences in expression of PMCL-associated genes, between control and disease samples, were obtained using paired t-test. B) Microarray data from brain cortex (n = 161) and substantia nigra (n = 35) obtained from Alzheimer’s disease (AD) and Parkinson’s disease (PD) patients respectively, was used along with corresponding controls to identify genes significantly down regulated in each condition. The chart shows the distribution of expected proportion of disease-related down-regulated genes for each condition, using 1 000 000 random samples of 81background genes each. Blue arrow indicates the actual proportion and associated probabilities of PMCL-associated genes that are also down-regulated in each indicated condition. Inset: significance of enrichment of down-regulated genes (expressed as –log(P value)) for each separate brain region in AD patients with the dashed line representing the adjusted significance threshold (EC: entorhinal cortex; PC: posterior Cingulate cortex; MTG: medial temporal gyrus; HIP: hippocampus; SFG: superior frontal gyrus and VCX: visual cortex). Note that the proportion of PMCL-associated genes down-regulated in AD

was significantly higher than expected in all regions except the primary visual cortex and superior frontal gyrus.

Taken together, these results demonstrate that, collectively, PMCL-associated genes are significantly down regulated in the cerebral cortex and substantia nigra of Alzheimer's disease and Parkinson's disease patients.

However, if the level of activation of PMCL-associated genes is functionally linked to changes in long term post-mitotic maintenance, we should also expect these genes to be downregulated in degenerative conditions not necessarily related to the nervous system but also involving reduced cell survival or compromised functional stability. To this end we looked at Hutchinson-Gilford progeria syndrome (HGPS), a condition involving a systemic failure of cell maintenance mechanisms liked to normal ageing, such as compromised DNA repair, genome instability and premature senescence (Burtner and Kennedy 2010; Coppede and Migliore 2010; Kudlow et al. 2007; Musich and Zou 2011) and used available microarray expression data derived from 2 biological samples derived from fibroblasts of HGPS patients with corresponding controls for each condition (n = 2, Dataset 4, Table 4.1). As shown in figure 4.4A, a paired t-test comparison between affected and control-derived fibroblasts revealed a statistically significant decrease in the average expression of PMCL-associated genes in affected samples, relative to their healthy counterparts. Using the same complementary approach described above, the proportion of PMCL-associated genes that were also down-regulated in HGPS-derived fibroblasts was also significantly higher than expected by chance ($1.35\text{x}10^{-4}$ respectively, Figure 4.4B)

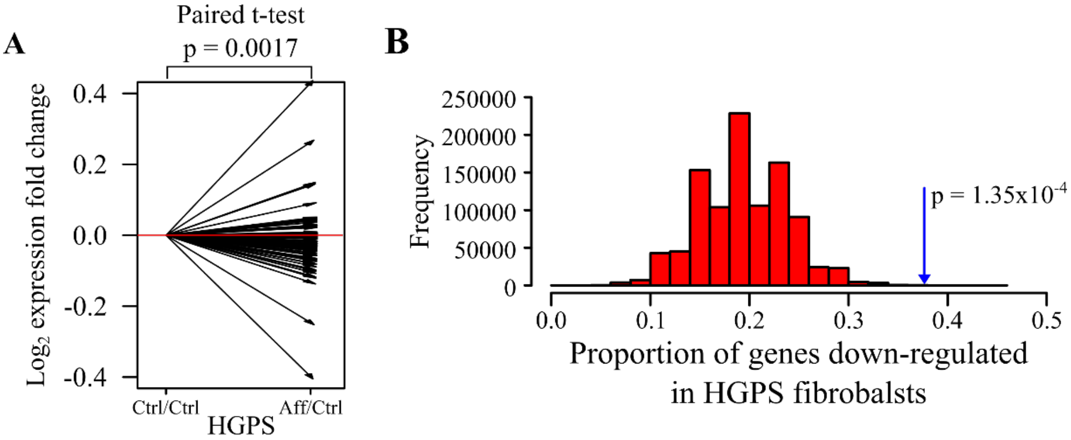

Figure 4.4. PMCL-associated genes are down regulated in Hutchinson-Gilford Progeria syndrome. We used available microarray expression data derived from 2 biological samples derived from fibroblasts of HGPS patients with corresponding controls for each condition (n = 2, Dataset 4, Table 4.1). A) Paired t-test comparison between affected and control-derived fibroblasts revealing a statistically significant decrease in the average expression of PMCL-associated genes in affected samples, relative to their healthy counterparts. B) Microarray data from HGPS-derived fibroblasts was also used along with corresponding controls to identify genes significantly down regulated in this condition. The chart shows the distribution of expected proportion of disease-related down-regulated genes using 1 000 000 random samples of 81background genes. Blue arrow indicates the actual proportion and associated probabilities of PMCL-associated genes down-regulated in HGPS (p= $1.35x10^{-4}$).

PMCL-associated genes display a reduced level of concerted expression in Alzheimer's and Parkinson's disease.

Genes involved in related biological pathways display correlated expression patterns reflecting their functional association (Eisen et al. 1998; Homouz and Kudlicki 2013). Gene co-expression analysis has been widely used to gain insights into the functional organization of transcriptomes across tissues, conditions and species (Obayashi and Kinoshita 2011; Oldham et al. 2006; Oldham et al. 2008; Saris et al. 2009; Torkamani et al. 2010; Usadel et al. 2009; J. Zhang et al. 2012). But apart from revealing functional interactions among groups of genes, gene co-expression could also reveal alterations of the underlying regulatory architecture associated to a global expression profile of a set of genes under particular pathological conditions. Accordingly, we used gene co-expression as an index of regulatory coordination to determine whether PMCL –associated genes display altered levels of correlated expression in AD and PD.

Using expression data derived from substantia nigra of PD and six cortical regions from AD patients and corresponding controls for each condition (Dataset 1) , we obtained the Pearson correlation coefficient between all possible pairs of PMCL-associated genes for each condition and associated controls (see methods) and carried out a paired Wilcoxon test comparison . As shown in figure 4.5A, this analysis reveals a strong and statistically significant reduction in the collective co-expression PMCL-associated genes in brain samples of AD patients relative to the coexpression of the same genes in control samples ($p = 1.369 \times 10^{-34}$). The same analysis carried out with PD data, reveals an equally strong reduction in correlated expression of PMCL-associated genes in the substantia nigra of PD patients relative to control samples (Figure 4.5B; $p = 7.0 \times 10^{-7}$).

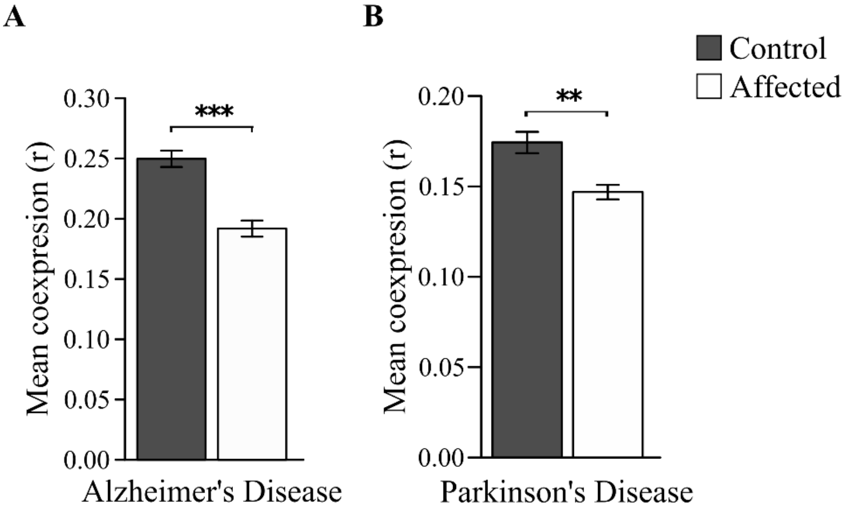

Figure 4.5. PMCL-associated genes display a reduced level of co-expression in in Alzheimer's and Parkinson's disease. Using expression data derived from substantia nigra of PD and six brain regions from AD patients and corresponding controls for each condition (Dataset 1), we obtained the Pearson correlation coefficient between all possible pairs of PMCL-associated genes for each condition and associated controls and carried out a paired Wilcoxon test comparison. A) Chart showing the average correlation coefficient (±S.E.M) between PMCL-associated genes in control and AD samples. B) Chart showing the average correlation coefficient (±S.E.M) between PMCL-associated genes in control and PD samples. Associated p values are indicated.

It is worth noting that the observed reduction in the overall level of co-expression of PMCL-associated genes in PD and AD is not the result of a general reduction in the level of coexpression of the background transcriptome as no significant reduction in coexpression was observed in 10 000 equally-sized random samples of background genes (not shown).

Together, these results demonstrate that PMCL-associated genes display significant reduction in their levels of regulatory coordination in in the brain of PD and AD patients relative to their control counterparts.

Down regulation of PMCL-associated genes in degenerative conditions is specifically linked to their association with cellular longevity.

The above results offer the opportunity to assess whether the observed down regulation of PMCL-associated genes in degenerative conditions is specifically linked to their underlying association with cellular longevity. Because the seven reference tissues used to identify these genes differ in more than one aspect, it is conceivable that alternative selections based on different cellular traits could have led to the exact same results, thereby demonstrating a lack of association between post-mitotic cellular longevity and the down regulation of these genes in degenerative conditions. Because different cellular traits would be typically associated to different rankings or orderings of the reference tissues, one way to assess the effect of potentially different phenotypes in the selection of genes and their down regulation in these conditions, is by looking at the effect of different permutations of these tissues on the resulting gene sets.

To these end, using the same strategy employed to identify PMCL-associated genes, we selected alternative sets of genes derived from each 5 040 possible permutations of the original cellular longevity values. For each permutation, we defined the degree of similarity with the original ordering as the correlation coefficient between the original longevity values and the permuted one. For all permutations above any given similarity value, we measured the proportion of permutations leading to gene sets significantly down regulated in each of the two neurodegenerative conditions examined. As shown in Figure 4.6, the proportion of down regulated gene sets remains close to zero for low minimal similarity values, suddenly increasing as the similarity value approaches one. These results demonstrate that the only permutations leading to the detection of gene sets significantly down regulated in degenerative conditions are those closely aligned with the real post-mitotic cell longevity values. In other words, these results demonstrate that the observed down regulation of PMCL-associated genes in degenerative conditions is specifically linked to their underlying association with cellular longevity.

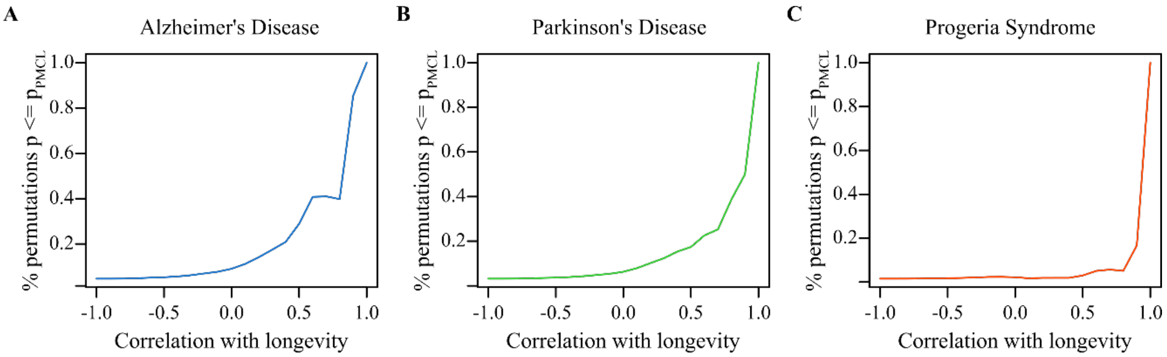

Figure 4.6. Down regulation of PMCL-associated genes in degenerative conditions is specifically linked to their association with cellular longevity. Following the same strategy to identify PMCL-associated genes, we identified alternative sets of genes highly correlated with each of the 5 040 possible permutations of the original seven cell longevity values. For each permutation, we defined the degree of similarity with the original ordering by means of their correlation coefficient. A-C; Each graph shows for all permutations above a given similarity value (x axis), the proportion of permutations leading to gene sets as significantly down regulated in each of the indicated degenerative conditions as the real ordering. Note that the proportion of significantly down regulated gene sets only increases when the minimal similarity between the permuted and the original ordering of longevity values approaches 1. AD: Alzheimer disease; PD: Parkinson’s disease.

PMCL-associated genes are enriched in specific biological processes and transcription factors targets.

If our candidate PMCL-associated genes are functionally related, we would expect them to share common pathways and biological processes. In order to identify possible pathways and biological processes significantly overrepresented among these genes, we conducted a gene ontology (GO) term enrichment analysis. We specifically looked at biological process categories contained in the GO slim subset of terms (http://geneontology.org) and Benjamini-Hochberg multiple testing corrections were carried out against the number of categories tested. Eight biological processes were found overrepresented: cytoskeleton-dependent intracellular transport, tRNA metabolic process, cell cycle, cell morphogenesis, protein folding, cell division, cellular amino acid metabolic process and ribosome biogenesis (Table 4.3).

If PMCL-associated genes are functionally related, we should also expect them to display a higher level of transcriptional coordination with each other in longer living tissues and should therefore be, to a significant extent, under concerted transcriptional regulation. Transcription factors (TF) are key components of regulatory cascades involved in coordinating gene expression. Enrichment of specific TF targets among PMCL-associated genes can provide additional insights into the general regulation of post-mitotic maintenance and functional stability. To this end, we used Transcription factor target annotations obtained from the Molecular Signatures Database (MSigDB v4.0) and found that our set of PMCL-associated genes is significantly enriched in genes with binding sites for HSF, ELK1, EFC (RFX1), USF and USF2 (Table 4.3), in addition to genes containing the SP1 binding motif V.SP1_01 (adj. p = 0.043). Taken together, these results demonstrate that distinct biological processes as well as specific transcription factors targets are statistically overrepresented among genes whose expression patterns are closely associated with changes in post-mitotic cell longevity and that these PMCL-associated genes are specifically down-regulated the brain cortex and hippocampus on the one hand, and substantia nigra, on the other, of AD and PD patients respectively.

4.4 Discussion

Terminally differentiated post-mitotic cells have different turnover and survival requirements. Whether these differences arise from equally different cell maintenance mechanisms engaged by different cell types or the differential activation of an otherwise common molecular repertoire is not known. Nowhere are these supporting mechanisms as critical as in the nervous system where the vast majority of nerve cells cannot be replaced and need to survive as long as the organism, reaching in humans even 100 years or more.

The specific regulatory or signalling events directing long term post-mitotic survival are known to differ across different tissues. However, the basic molecular events ensuring appropriate levels of DNA repair, protein turnover and stability as well as organelle integrity could, at least in principle, potentially recruit a common repertoire of molecular mechanism with the only difference being the level of activation of these same mechanisms in response to the survival requirements of different cell types and tissues. In this study, we have asked whether transcriptional alterations associated with two neurodegenerative conditions are associated with a potentially general cell longevity pathway in human tissues differentially engaged in different cell types with different survival requirements.

Along this lines, we found that genes downregulated in AD and PD also have a significant tendency to show increased levels of expression in line with variations of post-mitotic maintenance demands in other tissues. Following a reverse approach, we conducted a genome wide screening to identify a transcriptional signature of long term post-mitotic maintenance to determine whether these genes are also down regulated in AD and PD.

Because of the inevitable noise in the existing expression data and the limited number of tissues for which accurate data on cellular longevity is available, our ability to identify a hypothetical cell maintenance machinery specifically linked to variations in long term post-mitotic survival is necessarily limited. In spite of this limitation, by comparing genome-wide expression data in seven tissues ranging in cell longevity from 120 days to over 70 years in combination with Jack-knife correlations to rule out spurious effects of strong outliers, we detect at least 81 genes whose levels of expression are robustly correlated with cellular longevity. While a conceptually similar strategy has been previously used to scan for genes associated with increased cancer incidence in several tissues (Silva et al. 2011), using large scale expression data to scan for genes potentially involved in post-mitotic cell longevity has never been attempted before.

Given the low statistical power associated to the use of only seven tissues, our selection of PMCL-associated genes was based on their associated high Jack-knife correlation value rather than significance. In spite of this, we demonstrate that the resulting set of PMCL-associated genes and their specific nature is not the result of a potential multiple testing artefact. Indeed, using additional data from tissues not included originally in the identification of these genes as well as two additional independent expression databases, we found that PMCL-associated genes systematically display higher expression in progressively longer-living tissues. This result is, by no mean consistent with the expected random outcome of a multiple testing artefact.

Having identified a robust set of PMCL-associated genes, we found that these genes are significantly down regulated in the cerebral cortex and substantia nigra of Alzheimer’s disease and Parkinson’s disease patients respectively. Interestingly, PMCL associated genes showed no significant enrichment of down regulated genes in the visual cortex of Alzheimer’s disease patients (Figure 4.4B). This result is particularly significant given the fact that the visual cortex is known to show the least changes and is relatively spared from Alzheimer’s disease pathologies (Liang et al. 2007; Liang et al. 2008). In addition to the observed collective down regulation of PMCL-associated genes in AD and PD, an additional analysis of their level of concerted regulation (or co-expression) revealed a statistically significant reduction in average co-expression in these degenerative conditions relative to the average correlation of these genes in normal controls, further revealing an overall disruption of the concerted regulation of these genes in neurodegenerative pathologies.

We further demonstrate that the observed down regulation of our set of genes in each of these conditions is specifically related to their underlying association with cellular longevity. We did this by following the exact same procedure we followed to identify our PMCL-associated genes and obtained all possible alternative gene sets resulting from all possible permutations of cell longevity values in the original seven tissues. We showed that only those permutations that match the original PMCL-based ranking of the reference tissues lead to gene sets that are also down regulated in these conditions.

These results demonstrate that the down regulation of PMCL-associated genes in three separate degenerative conditions is specifically linked to the PMCL-associated ranking of the reference tissues originally used to identify these genes. In other words the collective down regulation of these genes in degenerative conditions, is the result their specific underlying association with post-mitotic cell longevity.

Transcriptome analyses allowed us to uncover a set of genes with a distinct pattern of expression cell types of increasing post-mitotic maintenance in the human body. These genes are enriched among deregulated genes in disease states. Importantly, our results are consistent with the existence of a generalised common molecular mechanism controlling basal post-mitotic maintenance. Because nerve cells survive as long as the organism, any consistent and systematic differences in overall life expectancy between individuals are likely to be accompanied by differential post-mitotic survival demands in nerve tissue. In this respect, the known sex dimorphism in life expectancy humans, with females living 6% longer lives than males (Clutton-Brock and Isvaran 2007; Kinsella 1998; Vina et al. 2005; Vina and Borras 2010) is likely to translate into corresponding differences in long-term neuronal maintenance. While the cellular and genetic mechanisms underlying sexual lifespan dimorphism are still poorly understood, proposed mechanism include differences in telomere dynamics (Barrett and Richardson 2011; Jemielity et al. 2007), differential response to oxidative stress (Ballard et al. 2007) and asymmetric inheritance of sex chromosomes and mitochondria (Camus et al. 2012; Gemmell et al. 2004). In order to test whether brain expression of PMCL associated genes reflect sexual dimorphism in life expectancy, we compared their expression using existing RNA-seq data from both male and female human individuals. Along the same lines we also investigated gene expression profiles for male and female macaques, as this species displays a more pronounced sexual lifespan dimorphism than humans; with females living on average 72% more than males in the wild (Clutton-Brock and Isvaran 2007), a difference potentially entailing substantially higher neuronal survival and functional stability demands in female macaques compared to males. RNA-seq data for both male and female individuals from these two species was obtained from the study performed by Brawland et al in 2011 (Brawand et al. 2011) (Dataset 5, Table 4.1). A paired t-test comparison of the expression levels of all PMCL-associated genes between males and females in humans revealed a statistically significant increase in the average expression difference of PMCL-associated genes in human females relative to males (Figure 4.7A). Interestingly, the same comparison of the expression levels of all PMCL-associated genes between females and males in macaques revealed a much more pronounced expression of PMCL-associated genes in female relative to males (Figure 4.7A). It is worth noting that dimorphic expression of PMCL-associated genes was much more pronounced in macaques where differences in lifespan between females and males in the wild are also much greater than in humans. It should be mentioned, however, that the pronounced dimorphic lifespan in macaques has been observed predominantly in wild populations and some studies in captivity have actually reported an inverse relationship (Mattison et al. 2012). This suggests that captivity conditions could have a detrimental effect in survival specifically affecting females (or a beneficial effect, specifically affecting males).

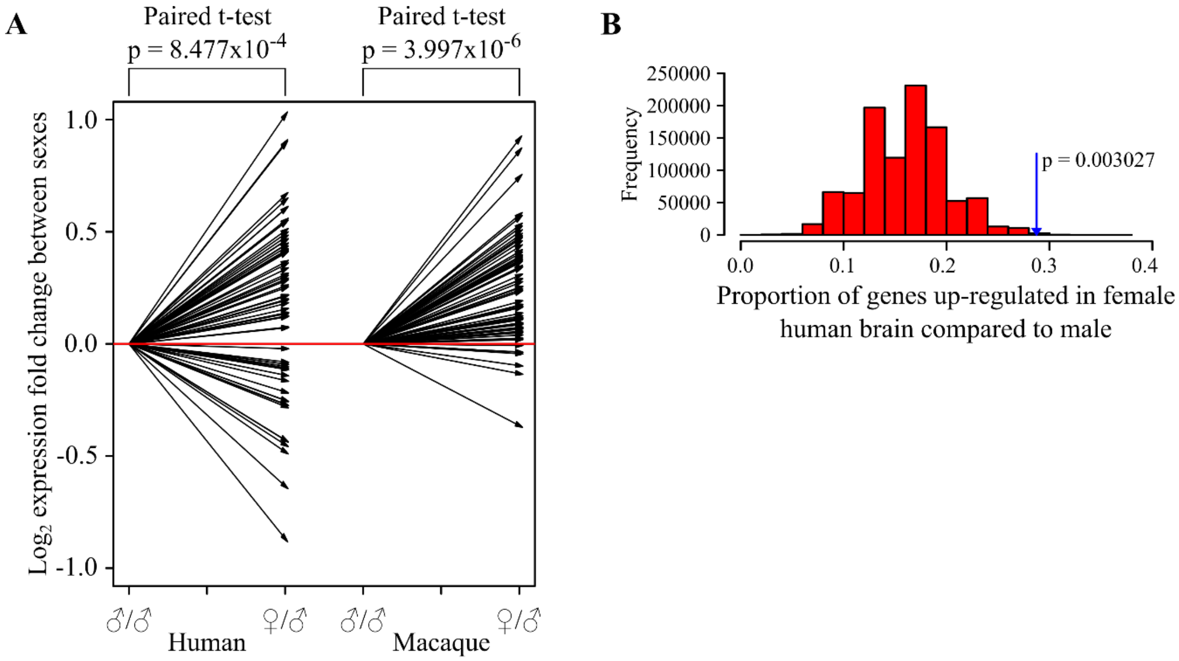

Figure 4.7. Sexual expression dimorphism of PMCL-associated genes in the brain reflects sexual differences in longevity in human and macaque. A) fold-change in gene expression for each PMCL-associated gene comparing males and females in human and macaque. Each arrow goes from –log$_2$ (♂/♂ expression) to log$_2$ (♀/♂ expression) for each particular PMCL-associated gene. p-values for the observed average differences in brain expression of PMCL-associated genes between the two sexes for each species were obtained using paired t-tests. B) Distribution of the expected proportion of genes up-regulated in females relative to males using 1 000 000 random samples of 81 genes. A linear model was used to detect genes significantly up-regulated in females using RNA-seq data from 30 different brain samples of 40 year old human subjects obtained from the BrainSpan dataset. The blue arrow indicates the actual proportion of PMCL-associated genes up-regulated in females relative to males.

Using a complementary approach based on a separate source of RNA-seq expression data derived from 15 brain regions from a 40 year old man and woman obtained from the Brainspan database, Dataset 6 (Table 4.1) we extracted the list of genes significantly up regulated in the female transcriptome (see methods). We then used this list to conduct an enrichment analysis aimed at detecting over-representation of female up-regulated genes among our set of PMCL-associated genes. The expected proportion was numerically calculated based on 1 000 000 equally-sized random gene samples drawn from the overall gene population. As shown in Figure 4.7B, the observed proportion of PMCL-associated genes up regulated in females relative to males is significantly higher than expected by chance (p = 0.003). These results demonstrate that in the female nervous system, where cell survival-related requirements are likely to be higher than in males, PMCL-associated genes are significantly up-regulated relative to their male counterpart.

In all, our results demonstrate that in the female nervous system, PMCL-associated genes are significantly up-regulated relative to their male counterpart possibly reflecting corresponding sex-related differences in long term neuronal maintenance requirements.

Both the down-regulation of PMCL-associated genes in Alzheimer's disease and Parkinson's disease as well as Progeria, together with their up-regulation in the female brain of both humans and macaques, suggest that these genes could constitute a potential signature of either enhanced or compromised functional stability both in neurons as well as other cell types.

Using gene ontology functional annotations and enrichment analysis we found that biological processes such as cytoskeletal-dependent transport, cell morphogenesis and protein folding are statistically overrepresented among our PMCL-associated even after correcting for multiple testing against all 69 functional categories tested. Crucially these genes are also enriched in targets of specific transcription factor further supporting the notion of these genes being part of a common pathway involved in long term cell survival and functional stability. Similar results were obtained when using a standard GO enrichment analysis tool such as WebGestalt.

Our screening captured genes involved in resistance against protein misfolding including prefoldins (*PFDN2*), ubiquitin esterases (*UCHL1*), chaperonins (*CCT7*), chaperons (*HSP90AB1*) and associated adaptor proteins (*STIP1*) as well as proteosomal subunits (*PSMC4*). The fact that these genes are increasingly up-regulated in long living tissues points towards the sustained activation of the unfolded protein response (UPR) and/or the proteasome pathway as a central component of the long-term survival machinery of long living tissues such as the nervous system. Along these lines, both UPR and the ubiquitin/proteasome system (UPS) have been proposed to be important players in the aging process in different species (Durieux et al. 2011; Kimata et al. 2006; Kruegel et al. 2011; Min et al. 2008; Morley and Morimoto 2004; Perez et al. 2009). Oxidative stress can cause protein misfolding and improperly folded proteins that are either retained within the lumen of the endoplasmic reticulum (ER) in complex with molecular chaperones or degraded through the 26S proteasome or through autophagy. Accumulation of misfolded proteins is also known to cause ER stress, which in turn can exacerbate oxidative stress (Gregersen and Bross 2010; Malhotra and Kaufman 2007). *HSP90* is known to modulate the unfolded protein response (UPR) (Marcu et al. 2002) and targeting *HSP90* can destabilise UPR induced cell death (Barrott and Haystead 2013; Davenport et al. 2008; Jackson 2013). Interestingly, mutants of *HSP90* are known to affect lifespan in *C. elegans*, *D. melanogaster* and *S. cerevisiae* (Chen and Wagner 2012; Morley and Morimoto 2004; Sakurai and Ota 2011).

Organismal ageing is a process that involves a progressive decrease in the capacity to adequately maintain tissue homeostasis (Bernardes de Jesus and Blasco 2012; Burton 2009; de Magalhaes and Faragher 2008; de Magalhaes et al. 2012; Dutta et al. 2012; Manayi et al. 2014; Terman et al. 2010). Being such a complex process, ageing involves a large number of changes at various physiological levels and could, at least in principle also involve the gradual breakdown in post-mitotic cell maintenance. With this in mind we looked into any potential overlaps between post-mitotic cell longevity genes and genes known to be associated with ageing. To this end, we examined the GenAge database of genes related to ageing (Tacutu et al. 2013), and after comparing with PMCL-associated genes a number of functional links between both sets of genes were apparent. For example, GenAge lists a number of genes, including E2F1, p53, CDKN1A, PPP1CA, known to be regulated by the transcriptional intermediary factor TRIM28 which we found among our PMCL-associated genes and is involved in development and DNA repair. Conversely, GenAge lists transcription factor SP1 and we found a significant overrepresentation of genes with a particular SP1 binding site among our gene set. We also identified COX8A, a cytochrome c oxidase, while GenAge includes cytochrome c oxidase (MT-CO1) and COXPD6, a pro-apoptotic factor involved in its release from the mitochondria. While GenAge contains the gene encoding for the catalytic subunit of the protein phosphatase 1 (PPP1CA) and several of its regulators (BRCA1, BCL2 and PTK2; all of them members of the PPP1R family) we identify another regulator, PPP1R7 among our PMCL- associated genes. Furthermore, both gene sets contain genes involved in the ubiquitin mediated proteolysis pathway (e.g. UCHL1, UBE2I, UBB, and USP14). We also identified HSP90AB1 and its co-chaperones CDC37 and STIP1, whereas GenAge points towards chaperones HSP90AA1, HSPD1, HSPA1A, HSPA1B, HSPA8, and STUB1 (Apweiler et al. 2014; Stelzer et al. 2011).

Prompted by the potential association between PMCL-associated genes and aging-related processes we looked at differentially expressed genes in the brain of mice subjected to caloric restriction (CR), an experimental dietary regime known to slow down ageing-related changes in many animal models (Barger et al. 2008). Using available expression data in mice we found a statistically significant overrepresentation of PMCL-associated genes among CR up-regulated genes (Dataset 7; p = 0.041). An even more pronounced effect was found when using human data (Mercken et al. 2013) derived from skeletal muscle of human individuals subjected to CR (Dataset 8; p = 0.0082).

Differences among neuronal populations in the production and/or clearance of abnormal proteins are thought to be key determinants of age-related neuronal vulnerability in Alzheimer's disease, Parkinson's disease (PD) and Huntington's disease (HD) (Lam et al. 2000; Mattson and Magnus 2006; McNaught et al. 2001). In this regard, several of the adverse consequences of ageing and neurodegenerative disorders on neuronal function, morphology and survival, as well as behavioural alteration, can be mimicked by pharmacological inhibition of proteasomes (Romero-Granados et al. 2011; Sullivan et al. 2004). Interestingly, loss of function of *UCH-L1* in mice is known to cause gracile axonal dystrophy (gad) phenotype resulting in sensory–motor ataxia (Saigoh et al. 1999). Importantly, these mutants also showed axonal degeneration and formation of spheroid bodies in nerve terminals and an accumulation of amyloid β-protein (Aβ) and ubiquitin-positive deposits, suggesting that *UCH-L1* is involved in neurodegenerative disorders. On the other hand, in amyloid pathogenesis, overexpression of Hsp70 and Hsp90 has been shown to decrease Aβ aggregation (C. G. Evans et al. 2006), reduce Aβ-mediated neuronal toxicity, and appears to enhance the chaperone-mediated clearance of amyloid precursor protein (APP) and its amyloidgenic Aβ derivatives (P. Kumar et al. 2007). Indeed, modulation of *HSP90* has been proposed as a therapeutic tool against Alzheimer’s disease (Zhao et al. 2012).

4.5 Conclusions

Taken together, our results show that genes abnormally down regulated in AD/PD are significantly enriched in genes whose expression levels are closely associated with increased post-mitotic cellular longevity across a variety of human tissues. In this regard, our results support the notion of a common molecular repertoire of cellular maintenance mechanisms shared by all terminally differentiated post-mitotic cells and show that these same mechanisms are differentially engaged in different cell types with different survival requirements. In addition, the observed down regulation of these genes in models of neuronal degeneration and reduced lifespan and/or compromised functional stability, identify PMCL-associated genes as robust molecular markers of either compromised or enhanced cell survival both in neural and non-neural tissues. This is the first genome-wide analysis suggesting the existence of generalised cell longevity pathways in human tissues that becomes compromised in neurodegenerative conditions. Identifying the underlying maintenance mechanisms that allow long living tissues, such as nerve cells, to preserve their functional and structural integrity for the entire lifetime of the organism is essential to understand both aging and neurodegeneration in addition to the unique cell survival capabilities of the human nervous system.

4.6 Materials and Methods

4.6.1 Cellular longevity estimates.

Cellular longevity estimates based on quantification of ^{14}C in genomic DNA from 7 somatic tissues (adipocyte, cardiac myocytes, cerebellum, pancreatic islet, skeletal muscle, leukocytes and small intestine) were obtained from Spalding et al (2005), and associated literature sources (Supplementary table 4.1).

4.6.2 Human tissues gene expression data.

GCRMA normalized cell type specific patterns of mRNA expression for seven tissues for which cell longevity data is available, were extracted from the Affymetrix GeneChip HG-133U part of the Human U133A/GNF1H Gene Atlas dataset, which comprises transcriptome data for 79 human tissue samples and cell lines (Dataset 2). While occipital cortex expression data was also available, only data from cerebellum was initially included in order to avoid unnecessary overrepresentation of nervous tissue in our initial tissue samples. Probe sets were mapped to Ensembl gene IDs via probe set annotations downloaded from the Ensembl's Biomart database (release71). Where more than one probe mapped to a single gene ID, expression measurements were averaged. Any probe matching more than one gene ID was eliminated from the analysis. Probes with zero variance in expression levels across tissues were excluded together with non-protein coding genes. This reduced our background population of genes to a total of 11 449 genes. In order to correct for variations in total signal across tissues, individual expression values were renormalized against the total expression signal per tissue. All the expression data obtained from the sources listed in Table 1 was processed in a similar way. Briefly, expression data from brain, muscle and skin from Gene Expression Omnibus (GEO) (GSE13162, GSE11681 and GSE42114 respectively, Dataset 3, Table 4.1) were selected due to the similarity of the microarray platforms and the availability of several normal replicas allowing a reliable assessment of co-expression. As before, we summarized to Ensembl gene ID all RMA-normalized expression values which were then normalized by the total intensity per sample. RNA-seq expression data (RPKM-normalized and summarized to gene ID as described above) was downloaded from Brainspan database ((J. A. Miller et al. 2014) http://www.brainspan.org/, Dataset 6, Table 4.1). Data for all 12 cortical areas present in this database across 20 different ages were extracted from this source for subsequent analyses. We further normalized individual expression values within samples against the total level of gene expression in each sample. Where more than one sample was available for the same age, expression values from equivalent samples were averaged. The same procedure was followed for the transcriptome data of the 15 cortical areas present for both 40 year old male and female samples used in Figure 4.5. RPKM normalized RNA-seq expression levels for human and macaque orthologous genes from both male and female individuals were obtained from Brawand et al. dataset (Brawand et al. 2011). Individual expression values were again normalized against total signal per sample (Dataset 5, Table 4.1). Microarray derived, RMA normalized values of gene expression values derived from substantia nigra of Parkinson's disease patients or Hutchinson Gilford Progeria Syndrome-derived fibroblasts and their corresponding controls were obtained from NCBI's GEO (Dataset 1 and 4, Table 4.1). Raw CEL files for arrays with gene expression levels in tissues with Alzheimer's disease were also downloaded from GEO. The later were RMA-normalized for consistency. We summarized per probe expression levels to ensemble gene IDs in the same manner as with the Human Gene Atlas data set and renormalized against the total expression signal in each array/sample. Finally, RMA-normalized microarray data from neocortex mice under CR and Z-normalized microarray data for skeletal muscle of individuals subject to CR were downloaded from GEO (Dataset 7, Table 4.1) and processed as previously described.

4.6.3 Co-expression analyses.

Co-expression analyses were carried out by obtaining the correlation coefficient across all possible pairs of PMCL-associated genes in brain samples of Alzheimer’s and Parkinson’s disease as well as corresponding control samples. To evaluate whether PMCL-associated genes were highly co-expressed, relative to background gene population, using any given dataset, the corresponding p-value was numerically determined by comparing the mean co-expression of PMCL-associated genes with the expected distribution of mean co-expression values computed from 100 000 random gene samples of the same size. Comparisons of mean co-expression of PMCL-associated genes across tissues and/or samples were carried out using paired t-tests.

4.6.4 Enrichment of disease down regulated genes.

Differential expression analysis was carried out using the disease expression datasets to compare disease against control conditions for each case of study using the LIMMA package in R (Smyth 2005). Significant biases in the proportion of disease-related down-regulated genes among our set of PMCL-associated genes was assessed by contrasting the observed proportion of these genes with the ones observed in at least 1 000 000 equally sized random sampled obtained from the background gene population. The test involving differentially down-regulated genes in 40 years old human males when compared to females was done following the same approach.

4.6.5 Functional enrichment analysis.

Biological Process GO Slim annotations where obtained from Ensembl’s (release 71) Biomart. Entrez IDs and Gene symbols annotations for Transcription factor target sites where obtained from the Molecular Signatures Database v4.0 (MSigDB) (http://www.broadinstitute.org/gsea/msigdb/index.jsp). These annotations are based on transcription factor binding sites defined in the TRANSFAC (version 7.4, http://www.gene-regulation.com/) database. Entrez IDs and Gene symbols where mapped to Ensembl IDs with a correspondence table downloaded from Ensembl’s Biomart. Gene sets sharing a binding site labelled as UNKNOWN where excluded from this analysis. To measure the enrichment in genes with any target binding site for a given transcriptional factor, we summarized TRANSFAC annotations by assigning a gene to a transcription factor if it contains any target for that TF in TRANSFAC annotations.

For consistency across all different enrichment analyses carried out, and in order to facilitate the use of the same Ensembl version throughout the study, we employed our own numerical methods to assess significant over-representation. Briefly, statistical enrichment of each analysed category (i.e., gene ontology, disease-down regulated genes, transcription factor targets, sex-specific differentially expressed genes, caloric restriction-associated genes etc.) among our set of PMCL-associated genes was assessed by performing a Z–test, where the expected representations and their standard deviations were obtained from 1 000 000 Monte Carlo simulations using random samples of 81 genes drawn from our curated set of 11 449 genes. Benjamini-Hochberg multiple testing corrections against the number of categories tested in each analysis was done (GO slim functional categories, n = 69, TRANSFAC specific factor target binding site, n = 501, TRASFAC summarized to transcription factors, n = 283). Categories with a resulting adj. $p < 0.05$ and with an excess of more than 1 PMCL-associated gene than expected, were deemed significantly enriched.

4.6.7 Statistical analysis.

All statistical analyses were carried out using the R statistical software package.

Authors’ contributions

AU and HG conceived and designed the study. ACM and JMS carried out the analyses presented. All authors contributed to the preparation of the manuscript.

Acknowledgements

We thank Leticia Ramirez-Lugo from the National University of Mexico for her technical support. This work was supported by CONACyT PhD scholarships to ACM and JMS, a Dorothy Hodgkin Research Fellowship and Royal Society Research Grant to AUO and Institutional support from the University of Lincoln to HG.

Supplementary Table S4.1

Tissue	Age (years)	Reference
Occipital cortex	72	(Bhardwaj et al. 2006)
Cerebellum	69.1	(K. L. Spalding et al. 2005)
Cardiomyocytes	66 (Occipital lobe - 6)	(Bergmann et al. 2009)
Pancreatic islets	42	(Perl et al. 2010)
Intestine (non epithelial cells)	15.9	(K. L. Spalding et al. 2005)
Intercostal skeletal muscle	15.1	(K. L. Spalding et al. 2005)
Adipocytes	9.5	(Kirsty L. Spalding et al. 2008)
Blood	0.33 (120 days Erythrocytes)	(K. L. Spalding et al. 2005; Whitehouse et al. 1982)

Table 4.3. GO slim terms and transcription factor target significantly enriched among PMCL-associated genes (Adjusted p-value < 0.05)

Gene ontology accession	Gene ontology term	O/E	Adj. p
GO:0030705	cytoskeleton-dependent intracellular transport	3 / 0.36	0.0003
GO:0006399	tRNA metabolic process	4 / 0.69	0.0013
GO:0007049	cell cycle	15 / 6.17	0.0024
GO:0000902	cell morphogenesis	11 / 4.41	0.0101
GO:0006457	protein folding	4 / 1.12	0.0491
GO:0051301	cell division	6 / 2.27	0.0491
GO:0006520	cellular amino acid metabolic process	6 / 2.25	0.0491
GO:0042254	ribosome biogenesis	3 / 0.80	0.0491

Transcriptional Factor Site	O/E	Adj. p
HSF	6 / 1.55	0.0416
USF2	5 / 1.27	0.0416
USF	10 / 3.93	0.0416
ELK1	15 / 6.87	0.0416
EFC (RFX1)	5 / 1.34	0.0494

5. General Discussion

While the detailed mechanisms in how a large brain underlies elevated cognitive capacity are still a matter of debate (Deary et al. 2010), one of its indubitable consequences is the ability to ponder on reasons and origins of this organ. The evolution of the brain is a process that has fascinated scientists for decades. Large brains confer cognitive advantages which can result in an increase in survival and a longer reproductive life, making them potentially advantageous for species (Allman et al. 1993; Gonzalez-Lagos et al. 2010; Isler and van Schaik 2009; Sol 2009a). On the other hand, large brains carry with them a high adaptive costs resulting from such factors as high metabolic demand, higher parental investment, diminished annual fertility and delayed reproductive age (Gonzalez-Lagos et al. 2010; Isler and van Schaik 2006a; W. R. Leonard et al. 2003; Roth and Dicke 2005; Weisbecker and Goswami 2010). Due to the complex interplay of factors both allowing for or resulting from a larger brain the precise nature of the genomic changes that account for differences in the size of the brain remain poorly understood (Dorus et al. 2004; Shi et al. 2006).

Throughout this thesis I have used a combination of comparative genomic approaches and transcriptomic analyses in order to further our understanding of the genomic footprint of two complex phenotypes, brain evolution and cellular longevity. This compendium represents the first genome wide scan for the association of changes in size of gene families, one of the main genetic drivers of phenotypic evolution (Lynch and Conery 2000), with the evolution of larger brains and larger neocortex in mammals. I identified a significant over-representation of GFS variations in line with increased encephalization and neocorticalization in mammals, and proved that this relationship is not accounted for by known correlates of these variables, nor is it a result of mere phylogenetic relatedness between the analysed species. I also found that these neural associated variations are particularly enriched in families with genes involved in processes such as immune system, cell-cell signalling and chemotaxis, and seem to be a response to the specific cellular, physiological and developmental demands of an increased brain size in mammals. These studies constitute a step in understanding the genetic footprints of the evolution of the brain and the neocortex in mammals.

Using an approach borrowed from comparative genomics, I also compare differences in post-mitotic cell longevity of seven different tissues with their transcriptomes in an attempt to shed a light on the basis of cellular longevity across tissues, particularly focusing on the genetic determinants of longevity in neurons. As a result, I identified a set of genes whose expression levels are closely associated with increased post-mitotic cellular longevity (PMCL) across a variety of human tissues ranging in longevity from 120 days to over 70 years. These genes are down regulated in the cerebral cortex and substantia nigra of Alzheimer's and Parkinson's disease patients, as well as in Hutchinson-Gilford progeria-derived fibroblasts, further suggesting their involvement in the regulation of cellular maintenance. Moreover, we found that sexual dimorphism in the expression patterns of PMCL-associated genes in the brain mirrors observed differences in average lifespan between sexes in humans and macaques, insinuating a link between differential demands in neuronal maintenance between males and females and level of activity of PMCL-associated genes. These results provide an insight into the machinery of post-mitotic maintenance of neural and non-neural tissues.

5.1 Variations in gene family size and encephalization

Genomic determinants underlying increased encephalization across mammalian lineages remains an open question. In chapter 2 I investigate if the large and frequent changes in the size of gene families observed in mammalian taxa play a part in shaping the differences in relative brain size among species. Using a genome-wide comparative approach, we examined changes in gene family size (GFS) and degree of encephalization in 39 fully sequenced mammalian species and found a highly significant over-representation of gene families displaying a positive association between their size and encephalization. This bias is particularly pronounced in families associated with specific biological functions. The most robust and consistent bias was observed in gene families associated with cell signalling, immune regulation and chemotaxis.

Both chemotaxis and cell signalling functions are known to play central roles in the development and function of the nervous system. Chemokines and their receptors play a crucial part in directing the proliferation and migration of immature neurons, glia and their precursors (reviewed in (Tran and Miller 2003)). Furthermore, chemokines and their receptors are important in neuroinflammatory diseases, strongly suggesting an important role in adult nervous system as well (De Groot and Woodroofe 2001). Chemokines where originally discovered and described in the immune system as regulators of leukocyte trafficking, inflammation, autoimmunity, angiogenesis and metastasis (Lira and Furtado

2012; Rossi and Zlotnik 2000); suggesting that this high overlap between chemotaxis and immune system could partially explain the observed enrichment of immune system-associated functions among gene families displaying the highest association between GFS and Ei. However, in recent years many other signalling and regulatory mechanisms originally described in the immune system, such as cytokines and immune related transcriptional regulators, have increasingly been found implicated in key neural-specific roles both in the developing and adult nervous system (Crampton et al. 2012; Gavalda et al. 2009; Gutierrez et al. 2005; Gutierrez et al. 2008; McKelvey et al. 2012; Nolan et al. 2011; O'Keeffe et al. 2008). Moreover, in the human cerebral cortex, immune system-related functions have been found to be significantly over-represented among genes displaying higher expression variability in the developing cerebral cortex than in the same tissues in adult (Sterner et al. 2012), hinting to a substantial involvement of immune-related signals during cortical development.

5.2 Changes in gene family size and neocorticalization.

Higher cognitive abilities such as thinking, consciousness and self-control reside in the neocortex. One of the key features of the latest evolutionary addition of the mammalian brain, the neocortex, it is multi-layered structure. Neocortical regions change in size, structure and occurrence depending on the species specific functional demands and is not universal across mammalian species (Krubitzer and Huffman 2000). The neocortex can be partitioned into different fields, through a process termed arealization, which has its basis in a combination of epigenetic and genetic mechanism (Alfano and Studer 2013; Dehay et al. 1996; D. D. O'Leary et al. 2007). In large-brained species the emergence of new areas has been suggested to deal with new functions, thus the molecular mechanism that are necessary for the formation of new regions might increase as the species brains get bigger.

Gene family size changes are the product of gene duplication and losses events, of particular importance are gene duplication events which may lead to the creation of new genes that perform a wider set of functions. Similarly, the increase in neocortical size has been linked to the creation of new areas in the neocortex, which are potentially areas that will perform new functions (Changizi 2001; Kaas et al. 2013). There has been propose that the duplication of association areas in the neocortex can be easily accomplished, so effortlessly, that even duplications may arise within the same species as it occurs among humans (Sereno and Huang 2006). How GFS changes and NR may then be associated? A hint is provided from the signalling molecule FGF8, which has been previously identify as a mediator of brain patterning during early development, and whose ectopic expression cause the duplication of the primary somatosensory area (S1) to create a new neocortical area (Fukuchi-Shimogori and Grove 2001).

As a general indication of the preferred gene function, the Gene Ontology enrichment analysis unveiled a connexion between Nr and gene family size. Along an Nr increased there has been found an increase in size of particular gene families that are significantly enriched in cell-cell signalling, chemotaxis and immune response biological processes. Cell-cell signalling and chemotaxis are functional sets of particular importance throughout the nervous system, and therefore is not difficult to think that the observed association between GFS and Nr reflects the functional demands that neocorticalization inflicts on the nervous system. If we look with further detail, cell- cell signalling annotations includes relevant biological functions such as synaptic signalling and neurotransmission. It is not surprising then that the enlarged association areas, in particular in human, are enriched in number synapses and preserve a particular distribution as indicated by increased density of dendritic spines and an elaborate dendritic branching pattern. As an example, among neocortically-enlarged hominids, there has been found a partial duplication event in the gene family of SRGAP2, which codes for a highly conserved protein expressed early in development required for spine maturation, neuronal migration, differentiation and neurite outgrowth (Charrier et al. 2012). A recent human specific duplication event in the GABA receptor family has also been identified; being GABA one of the main inhibitory neurotransmitter whose role is central for cell-cell signalling interactions between neural cells in mammalian species. Within the mammalian cortex, GABAergic neurons encompass one of the two major neuronal classes, where they control cortical plasticity, and have been shown to increase in number and complexity during human evolution (E. G. Jones 1993; Letinic et al. 2002). Lastly, as we have suggested previously, the overrepresentation of immune response functional annotations could be due to the gradual integration of the immune signalling system to the mammalian nervous system in order to meet the demands of an increasingly large brain (Castillo-Morales et al. 2014). Alternatively and or complementarily, given that Nr is positively associated to group size, perhaps it is increased exposure to pathogens corresponding with increased exposure to conspecifics which has increased demands on the immune system (Pasquaretta et al. 2014).

5.3. Molecular basis of neuronal post mitotic maintenance.

In Chapter 4 I explore the underlying mechanisms for sustaining post-mitotic cells in the human brain by comparing transcription profiles of cell types with different cell longevity. Terminally differentiated post-mitotic cells have different turnover and survival requirements. These differences can arise from differing cell maintenance mechanisms unique to each cell type or by the differential activation of a common molecular repertoire. Nowhere would these supporting mechanisms play a particularly critical role as in the human nervous system, where the vast majority of nerve cells cannot be replaced and need to survive as long as the organism, reaching even more than 100 years of age in some cases.

In this regard, we found that genes downregulated in AD and PD show increased levels of expression in line with differences of post-mitotic cell lifespan in other tissues. We carried out a genome wide screening to identify a transcriptional signature of long term post-mitotic maintenance to determine whether these genes are also down regulated in AD and PD. We detect at least 81 genes whose levels of expression are robustly correlated with cellular longevity, providing the first attempt to scan for genes potentially involved in post-mitotic cell longevity. Conversely, we also found that these genes are significantly under expressed in the cerebral cortex and substantia nigra of Alzheimer’s disease and Parkinson’s disease patients respectively. Remarkably, PMCL-associated genes are not particularly enrichment on down regulated genes in the visual cortex of Alzheimer’s disease patients. This result is particularly significant due to the fact that this region of the brain is known to be relatively spared from Alzheimer’s disease pathologies (Liang et al. 2007; Liang et al. 2008). Additionally, PMCL-associated genes appear to be collectively dysregulated, as expressed by a strong reduction in their level of co-expression in these degenerative conditions relative to the same tissues in non-affected people, further suggesting a general disruption of the cell maintenance machinery genes in neurodegenerative pathologies.

The involvement of the PMCL-associated genes in cell homeostasis in neural and non-neural cell types is also strongly supported by similar observations when analysing fibroblasts of Hutchinson-Gilford progeria syndrome patients, a genetic condition characterized by the dramatic, rapid appearance of aging beginning in childhood, and in which cellular senescence and tissue homeostasis is particularly disrupted (Bridger and Kill 2004). These results are consistent with the existence of a generalised molecular mechanisms for post-mitotic maintenance common to all tissue types.

Because neurons survive as long as the organism, any systematic differences in overall life expectancy between individuals are likely to be accompanied by differential activity of post-mitotic survival mechanisms. In this respect, the known sexual dimorphism in life expectancy in humans, with females living 6% longer lives than males (Clutton-Brock and Isvaran 2007; Kinsella 1998; Vina et al. 2005; Vina and Borras 2010) is likely to translate into corresponding variations in neuronal maintenance. In order to test whether brain expression of PMCL associated genes reflect the aforementioned sexual dimorphism in life expectancy, we compared their expression in both male and female human individuals and found that expression of PMCL-associated genes is significantly higher in human females relative to males. In the same manner we also studied gene expression profiles for male and female macaques, as this species displays a more pronounced sexual lifespan dimorphism than humans; with females living on average 72% more than males in wild populations (Clutton-Brock and Isvaran 2007). Interestingly, our results revealed a much more pronounced expression dimorphism of PMCL-associated genes in macaque female relative to males than the one observed in humans, where differences in lifespan between females and males in the wild are also more noticeable. All these results taken together suggest that PMCL-associated genes could constitute a potential signature of enhanced functional homeostasis in both neural and non-neural cell types.

Functional characterization of PMCL-associated genes carried out using Gene ontology enrichment analysis found that biological processes such as cytoskeletal-dependent transport, cell morphogenesis and protein folding are statistically overrepresented among these group of genes. Of particular interest are genes involved in resistance against protein misfolding identified among PMCL-associated genes, such as prefoldins (*PFDN2*), ubiquitin esterases (*UCHL1*), chaperonins (*CCT7*), chaperons (*HSP90AB1*) and associated adaptor proteins (*STIP1*) as well as proteosomal subunits (*PSMC4*); since both unfolded protein response (UPR) and the ubiquitin/proteasome system have been proposed to play an important part in the aging process in different species (Durieux et al. 2011; Kimata et al. 2006; Kruegel et al. 2011; Min et al. 2008; Morley and Morimoto

2004; Perez et al. 2009). Oxidative stress can lead to improperly folded proteins that are either retained within the lumen of the endoplasmic reticulum (ER) in complex with molecular chaperones or degraded through the 26S proteasome or through autophagy. Accumulation of misfolded proteins is also known to cause ER stress, which in turn can exacerbate oxidative stress entering a positive feedback loop ending in increased cell death (Gregersen and Bross 2010; Malhotra and Kaufman 2007). *HSP90* is known to modulate UPR (Marcu et al. 2002) and targeting *HSP90* can destabilise UPR induced cell death (Barrott and Haystead 2013; Davenport et al. 2008; Jackson 2013). Interestingly, mutants of *HSP90* are known to affect lifespan in *C. elegans, D. melanogaster* and *S. cerevisiae.*

Senescence at the cellular level is mostly studied in proliferating cells where the conditions under which cells stop dividing are assessed. In many tissues, however, cells have low turnover rates with neurons being required to live as long as the organism. The mechanisms enabling the maintenance of non-dividing cells remain largely unexplored as experimental models are impractical. Furthermore, as most animal models have short generation time and low lifespans it is possible that the molecular basis of post-mitotic cell maintenance in long lived species might not be shared in short lived organisms as rodent models.

The approach used in this study took advantage of the characterisation of cell turnover obtained for a number of human cells. The characterisation of further cell turnover estimates for other tissues will allow to further refine the list of genes involved in the regulation of cell maintenance.

Interestingly, my results are suggestive of the existence of a set of genes involved in post-mitotic cell maintenance across cell types. This study is the first evidence supporting a general post-mitotic maintenance mechanism.

Taken together, the results presented in this thesis provide insights into the molecular basis of brain size and morphology evolution as well as the underlying molecular mechanisms for post-mitotic cell which allows neurons in long lived organisms to survive.

Gene duplications that give origin to an increment in GFS can result in three different outcomes: nonfunctionalization of one copy by degenerative mutations, neofunctionalization or subfunctionalization (Lynch and Conery 2000). While nonfunctionalization represent the most common consequence of this phenomena due to the deleterious character of most mutations, the other two outcomes are the ones that are most likely to result in phenotype evolution, and as such restricting our analysis to these events might be key in order to capture functional changes in the genome deriving in the evolution of the brain. Furthermore, either due to positive selection, relaxed constrains or random drift, duplicated genes experience evolutionary rate acceleration that shapes subsequent evolution of gene duplicates. Tracking the rate of nonsynonymous to synonymous substitutions (dN/dS) of members of families with GFS associated to the evolution of the brain will permit us to evaluate functional consequences of these events. While most genome wide approaches to scan genes with evolutionary rates associated with larger brains focus on sets of relatively closely related taxa, such as primates or cetacean, executing this kinds of analysis along the whole of mammalian evolution will shed a light on the common genetic mechanisms that convergently contributed to shape the central nervous system in different mammalian taxa. Moreover, increase availability of both genomic data as well as measurements of other phenotypes more closely related to cognitive ability, such as total neuron number or neural connectivity will be of the utmost importance to the advancement of evolutionary neurobiology.

While the transcriptome-wide analysis here presented suggests the existence of a generalised cell longevity pathways in human tissues that becomes compromised in neurodegenerative conditions, identifying the precise mechanisms that allow long living tissues, particularly neurons, to maintain homeostasis for the entire lifetime of the organism is essential to understand both aging and neurodegeneration. In order to achieve such a goal, experimental follow ups need to address the hypothesis built in these thesis. This could be achieved by experimentally altering the activity of PMCL-associated genes on an *in vitro* system such as hESC-derived neurons using RNAi transfection, or in a short lived *in vivo* system, such as the nematode *C. elegans* or the turquoise killifish (*Nothobranchius furzeri*) with genome editing tools such as CRISPR/Cas system.

6. References

Adolphs, R., et al. (1994), 'Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala', *Nature*, 372 (6507), 669-72.

Aiello, L. C. and Dunbar, R. I. M. (1993), 'Neocortex Size, Group-Size, and the Evolution of Language', *Current Anthropology,* 34 (2), 184-93.

Aiello, L. C. and Wheeler, Peter (1995), 'The Expensive-Tissue Hypothesis: The Brain and the Digestive System in Human and Primate Evolution', *Current Anthropology,* 36 (2), 199-221.

Alfano, C. and Studer, M. (2013), 'Neocortical arealization: evolution, mechanisms, and open questions', *Dev Neurobiol,* 73 (6), 411-47.

Allen, J. S., Bruss, J., and Damasio, H. (2005), 'The aging brain: the cognitive reserve hypothesis and hominid evolution', *Am J Hum Biol,* 17 (6), 673-89.

Allfrey, V. G., Faulkner, R., and Mirsky, A. E. (1964), 'ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS', *Proceedings of the National Academy of Sciences of the United States of America,* 51 (5), 786-94.

Allison, T., et al. (1994), 'Face recognition in human extrastriate cortex', *J Neurophysiol,* 71 (2), 821-5.

Allman, J., McLaughlin, T., and Hakeem, A. (1993), 'Brain weight and life-span in primate species', *Proc Natl Acad Sci U S A,* 90 (1), 118-22.

Anthony, Raoul (1938), 'Anatomie comparée du cerveau (Doin, Paris 1928) -- Essai de recherche d'une expression anatomique approximative du degré d'organisation cérébrale, autre que le poids de l'encéphale comparé au poids du corps', *Bull Mem Soc Anthropol Paris,* 9 (9-1-3), 17-67.

Apweiler, R., et al. (2014), 'Activities at the Universal Protein Resource (UniProt)', *Nucleic Acids Research,* 42 (D1), D191-D98.

Aristide, L., et al. (2015), 'Encephalization and diversification of the cranial base in platyrrhine primates', *J Hum Evol,* 81, 29-40.

Azevedo, F. A., et al. (2009), 'Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain', *J Comp Neurol,* 513 (5), 532-41.

Ballard, J William O, et al. (2007), 'Sex differences in survival and mitochondrial bioenergetics during aging in Drosophila', *Aging Cell,* 6 (5), 699-708.

Barger, J. L., et al. (2008), 'A low dose of dietary resveratrol partially mimics caloric restriction and retards aging parameters in mice', *PLoS One,* 3 (6), e2264.

Barker, F. G. (1995), 'Phineas among the Phrenologists - the American Crowbar Case and Nineteenth-Century Theories of Cerebral Localization', *Journal of Neurosurgery,* 82 (4), 672-82.

Barrett, Emma LB and Richardson, David S (2011), 'Sex differences in telomeres and lifespan', *Aging Cell,* 10 (6), 913-21.

Barrickman, N. L. and Lin, M. J. (2010), 'Encephalization, expensive tissues, and energetics: An examination of the relative costs of brain size in strepsirrhines', *Am J Phys Anthropol,* 143 (4), 579-90.

Barrickman, N. L., et al. (2008), 'Life history costs and benefits of encephalization: a comparative test using data from long-term studies of primates in the wild', *J Hum Evol,* 54 (5), 568-90.

Barrott, J. J. and Haystead, T. A. (2013), 'Hsp90, an unlikely ally in the war on cancer', *FEBS J,* 280 (6), 1381-96.

Barton, R. A. (1996), 'Neocortex Size and Behavioural Ecology in Primates', *Proceedings: Biological Sciences,* 263 (1367), 173-77.

--- (2006), 'Primate brain evolution: Integrating comparative, neurophysiological, and ethological data', *Evolutionary Anthropology: Issues, News, and Reviews,* 15 (6), 224-36.

Barton, R. A. and Capellini, I. (2011), 'Maternal investment, life histories, and the costs of brain growth in mammals', *Proc Natl Acad Sci U S A,* 108 (15), 6169-74.

Benson, Mj , et al. (2008), 'Cutting edge: the dependence of plasma cells and independence of memory B cells on BAFF and APRIL', *J Immunol,* 180 (6), 3655-9.

Bergmann, O., et al. (2009), 'Evidence for cardiomyocyte renewal in humans', *Science,* 324 (5923), 98-102.

Bergstresser, P. R. and Taylor, J. R. (1977), 'Epidermal 'turnover time'--a new examination', *Br J Dermatol,* 96 (5), 503-9.

Bernardes de Jesus, B. and Blasco, M. A. (2012), 'Assessing cell and organ senescence biomarkers', *Circ Res,* 111 (1), 97-109.

Bhardwaj, R. D., et al. (2006), 'Neocortical neurogenesis in humans is restricted to development', *Proc Natl Acad Sci U S A,* 103 (33), 12564-8.

Bishop, Nicholas A., Lu, Tao, and Yankner, Bruce A. (2010), 'Neural mechanisms of ageing and cognitive decline', *Nature,* 464 (7288), 529-35.

Boehm, T. (2012), 'Evolution of vertebrate immunity', *Curr Biol,* 22 (17), R722-32.

Bond, Jacquelyn, et al. (2005), 'A centrosomal mechanism involving CDK5RAP2 and CENPJ controls brain size', *Nat Genet,* 37 (4), 353-55.

Boyle, A. P., et al. (2014), 'Comparative analysis of regulatory information and circuits across distant species', *Nature,* 512 (7515), 453-6.

Brawand, D., et al. (2011), 'The evolution of gene expression levels in mammalian organs', *Nature,* 478 (7369), 343-8.

Bridger, J. M. and Kill, I. R. (2004), 'Aging of Hutchinson-Gilford progeria syndrome fibroblasts is characterised by hyperproliferation and increased apoptosis', *Experimental Gerontology,* 39 (5), 717-24.

Burkart, J. M., Hrdy, S. B., and Van Schaik, C. P. (2009), 'Cooperative Breeding and Human Cognitive Evolution', *Evolutionary Anthropology,* 18 (5), 175-86.

Burki, F. and Kaessmann, H. (2004), 'Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux', *Nat Genet,* 36 (10), 1061-3.

Burtner, C. R. and Kennedy, B. K. (2010), 'Progeria syndromes and ageing: what is the connection?', *Nat Rev Mol Cell Biol,* 11 (8), 567-78.

Burton, D. G. (2009), 'Cellular senescence, ageing and disease', *Age (Dordr),* 31 (1), 1-9.

Bush, E. C. and Lahn, B. T. (2005), 'Selective constraint on noncoding regions of hominid genomes', *PLoS Comput Biol,* 1 (7), e73.

Bustamante, C. D., et al. (2005), 'Natural selection on protein-coding genes in the human genome', *Nature,* 437 (7062), 1153-7.

Byrne, Richard W. and Whiten, Andrew (1988), *Machiavellian intelligence : social expertise and the evolution of intellect in monkeys, apes, and humans* (Oxford science publications; Oxford

New York: Clarendon Press ;

Oxford University Press) xiv, 413 p.

Caceres, M., et al. (2007), 'Increased cortical expression of two synaptogenic thrombospondins in human brain evolution', *Cereb Cortex,* 17 (10), 2312-21.

Camus, M Florencia, Clancy, David J, and Dowling, Damian K (2012), 'Mitochondria, maternal inheritance, and male aging', *Current Biology*.

Capra, J. A., et al. (2013), 'Many human accelerated regions are developmental enhancers', *Philos Trans R Soc Lond B Biol Sci,* 368 (1632), 20130025.

Cassese, G , et al. (2003), '- Plasma cell survival is mediated by synergistic effects of cytokines and adhesion-dependent signals', *J Immunol,* 171 (4), 1684-90.

Castillo-Morales, A., et al. (2014), 'Increased brain size in mammals is associated with size variations in gene families with cell signalling, chemotaxis and immune-related functions', *Proceedings of the Royal Society B-Biological Sciences,* 281 (1775).

Changizi, M. A. (2001), 'Principles underlying mammalian neocortical scaling', *Biol Cybern,* 84 (3), 207-15.

Charrier, C., et al. (2012), 'Inhibition of SRGAP2 Function by Its Human-Specific Paralogs Induces Neoteny during Spine Maturation', *Cell,* 149 (4), 923-35.

Charvet, C. J. and Finlay, B. L. (2012), 'Embracing covariation in brain evolution: large brains, extended development, and flexible primate social systems', *Prog Brain Res,* 195, 71-87.

Chen-Plotkin, A. S., et al. (2008), 'Variations in the progranulin gene affect global gene expression in frontotemporal lobar degeneration', *Hum Mol Genet,* 17 (10), 1349-62.

Chen, B. and Wagner, A. (2012), 'Hsp90 is important for fecundity, longevity, and buffering of cryptic deleterious variation in wild fly populations', *BMC Evol Biol,* 12, 25.

Choi, S. S. and Lahn, B. T. (2003), 'Adaptive evolution of MRG, a neuron-specific gene family implicated in nociception', *Genome Res,* 13 (10), 2252-9.

Clutton-Brock, T. H. and Isvaran, K. (2007), 'Sex differences in ageing in natural populations of vertebrates', *Proc Biol Sci,* 274 (1629), 3097-104.

Cole, G. M. and Frautschy, S. A. (2007), 'The role of insulin and neurotrophic factor signaling in brain aging and Alzheimer's Disease', *Exp Gerontol,* 42 (1-2), 10-21.

Connor, R. C. (2007), 'Dolphin social intelligence: complex alliance relationships in bottlenose dolphins and a consideration of selective environments for extreme brain size evolution in mammals', *Philos Trans R Soc Lond B Biol Sci,* 362 (1480), 587-602.

Consortium, Encode Project, et al. (2007), 'Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project', *Nature,* 447 (7146), 799-816.

Cooper, D. N. and Kehrer-Sawatzki, H. (2011), 'Exploring the potential relevance of human-specific genes to complex disease', *Hum. Gen.,* 5, 99-107.

Coppede, F and Migliore, L (2010), 'DNA repair in premature aging disorders and neurodegeneration', *Curr Aging Sci,* 3 (1), 3-19.

Cordain, Loren, Watkins, Bruce A, and Mann, Neil J (2001), 'Fatty acid composition and energy density of foods available to African hominids'.

Covas, R. and Griesser, M. (2007), 'Life history and the evolution of family living in birds', *Proc Biol Sci,* 274 (1616), 1349-57.

Crampton, S. J., et al. (2012), 'Exposure of foetal neural progenitor cells to IL-1beta impairs their proliferation and alters their differentiation - a role for maternal inflammation?', *J Neurochem,* 120 (6), 964-73.

Crawford, M. A., et al. (1999), 'Evidence for the unique function of docosahexaenoic acid during the evolution of the modern hominid brain', *Lipids,* 34 (1), S39-S47.

Crick, Francis and Koch, Christof (1990), 'Towards a neurobiological theory of consciousness', *Seminars in the Neurosciences* (2: Saunders Scientific Publications), 263-75.

Cunningham, F., et al. (2015), 'Ensembl 2015', *Nucleic Acids Res,* 43 (Database issue), D662-9.

Davenport, E. L., Morgan, G. J., and Davies, F. E. (2008), 'Untangling the unfolded protein response', *Cell Cycle,* 7 (7), 865-9.

De Groot, C. J. and Woodroofe, M. N. (2001), 'The role of chemokines and chemokine receptors in CNS inflammation', *Prog Brain Res,* 132, 533-44.

de Magalhaes, J. P. and Faragher, R. G. A. (2008), 'Cell divisions and mammalian aging: integrative biology insights from genes that regulate longevity', *Bioessays,* 30 (6), 567-78.

de Magalhaes, J. P., et al. (2012), 'Genome-Environment Interactions That Modulate Aging: Powerful Targets for Drug Discovery', *Pharmacological Reviews,* 64 (1), 88-101.

de Sousa, Alexandra A. and Proulx, Michael J. (2014), 'What can volumes reveal about human brain evolution? A framework for bridging behavioral, histometric, and volumetric perspectives', *Front Neuroanat,* 8.

Deaner, R. O., Nunn, C. L., and van Schaik, C. P. (2000), 'Comparative tests of primate cognition: different scaling methods produce different results', *Brain Behav Evol,* 55 (1), 44-52.

Deaner, R. O., Barton, R. A., and Van Schaik, C. (2003), 'Primate brains and life histories: renewing the connection', in Kappeler P.M and Pereira M.E (eds.), *Primates Life Histories and Socioecology* (Chicago, IL: The University of Chicago Press), 233-65.

Deaner, R. O., et al. (2007), 'Overall brain size, and not encephalization quotient, best predicts cognitive ability across non-human primates', *Brain Behav Evol,* 70 (2), 115-24.

Deary, I. J., Penke, L., and Johnson, W. (2010), 'The neuroscience of human intelligence differences', *Nat Rev Neurosci,* 11 (3), 201-11.

DeFelipe, Javier (2011), 'The evolution of the brain, the human nature of cortical circuits and intellectual creativity', *Frontiers in Neuroanatomy,* 5.

Dehay, C., et al. (1996), 'Contribution of thalamic input to the specification of cytoarchitectonic cortical fields in the primate: effects of bilateral enucleation in the fetal monkey on the boundaries, dimensions, and gyrification of striate and extrastriate cortex', *J Comp Neurol,* 367 (1), 70-89.

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

Demuth, J. P., et al. (2006), 'The evolution of mammalian gene families', *PLoS One,* 1, e85.

Dorus, S., et al. (2004), 'Accelerated evolution of nervous system genes in the origin of Homo sapiens', *Cell,* 119 (7), 1027-40.

Dorus, S., et al. (2006), 'Sonic Hedgehog, a key development gene, experienced intensified molecular evolution in primates', *Human Molecular Genetics,* 15 (13), 2031-37.

Drachman, D. A. (1997), 'Aging and the brain: a new frontier', *Ann Neurol,* 42 (6), 819-28.

Dror, Y. and Hopp, M. (2014), 'Hair for brain trade-off, a metabolic bypass for encephalization', *Springerplus,* 3, 562.

Dukas, R. (2004), 'Evolutionary biology of animal cognition', *Annual Review of Ecology Evolution and Systematics,* 35, 347-74.

Dumas, L. (2012), 'DUF1220-domain copy number implicated in human brain-size pathology and evolution', *Am. J. Hum. Genet.,* 91, 444-54.

Dunbar, R. I. (1992), 'Neocortex Size as a Constraint on Group-Size in Primates', *Journal of Human Evolution,* 22 (6), 469-93.

--- (1993), 'Coevolution of Neocortical Size, Group-Size and Language in Humans', *Behavioral and Brain Sciences,* 16 (4), 681-94.

--- (2009), 'The social brain hypothesis and its implications for social evolution', *Ann Hum Biol,* 36 (5), 562-72.

Dunbar, R. I. and Shultz, S. (2007a), 'Evolution in the social brain', *Science,* 317 (5843), 1344-7.

--- (2007b), 'Understanding primate brain evolution', *Philos Trans R Soc Lond B Biol Sci,* 362 (1480), 649-58.

Durieux, J., Wolff, S., and Dillin, A. (2011), 'The cell-non-autonomous nature of electron transport chain-mediated longevity', *Cell,* 144 (1), 79-91.

Dutta, D., et al. (2012), 'Contribution of Impaired Mitochondrial Autophagy to Cardiac Aging Mechanisms and Therapeutic Opportunities', *Circulation Research,* 110 (8), 1125-38.

Eccles, JohnC (1994), 'The Evolution of Consciousness', *How the SELF Controls Its BRAIN* (Springer Berlin Heidelberg), 113-24.

Eisen, M. B., et al. (1998), 'Cluster analysis and display of genome-wide expression patterns', *Proc Natl Acad Sci U S A,* 95 (25), 14863-8.

Emery, N. J. and Clayton, N. S. (2004), 'The mentality of crows: convergent evolution of intelligence in corvids and apes', *Science,* 306 (5703), 1903-7.

Enard, W. (2002), 'Molecular evolution of FOXP2, a gene involved in speech and language', *Nature,* 418, 869-72.

--- (2011), 'FOXP2 and the role of cortico-basal ganglia circuits in speech and language evolution', *Curr Opin Neurobiol,* 21 (3), 415-24.

Enard, W., et al. (2002), 'Molecular evolution of FOXP2, a gene involved in speech and language', *Nature,* 418 (6900), 869-72.

Enard, W., et al. (2009), 'A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice', *Cell,* 137 (5), 961-71.

Enticott, P. G., et al. (2008), 'Mirror neuron activation is associated with facial emotion processing', *Neuropsychologia,* 46 (11), 2851-54.

Evans, C. G., Wisen, S., and Gestwicki, J. E. (2006), 'Heat shock proteins 70 and 90 inhibit early stages of amyloid beta-(1-42) aggregation in vitro', *J Biol Chem,* 281 (44), 33182-91.

Evans, P. D. (2004), 'Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans', *Hum. Mol. Genet.,* 13, 489-94.

Faheem, M., et al. (2015), 'Molecular genetics of human primary microcephaly: an overview', *BMC Med Genomics*, 8 Suppl 1, S4.

Falconer, Douglas S. and Mackay, Trudy F. C. (1996), *Introduction to quantitative genetics* (4 edn.; Harlow, UK: Longmans Green).

Faria, R. and Navarro, A. (2010), 'Chromosomal speciation revisited: rearranging theory with pieces of evidence', *Trends Ecol Evol*, 25 (11), 660-9.

Feder, J. L., Nosil, P., and Flaxman, S. M. (2014), 'Assessing when chromosomal rearrangements affect the dynamics of speciation: implications from computer simulations', *Front Genet*, 5, 295.

Fedrigo, O., et al. (2011), 'A potential role for glucose transporters in the evolution of human brain size', *Brain Behav Evol*, 78 (4), 315-26.

Felsenstein, Joseph (1985), 'Phylogenies and the Comparative Method', *The American Naturalist*, 125 (1), 1-15.

Ferland, R. J., et al. (2004), 'Abnormal cerebellar development and axonal decussation due to mutations in AHI1 in Joubert syndrome', *Nature Genetics*, 36 (9), 1008-13.

Finarelli, J. A. (2010), 'Does encephalization correlate with life history or metabolic rate in Carnivora?', *Biol Lett*, 6 (3), 350-3.

Fish, J. L. and Lockwood, C. A. (2003), 'Dietary constraints on encephalization in primates', *Am J Phys Anthropol*, 120 (2), 171-81.

Fishel, M. L., Vasko, M. R., and Kelley, M. R. (2007), 'DNA repair in neurons: so if they don't divide what's to repair?', *Mutat Res*, 614 (1-2), 24-36.

Flicek, P., et al. (2012), 'Ensembl 2012', *Nucleic Acids Res*, 40 (Database issue), D84-90.

Florio, Marta, et al. (2015), 'Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion', *Science*, 347 (6229), 1465-70.

Fortna, A., et al. (2004), 'Lineage-specific gene duplication and loss in human and great ape evolution', *PLoS Biol*, 2 (7), E207.

Freckleton, R. P., Harvey, P. H., and Pagel, M. (2002), 'Phylogenetic analysis and comparative data: a test and review of evidence', *Am Nat*, 160 (6), 712-26.

Fukuchi-Shimogori, T. and Grove, E. A. (2001), 'Neocortex patterning by the secreted signaling molecule FGF8', *Science*, 294 (5544), 1071-4.

Gallese, V., et al. (2002), 'Action representation and the inferior parietal lobule', *Common Mechanisms in Perception and Action*, 19, 334-55.

Garland, T., Jr., Bennett, A. F., and Rezende, E. L. (2005), 'Phylogenetic approaches in comparative physiology', *J Exp Biol*, 208 (Pt 16), 3015-35.

Gavalda, N., Gutierrez, H., and Davies, A. M. (2009), 'Developmental regulation of sensory neurite growth by the tumor necrosis factor superfamily member LIGHT', *J Neurosci,* 29 (6), 1599-607.

Gaya-Vidal, M. and Alba, M. M. (2014), 'Uncovering adaptive evolution in the human lineage', *BMC Genomics,* 15, 599.

Gemmell, Neil J, Metcalf, Victoria J, and Allendorf, Fred W (2004), 'Mother's curse: the effect of mtDNA on individual fitness and population viability', *Trends in ecology & evolution,* 19 (5), 238-44.

Gerstein, M. B., et al. (2014), 'Comparative analysis of the transcriptome across distant species', *Nature,* 512 (7515), 445-8.

Gilbert, Paul, Price, John, and Allan, Steven (1995), 'Social comparison, social attractiveness and evolution: How might they be related?', *New Ideas in Psychology,* 13 (2), 149-65.

Gilbert, S. L., Dobyns, W. B., and Lahn, B. T. (2005), 'Genetic links between brain development and brain evolution', *Nature Reviews Genetics,* 6 (7), 581-90.

Gokcumen, O., et al. (2011), 'Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection', *Genome Biol,* 12 (5), R52.

Gonzalez-Lagos, C., Sol, D., and Reader, S. M. (2010), 'Large-brained mammals live longer', *J Evol Biol,* 23 (5), 1064-74.

Goto, H., et al. (2009), 'The identification and functional implications of human-specific "fixed" amino acid substitutions in the glutamate receptor family', *BMC Evol Biol,* 9, 224.

Gregersen, N. and Bross, P. (2010), 'Protein Misfolding and Cellular Stress: An Overview', *Protein Misfolding and Cellular Stress in Disease and Aging: Concepts and Protocols,* 648, 3-23.

Greig, L. C., et al. (2013), 'Molecular logic of neocortical projection neuron specification, development and diversity', *Nat Rev Neurosci,* 14 (11), 755-69.

Grober, E., et al. (1992), 'Skill learning and repetition priming in Alzheimer's disease', *Neuropsychologia,* 30 (10), 849-58.

Gulati, N., et al. (2013), 'Creation of Differentiation-Specific Genomic Maps of Human Epidermis through Laser Capture Microdissection', *J Invest Dermatol,* 133 (11), 2640-42.

Gutierrez, H., et al. (2005), 'NF-kappaB signalling regulates the growth of neural processes in the developing PNS and CNS', *Development,* 132 (7), 1713-26.

Gutierrez, H., et al. (2011), 'Protein Amino Acid Composition: A Genomic Signature of Encephalization in Mammals', *PLoS ONE,* 6 (11), e27261.

Gutierrez, H., et al. (2008), 'Nuclear factor kappa B signaling either stimulates or inhibits neurite growth depending on the phosphorylation status of p65/RelA', *J Neurosci,* 28 (33), 8246-56.

Hahn, M. W., Han, M. V., and Han, S. G. (2007), 'Gene family evolution across 12 Drosophila genomes', *PLoS Genet,* 3 (11), e197.

Hahn, M. W., et al. (2005), 'Estimating the tempo and mode of gene family evolution from comparative genomic data', *Genome Res,* 15 (8), 1153-60.

Han, M. V., et al. (2009), 'Adaptive evolution of young gene duplicates in mammals', *Genome Res,* 19 (5), 859-67.

Harrington, Anthony W and Ginty, David D (2013), 'Long-distance retrograde neurotrophic factor signalling in neurons', *Nature Reviews Neuroscience,* 14 (3), 177-87.

Hartwell, L. H., et al. (1999), 'From molecular to modular cell biology', *Nature,* 402 (6761 Suppl), C47-52.

Harvey, P. H., Clutton-Brock, T. H., and Mace, G. M. (1980), 'Brain size and ecology in small mammals and primates', *Proc Natl Acad Sci U S A,* 77 (7), 4387-9.

Hawrylycz, M. J., et al. (2012), 'An anatomically comprehensive atlas of the adult human brain transcriptome', *Nature,* 489 (7416), 391-9.

Herculano-Houzel, S. (2011), 'Brains matter, bodies maybe not: the case for examining neuron numbers irrespective of body size', *Ann N Y Acad Sci,* 1225, 191-9.

Herculano-Houzel, S., et al. (2007), 'Cellular scaling rules for primate brains', *Proc Natl Acad Sci U S A,* 104 (9), 3562-7.

Heuer, E., et al. (2012), 'Nonhuman primate models of Alzheimer-like cerebral proteopathy', *Curr Pharm Des,* 18 (8), 1159-69.

Hill, William G. (2010), 'Understanding and using quantitative genetic variation', *Philosophical Transactions of the Royal Society B,* 365, 73 - 85.

Ho, J. W., et al. (2014), 'Comparative analysis of metazoan chromatin organization', *Nature,* 512 (7515), 449-52.

Holland, L. Z. and Short, S. (2008), 'Gene duplication, co-option and recruitment during the origin of the vertebrate brain from the invertebrate chordate brain', *Brain Behav Evol,* 72 (2), 91-105.

Homouz, D. and Kudlicki, A. S. (2013), 'The 3D organization of the yeast genome correlates with co-expression and reflects functional relations between genes', *PLoS One*, 8 (1), e54699.

Hoover, K. C. (2013), 'Evolution of olfactory receptors', *Methods Mol Biol*, 1003, 241-9.

http://brainspan.org 'BrainSpan Atlas of the Developing Human Brain'.

http://www.ensembl.org 'Ensembl release 76', 03/09/2014.

http://www.timetree.org/ 'TimeTree2', 05/01/2015.

Huber, R., et al. (1997), 'Microhabitat use, trophic patterns, and the evolution of brain structure in African cichlids', *Brain Behavior and Evolution*, 50 (3), 167-82.

Hubisz, M. J. and Pollard, K. S. (2014), 'Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution', *Curr Opin Genet Dev*, 29, 15-21.

Hughes, A. L. and Friedman, R. (2004), 'Shedding genomic ballast: extensive parallel loss of ancestral gene families in animals', *J Mol Evol*, 59 (6), 827-33.

Iacoboni, Marco (2005), 'Neural mechanisms of imitation', *Current Opinion in Neurobiology*, 15 (6), 632-37.

Iizuka, H. (1994), 'Epidermal turnover time', *J Dermatol Sci*, 8 (3), 215-7.

Isler, K. (2011), 'Energetic trade-offs between brain size and offspring production: Marsupials confirm a general mammalian pattern', *Bioessays*, 33 (3), 173-79.

Isler, K. and van Schaik, C. P. (2006a), 'Metabolic costs of brain size evolution', *Biol Lett*, 2 (4), 557-60.

Isler, K. and van Schaik, C. (2006b), 'Costs of encephalization: the energy trade-off hypothesis tested on birds', *J Hum Evol*, 51 (3), 228-43.

Isler, K. and van Schaik, C. P. (2009), 'The Expensive Brain: a framework for explaining evolutionary changes in brain size', *J Hum Evol*, 57 (4), 392-400.

Jackson, S. E. (2013), 'Hsp90: structure and function', *Top Curr Chem*, 328, 155-240.

Jaiswal, M, et al. (2012), 'Probing Mechanisms That Underlie Human Neurodegenerative Diseases in Drosophila', *Annual Review of Genetics*, 46, 371-96.

Jansen, R. C. and Nap, J. P. (2001), 'Genetical genomics: the added value from segregation', *Trends Genet*, 17 (7), 388-91.

Jemielity, Stephanie, et al. (2007), 'Short telomeres in short-lived males: what are the molecular and evolutionary causes?', *Aging Cell*, 6 (2), 225-33.

Jerison, H. J. (1973), *Evolution of the brain and intelligence* (New York,: Academic Press) xiv, 482 p.

--- (1985), 'Animal Intelligence as Encephalization', *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences,* 308 (1135), 21-35.

--- (1990), 'Fossil Evidence on the Evolution of the Neocortex', in EdwardG Jones and Alan Peters (eds.), *Comparative Structure and Evolution of Cerebral Cortex, Part I* (Cerebral Cortex, 8A: Springer US), 285-309.

Jiang, Ao, et al. (2015), 'Seasonality and Age is Positively Related to Brain Size in Andrew’s Toad (Bufo andrewsi)', *Evolutionary Biology,* 1-10.

Johnson, R., et al. (2010), *Human accelerated region 1 noncoding RNA is repressed by REST in Huntington's disease* (41) 269-74.

Jones, E. G. (1993), 'Gabaergic Neurons and Their Role in Cortical Plasticity in Primates', *Cerebral Cortex,* 3 (5), 361-72.

Jones, K. E. and MacLarnon, A. M. (2004), 'Affording larger brains: testing hypotheses of mammalian brain evolution on bats', *Am Nat,* 164 (1), E20-31.

Kaas, J. H. (1989), 'The evolution of complex sensory systems in mammals', *J Exp Biol,* 146, 165-76.

--- (2006), 'Evolution of the neocortex', *Current Biology,* 16 (21), R910-R14.

--- (2011), 'Neocortex in early mammals and its subsequent variations', *Ann N Y Acad Sci,* 1225, 28-36.

--- (2013), 'The Evolution of Brains from Early Mammals to Humans', *Wiley Interdiscip Rev Cogn Sci,* 4 (1), 33-45.

Kaas, J. H., Gharbawie, O. A., and Stepniewska, I. (2013), 'Cortical networks for ethologically relevant behaviors in primates', *Am J Primatol,* 75 (5), 407-14.

Kajiya, K., et al. (2001), 'Molecular bases of odor discrimination: Reconstitution of olfactory receptors that recognize overlapping sets of odorants', *Journal of Neuroscience,* 21 (16), 6018-25.

Kaufman, J. A., Hladik, C. M. , and Pasquet, P. (2003), 'On the Expensive-Tissue Hypothesis: Independent Support from Highly Encephalized Fish', *Current Anthropology,* 44 (5), 705-07.

Keeney, J. G., Dumas, L., and Sikela, J. M. (2014a), 'The case for DUF1220 domain dosage as a primary contributor to anthropoid brain expansion', *Frontiers in Human Neuroscience,* 8, 427.

Keeney, J. G., et al. (2014b), 'DUF1220 protein domains drive proliferation in human neural stem cells and are associated with increased cortical volume in anthropoid primates', *Brain Struct Funct.*

Kelley, J. L. and Swanson, W. J. (2008), 'Positive selection in the human genome: from genome scans to biological significance', *Annu Rev Genomics Hum Genet,* 9, 143-60.

Kimata, Y., et al. (2006), 'Yeast unfolded protein response pathway regulates expression of genes for anti-oxidative stress and for cell surface proteins', *Genes Cells,* 11 (1), 59-69.

Kinsella, K.; Gist YJ. (1998), 'Gender and aging: mortality and health', *US Department of Commerce Economics and Statistics Administration Bureau of the Census, Washington, DC*

Kirkpatrick, M. and Barton, N. (2006), 'Chromosome inversions, local adaptation and speciation', *Genetics,* 173 (1), 419-34.

Klein, R. J., et al. (2005), 'Complement factor H polymorphism in age-related macular degeneration', *Science,* 308 (5720), 385-9.

Kole, A. J., Annis, R. P., and Deshmukh, M. (2013), 'Mature neurons: equipped for survival', *Cell Death Dis,* 4, e689.

Konopka, G., et al. (2009), 'Human-specific transcriptional regulation of CNS development genes by FOXP2', *Nature,* 462 (7270), 213-7.

Kosiol, Carolin, et al. (2008), 'Patterns of Positive Selection in Six Mammalian Genomes', *PLoS Genet,* 4 (8), e1000144.

Kotrschal, K., Van Staaden, M. J., and Huber, R. (1998), 'Fish brains: evolution and environmental relationships', *Reviews in Fish Biology and Fisheries,* 8 (4), 373-408.

Kozlovsky, D. Y., et al. (2014), 'Chickadees with bigger brains have smaller digestive tracts: a multipopulation comparison', *Brain Behav Evol,* 84 (3), 172-80.

Kraft, P. and Hunter, D. J. (2009), 'Genetic risk prediction--are we there yet?', *N Engl J Med,* 360 (17), 1701-3.

Krubitzer, L. and Huffman, K. J. (2000), 'Arealization of the neocortex in mammals: genetic and epigenetic contributions to the phenotype', *Brain Behav Evol,* 55 (6), 322-35.

Kruegel, U., et al. (2011), 'Elevated proteasome capacity extends replicative lifespan in Saccharomyces cerevisiae', *PLoS Genet,* 7 (9), e1002253.

Krylov, D. M., et al. (2003), 'Gene Loss, Protein Sequence Divergence, Gene Dispensability, Expression Level, and Interactivity Are Correlated in Eukaryotic Evolution', *Genome Research,* 13 (10), 2229-35.

Kudlow, B. A., Kennedy, B. K., and Monnat, R. J., Jr. (2007), 'Werner and Hutchinson-Gilford progeria syndromes: mechanistic basis of human progeroid diseases', *Nat Rev Mol Cell Biol,* 8 (5), 394-404.

Kumar, P., et al. (2007), 'CHIP and HSPs interact with beta-APP in a proteasome-dependent manner and influence Abeta metabolism', *Hum Mol Genet,* 16 (7), 848-64.

Kumar, S. and Hedges, S. B. (2011), 'TimeTree2: species divergence times on the iPhone', *Bioinformatics,* 27 (14), 2023-4.

Kuzawa, C. W. (1998), 'Adipose tissue in human infancy and childhood: an evolutionary perspective', *Am J Phys Anthropol,* Suppl 27, 177-209.

Lai, C. S. L., et al. (2001), 'A forkhead-domain gene is mutated in a severe speech and language disorder', *Nature,* 413 (6855), 519-23.

Lam, YA, et al. (2000), 'Inhibition of the ubiquitin-proteasome system in Alzheimer's disease', *Proc Natl Acad Sci U S A,* 97 (18), 9902-6.

Lanni, C., et al. (2010), 'The expanding universe of neurotrophic factors: therapeutic potential in aging and age-associated disorders', *Curr Pharm Des,* 16 (6), 698-717.

Lassek, William D. and Gaulin, Steven J. C. (2008), 'Waist-hip ratio and cognitive ability: is gluteofemoral fat a privileged store of neurodevelopmental resources?', *Evolution and Human Behavior,* 29 (1), 26-34.

Lee, H. K., et al. (2004), 'Coexpression analysis of human genes across many microarray data sets', *Genome Res,* 14 (6), 1085-94.

Lefebvre, L., Reader, S. M., and Sol, D. (2004), 'Brains, innovations and evolution in birds and primates', *Brain Behav Evol,* 63 (4), 233-46.

Lefebvre, Louis, et al. (1997), 'Feeding innovations and forebrain size in birds', *Animal Behaviour,* 53 (3), 549-60.

Lemaitre, J. F., et al. (2009), 'Sperm competition and brain size evolution in mammals', *J Evol Biol,* 22 (11), 2215-21.

Leonard, W. R., Snodgrass, J. J., and Robertson, M. L. (2007), 'Effects of brain evolution on human nutrition and metabolism', *Annu Rev Nutr,* 27, 311-27.

Leonard, W. R., et al. (2003), 'Metabolic correlates of hominid brain evolution', *Comp Biochem Physiol A Mol Integr Physiol,* 136 (1), 5-15.

Leonard, WilliamR, Snodgrass, J. Josh, and Robertson, MarciaL (2011), 'Diet and Brain Evolution: Nutritional Implications of Large Human Brain Size', in Victor R. Preedy, Ronald Ross Watson, and Colin R. Martin (eds.), *Handbook of Behavior, Food and Nutrition* (Springer New York), 3-15.

Letinic, K., Zoncu, R., and Rakic, P. (2002), 'Origin of GABAergic neurons in the human neocortex', *Nature,* 417 (6889), 645-49.

Li, J. B., et al. (2004), 'Comparative and basal genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene', *Cell,* 117 (4), 541-52.

Liang, W. S., et al. (2008), 'Altered neuronal gene expression in brain regions differentially affected by Alzheimer's disease: a reference data set', *Physiol Genomics,* 33 (2), 240-56.

Liang, W. S., et al. (2007), 'Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain', *Physiol Genomics,* 28 (3), 311-22.

Lira, S. A. and Furtado, G. C. (2012), 'The biology of chemokines and their receptors', *Immunologic Research,* 54 (1-3), 111-20.

Liu, G. H., et al. (2011), 'Recapitulation of premature ageing with iPSCs from Hutchinson-Gilford progeria syndrome', *Nature,* 472 (7342), 221-5.

Loerch, P. M., et al. (2008), 'Evolution of the aging brain transcriptome and synaptic regulation', *PLoS One,* 3 (10), e3329.

Long, A. D. and Langley, C. H. (1999), 'The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits', *Genome Res,* 9 (8), 720-31.

Lynch, M. and Walsh, B. (1998), *Genetics and analysis of quantitative traits* (Sunderland, Mass.: Sinauer) xvi, 980 p.

Lynch, M. and Conery, J. S. (2000), 'The evolutionary fate and consequences of duplicate genes', *Science,* 290 (5494), 1151-55.

Magrassi, L., Leto, K., and Rossi, F. (2013), 'Lifespan of neurons is uncoupled from organismal lifespan', *Proceedings of the National Academy of Sciences of the United States of America,* 110 (11), 4374-79.

Malhotra, J. D. and Kaufman, R. J. (2007), 'Endoplasmic reticulum stress and oxidative stress: a vicious cycle or a double-edged sword?', *Antioxid Redox Signal,* 9 (12), 2277-93.

Malnic, B., et al. (1999), 'Combinatorial receptor codes for odors', *Cell,* 96 (5), 713-23.

Manayi, A., et al. (2014), 'Methods for the discovery of new anti-aging products - targeted approaches', *Expert Opinion on Drug Discovery,* 9 (4), 383-405.

Manolio, T. A., et al. (2009), 'Finding the missing heritability of complex diseases', *Nature,* 461 (7265), 747-53.

Marcu, M. G., et al. (2002), 'Heat shock protein 90 modulates the unfolded protein response by stabilizing IRE1alpha', *Mol Cell Biol,* 22 (24), 8506-13.

Marjoram, P., Zubair, A., and Nuzhdin, S. V. (2014), 'Post-GWAS: where next? More samples, more SNPs or more biology?', *Heredity (Edinb),* 112 (1), 79-88.

Marques, A. C. and Ponting, C. P. (2014), 'Intergenic lncRNAs and the evolution of gene expression', *Curr Opin Genet Dev,* 27, 48-53.

Martin, R. D. (1981), 'Relative brain size and basal metabolic rate in terrestrial vertebrates', *Nature,* 293 (5827), 57-60.

--- (1983), *Human brain evolution in an ecological context* (Fifty-second James Arthur lecture on the evolution of the human brain; New York: American Museum of Natural History) 58 p.

--- (1990), *Primate origins and evolution : a phylogenetic reconstruction* (Princeton, N.J.: Princeton University Press) xiv, 804 p.

Martin, RD (1996), *Scaling of the Mammalian Brain: the Maternal Energy Hypothesis* (11) 149-56.

Mattison, J. A., et al. (2012), 'Impact of caloric restriction on health and survival in rhesus monkeys from the NIA study', *Nature,* 489 (7415), 318-21.

Mattson, M. P. and Magnus, T. (2006), 'Ageing and neuronal vulnerability', *Nat Rev Neurosci,* 7 (4), 278-94.

McKelvey, L., et al. (2012), 'The intracellular portion of GITR enhances NGF-promoted neurite growth through an inverse modulation of Erk and NF-kappaB signalling', *Biol Open,* 1 (10), 1016-23.

McLean, C. Y. (2011), 'Human-specific loss of regulatory DNA and the evolution of human-specific traits', *Nature,* 471, 216-19.

McNaught, K. S., et al. (2001), 'Failure of the ubiquitin-proteasome system in Parkinson's disease', *Nat Rev Neurosci,* 2 (8), 589-94.

Medina, Loreta and Reiner, Anton (2000), 'Do birds possess homologues of mammalian primary visual, somatosensory and motor cortices?', *Trends in Neurosciences,* 23 (1), 1-12.

Mercken, E. M., et al. (2013), 'Calorie restriction in humans inhibits the PI3K/AKT pathway and induces a younger transcription profile', *Aging Cell,* 12 (4), 645-51.

Meredith, R. W., et al. (2011), 'Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification', *Science,* 334 (6055), 521-4.

Miller, J. A., et al. (2014), 'Transcriptional landscape of the prenatal human brain', *Nature,* 508 (7495), 199-206.

Miller, V. M. and Best, P. J. (1980), 'Spatial Correlates of Hippocampal Unit-Activity Are Altered by Lesions of the Fornix and Entorhinal Cortex', *Brain Research,* 194 (2), 311-23.

Milner, A. C. and Walsh, S. A. (2009), 'Avian brain evolution: new data from Palaeogene birds (Lower Eocene) from England', *Zoological Journal of the Linnean Society,* 155 (1), 198-219.

Min, J. N., et al. (2008), 'CHIP deficiency decreases longevity, with accelerated aging phenotypes accompanied by altered protein quality control', *Molecular and Cellular Biology,* 28 (12), 4018-25.

Mink, J. W., Blumenschine, R. J., and Adams, D. B. (1981), 'Ratio of central nervous system to body metabolism in vertebrates: its constancy and functional basis', *Am J Physiol,* 241 (3), R203-12.

Molenberghs, P., Cunnington, R., and Mattingley, J. B. (2009), 'Is the mirror neuron system involved in imitation? A short review and meta-analysis', *Neuroscience and Biobehavioral Reviews,* 33 (7), 975-80.

Molnár, Z., Tavare, A., and Cheung, A. F. P. (2007), '3.02 - The Origin of Neocortex: Lessons from Comparative Embryology', in Jon H. Kaas (ed.), *Evolution of Nervous Systems* (Oxford: Academic Press), 13-26.

Montgomery, S. H. and Mundy, N. I. (2012a), 'Positive selection on NIN, a gene involved in neurogenesis, and primate brain evolution', *Genes Brain Behav,* 11 (8), 903-10.

--- (2012b), 'Evolution of ASPM is associated with both increases and decreases in brain size in primates', *Evolution,* 66 (3), 927-32.

--- (2014), 'Microcephaly genes evolved adaptively throughout the evolution of eutherian mammals', *BMC Evol Biol,* 14, 120.

Montgomery, S. H., et al. (2011), 'Adaptive evolution of four microcephaly genes and the evolution of brain size in anthropoid primates', *Mol Biol Evol,* 28 (1), 625-38.

Moran, L. B., et al. (2006), 'Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease', *Neurogenetics,* 7 (1), 1-11.

Morley, J. F. and Morimoto, R. I. (2004), 'Regulation of longevity in Caenorhabditis elegans by heat shock factor and molecular chaperones', *Mol Biol Cell,* 15 (2), 657-64.

Musich, PR and Zou, Y (2011), 'DNA-damage accumulation and replicative arrest in Hutchinson-Gilford progeria syndrome', *Biochem Soc Trans,* 39 (6), 1764-9.

Navarrete, A., van Schaik, C. P., and Isler, K. (2011), 'Energetics and the evolution of human brain size', *Nature,* 480 (7375), 91-3.

Nestor, A., Plaut, D. C., and Behrmann, M. (2011), 'Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis', *Proceedings of the National Academy of Sciences of the United States of America,* 108 (24), 9998-10003.

Nielsen, R., et al. (2007), 'Recent and ongoing selection in the human genome', *Nat Rev Genet,* 8 (11), 857-68.

Niimura, Y., Matsui, A., and Touhara, K. (2014), 'Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals', *Genome Res,* 24 (9), 1485-96.

Noback, C.R., et al. (2005), *The Human Nervous System: Structure and Function* (Humana Press).

Nolan, A. M., Nolan, Y. M., and O'Keeffe, G. W. (2011), 'IL-1beta inhibits axonal growth of developing sympathetic neurons', *Mol Cell Neurosci,* 48 (2), 142-50.

Nomura, T., Gotoh, H., and Ono, K. (2013), 'Changes in the regulation of cortical neurogenesis contribute to encephalization during amniote brain evolution', *Nat Commun,* 4, 2206.

Northcutt, R. G. and Kaas, J. H. (1995), 'The emergence and evolution of mammalian neocortex', *Trends Neurosci,* 18 (9), 373-9.

O'Connor, BP , et al. (2004), 'BCMA is essential for the survival of long-lived bone marrow plasma cells', *J Exp Med,* 199 (1), 91-8.

O'Keeffe, G. W., et al. (2008), 'NGF-promoted axon growth and target innervation requires GITRL-GITR signaling', *Nat Neurosci,* 11 (2), 135-42.

O'Leary, D. D., Chou, S. J., and Sahara, S. (2007), 'Area patterning of the mammalian cortex', *Neuron,* 56 (2), 252-69.

O'Leary, M. A., et al. (2013), 'The placental mammal ancestor and the post-K-Pg radiation of placentals', *Science,* 339 (6120), 662-7.

Obayashi, T. and Kinoshita, K. (2011), 'COXPRESdb: a database to compare gene coexpression in seven model animals', *Nucleic Acids Res,* 39 (Database issue), D1016-22.

Oldham, M. C., Horvath, S., and Geschwind, D. H. (2006), 'Conservation and evolution of gene coexpression networks in human and chimpanzee brains', *Proc Natl Acad Sci U S A,* 103 (47), 17973-8.

Oldham, M. C., et al. (2008), 'Functional organization of the transcriptome in human brain', *Nat Neurosci,* 11 (11), 1271-82.

Ozaki, K., et al. (2002), 'Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction', *Nat Genet,* 32 (4), 650-4.

Pagel, M. (1999), 'Inferring the historical patterns of biological evolution', *Nature,* 401 (6756), 877-84.

Pasquaretta, C., et al. (2014), 'Social networks in primates: smart and tolerant species have more efficient networks', *Sci Rep,* 4, 7600.

Passingham, R. E. and Wise, S. P. (2012), *The Neurobiology of the Prefrontal Cortex. Anatomy, Evolution, and the Origin of Insight* (Oxford: Oxford University Press).

Pavlicek, A. and Jurka, J. (2006), 'Positive selection on the nonhomologous end-joining factor Cernunnos-XLF in the human lineage', *Biol Direct,* 1, 15.

Pelvig, D. P., et al. (2008), 'Neocortical glial cell numbers in human brains', *Neurobiology of Aging,* 29 (11), 1754-62.

Perez-Enciso, M., Quevedo, J. R., and Bahamonde, A. (2007), 'Genetical genomics: use all data', *BMC Genomics,* 8, 69.

Perez, V. I., et al. (2009), 'Protein stability and resistance to oxidative stress are determinants of longevity in the longest-living rodent, the naked mole-rat', *Proc Natl Acad Sci U S A,* 106 (9), 3059-64.

Perl, S., et al. (2010), 'Significant human beta-cell turnover is limited to the first three decades of life as determined by in vivo thymidine analog incorporation and radiocarbon dating', *J Clin Endocrinol Metab,* 95 (10), E234-9.

Pfefferle, A. D. (2011), 'Comparative expression analysis of the phosphocreatine circuit in extant primates: implications for human brain evolution', *J. Hum. Evol.,* 60, 205-12.

Pitnick, S., Jones, K. E., and Wilkinson, G. S. (2006), 'Mating system and brain size in bats', *Proc Biol Sci,* 273 (1587), 719-24.

Platek, S. M., et al. (2008), 'Neural correlates of self-face recognition: An effect-location meta-analysis', *Brain Research,* 1232, 173-84.

Platek, Steven M., et al. (2004), 'Where am I? The neurological correlates of self and other', *Cognitive Brain Research,* 19 (2), 114-22.

Platt, A., Vilhjalmsson, B. J., and Nordborg, M. (2010), 'Conditions under which genome-wide association studies will be positively misleading', *Genetics*, 186 (3), 1045-52.

Plummer, T. (2004), 'Flaked stones and old bones: biological and cultural evolution at the dawn of technology', *Am J Phys Anthropol*, Suppl 39, 118-64.

Pollard, K. S. (2006), 'An RNA gene expressed during cortical development evolved rapidly in humans', *Nature*, 443, 167-72.

Pond, Caroline M. (1998), *The fats of life* (Cambridge ; New York: Cambridge University Press) 337 p.

Ponting, Chris P. (2008), 'The functional repertoires of metazoan genomes', *Nat Rev Genet*, 9 (9), 689-98.

Poucet, B., et al. (2003), 'Place cells, neocortex and spatial navigation: a short review', *Journal of Physiology-Paris*, 97 (4-6), 537-46.

Prabhakar, S., et al. (2008), 'Human-specific gain of function in a developmental enhancer', *Science*, 321 (5894), 1346-50.

Prins, P., Smant, G., and Jansen, R. C. (2012), 'Genetical genomics for evolutionary studies', *Methods Mol Biol*, 856, 469-85.

Pritchard, J. K., Pickrell, J. K., and Coop, G. (2010), 'The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation', *Curr Biol*, 20 (4), R208-15.

Rakic, P. (2009), 'Evolution of the neocortex: a perspective from developmental biology', *Nature Reviews Neuroscience*, 10 (10), 724-35.

Reader, S. M. and Laland, K. N. (2002), 'Social intelligence, innovation, and enhanced brain size in primates', *Proc Natl Acad Sci U S A*, 99 (7), 4436-41.

Reader, Simon M., Hager, Yfke, and Laland, Kevin N. (2011), *The evolution of primate general and cultural intelligence* (366) 1017-27.

Reilly, Steven K., et al. (2015), 'Evolutionary changes in promoter and enhancer activity during human corticogenesis', *Science*, 347 (6226), 1155-59.

Ricklefs, R. E. (2004), 'The cognitive face of avian life histories - The 2003 Margaret Morse Nice Lecture', *Wilson Bulletin*, 116 (2), 119-33.

Rimol, L. M. (2010), 'Sex-dependent association of common variants of microcephaly genes with brain structure', *Proc. Natl Acad. Sci. USA*, 107, 384-88.

Rizzolatti, G., et al. (1996), 'Premotor cortex and the recognition of motor actions', *Brain Res Cogn Brain Res*, 3 (2), 131-41.

Rockman, M. V., et al. (2005), 'Ancient and recent positive selection transformed opioid cis-regulation in humans', *PLoS Biol*, 3 (12), e387.

Romero-Granados, R., et al. (2011), 'Postnatal proteasome inhibition induces neurodegeneration and cognitive deficiencies in adult mice: a new model of neurodevelopment syndrome', *PLoS One,* 6 (12), e28927.

Rossi, D. and Zlotnik, A. (2000), 'The biology of chemokines and their receptors', *Annual Review of Immunology,* 18, 217-43.

Roth, G. and Dicke, U. (2005), 'Evolution of the brain and intelligence', *Trends Cogn Sci,* 9 (5), 250-7.

Rubin, G. M., et al. (2000), 'Comparative genomics of the eukaryotes', *Science,* 287 (5461), 2204-15.

Sacher, G. A. and Staffeldt, E. F. (1974), 'Relation of Gestation Time to Brain Weight for Placental Mammals: Implications for the Theory of Vertebrate Growth', *The American Naturalist,* 108 (963), 593-615.

Saenz, A., et al. (2008), 'Gene expression profiling in limb-girdle muscular dystrophy 2A', *PLoS One,* 3 (11), e3750.

Saigoh, K., et al. (1999), 'Intragenic deletion in the gene encoding ubiquitin carboxy-terminal hydrolase in gad mice', *Nat Genet,* 23 (1), 47-51.

Sakurai, H. and Ota, A. (2011), 'Regulation of chaperone gene expression by heat shock transcription factor in Saccharomyces cerevisiae: Importance in normal cell growth, stress resistance, and longevity', *Febs Letters,* 585 (17), 2744-48.

Saris, C. G., et al. (2009), 'Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients', *BMC Genomics,* 10, 405.

Saus, E., et al. (2010), 'Comprehensive copy number variant (CNV) analysis of neuronal pathways genes in psychiatric disorders identifies rare variants within patients', *J Psychiatr Res,* 44 (14), 971-8.

Sawaguchi, Toshiyuki and Kudo, Hiroko (1990), 'Neocortical development and social structure in primates', *Primates,* 31 (2), 283-89.

Schuck-Paim, C., Alonso, W. J., and Ottoni, E. B. (2008), 'Cognition in an ever-changing world: Climatic variability is associated with brain size in neotropical parrots', *Brain Behavior and Evolution,* 71 (3), 200-15.

Schulte-Ruther, M., et al. (2007), 'Mirror neuron and theory of mind mechanisms involved in face-to-face interactions: A functional magnetic resonance imaging approach to empathy', *Journal of Cognitive Neuroscience,* 19 (8), 1354-72.

Sereno, M. I. and Huang, R. S. (2006), 'A human parietal face area contains aligned head-centered visual and tactile maps', *Nat Neurosci,* 9 (10), 1337-43.

Shaw, P., et al. (2008), 'Neurodevelopmental trajectories of the human cerebral cortex', *J Neurosci,* 28 (14), 3586-94.

Shi, P., Bakewell, M. A., and Zhang, J. (2006), 'Did brain-specific genes evolve faster in humans than in chimpanzees?', *Trends Genet,* 22 (11), 608-13.

Sholtis, S. J. and Noonan, J. P. (2010), 'Gene regulation and the origins of human biological uniqueness', *Trends Genet,* 26 (3), 110-8.

Shultz, S. and Dunbar, R. I. (2007), 'The evolution of the social brain: anthropoid primates contrast with other vertebrates', *Proc Biol Sci,* 274 (1624), 2429-36.

Shultz, S., et al. (2004), 'A community-level evaluation of the impact of prey behavioural and ecological characteristics on predator diet composition', *Proc Biol Sci,* 271 (1540), 725-32.

Shultz, S., et al. (2005), 'Brain size and resource specialization predict long-term population trends in British birds', *Proceedings of the Royal Society B-Biological Sciences,* 272 (1578), 2305-11.

Silva, A. S., et al. (2011), 'Gathering insights on disease etiology from gene expression profiles of healthy tissues', *Bioinformatics*, 27 (23), 3300-5.

Smyth, G. K. (2005), 'Limma: linear models for microarray data', in R. Gentleman, et al. (eds.), *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (Springer, New York), 397-420.

Snodgrass, J. Josh, Leonard, WilliamR, and Robertson, MarciaL (2009), 'The Energetics of Encephalization in Early Hominids', in Jean-Jacques Hublin and MichaelP Richards (eds.), *The Evolution of Hominin Diets* (Vertebrate Paleobiology and Paleoanthropology: Springer Netherlands), 15-29.

Sol, D. (2009a), 'The cognitive-buffer hypothesis for the evolution of large brains', in Reuven Dukas and John M. Ratcliffe (eds.), *Cognitive ecology II* (Chicago: University of Chicago Press), 372 p.

--- (2009b), 'Revisiting the cognitive buffer hypothesis for the evolution of large brains', *Biol Lett,* 5 (1), 130-3.

Sol, D. and Price, T. D. (2008), 'Brain size and the diversification of body size in birds', *Am Nat,* 172 (2), 170-7.

Sol, D., Timmermans, S., and Lefebvre, L. (2002), 'Behavioural flexibility and invasion success in birds', *Animal Behaviour,* 63, 495-502.

Sol, D., et al. (2007), 'Big-brained birds survive better in nature', *Proc Biol Sci,* 274 (1611), 763-9.

Sol, D., et al. (2008), 'Brain size predicts the success of mammal species introduced into novel environments', *American Naturalist*, 172, S63-S71.

Sol, D., et al. (2005), 'Big brains, enhanced cognition, and response of birds to novel environments', *Proc Natl Acad Sci U S A*, 102 (15), 5460-5.

Soshnikova, N., et al. (2013), 'Duplications of hox gene clusters and the emergence of vertebrates', *Dev Biol*, 378 (2), 194-9.

Spalding, K. L., et al. (2005), 'Retrospective birth dating of cells in humans', *Cell*, 122 (1), 133-43.

Spalding, Kirsty L., et al. (2008), 'Dynamics of fat cell turnover in humans', *Nature*, 453 (7196), 783-87.

Spaulding, Shannon (2013), 'Mirror Neurons and Social Cognition', *Mind & Language*, 28 (2), 233-57.

Stedman, H. H., et al. (2004), 'Myosin gene mutation correlates with anatomical changes in the human lineage', *Nature*, 428 (6981), 415-18.

Stelzer, G., et al. (2011), 'In-silico human genomics with GeneCards', *Hum Genomics*, 5 (6), 709-17.

Sterner, K. N., et al. (2012), 'Dynamic gene expression in the human cerebral cortex distinguishes children from adults', *PLoS One*, 7 (5), e37714.

Su, A. I., et al. (2004), 'A gene atlas of the mouse and human protein-encoding transcriptomes', *Proc Natl Acad Sci U S A*, 101 (16), 6062-7.

Sugiura, M., et al. (2005), 'Cortical mechanisms of visual self-recognition', *Neuroimage*, 24 (1), 143-49.

Sullivan, P. G., et al. (2004), 'Proteasome inhibition alters neural mitochondrial homeostasis and mitochondria turnover', *J Biol Chem*, 279 (20), 20699-707.

Tacutu, R., et al. (2013), 'Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing', *Nucleic Acids Res*, 41 (Database issue), D1027-33.

Terman, A., et al. (2010), 'Mitochondrial Turnover and Aging of Long-Lived Postmitotic Cells: The Mitochondrial-Lysosomal Axis Theory of Aging', *Antioxidants & Redox Signaling*, 12 (4), 503-35.

Torkamani, A., et al. (2010), 'Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia', *Genome Res*, 20 (4), 403-12.

Tran, P. B. and Miller, R. J. (2003), 'Chemokine receptors: signposts to brain development and disease', *Nat Rev Neurosci,* 4 (6), 444-55.

Tsuboi, M., et al. (2015), 'Comparative support for the expensive tissue hypothesis: Big brains are correlated with smaller gut and greater parental investment in Lake Tanganyika cichlids', *Evolution,* 69 (1), 190-200.

Usadel, B., et al. (2009), 'Co-expression tools for plant biology: opportunities for hypothesis generation and caveats', *Plant Cell Environ,* 32 (12), 1633-51.

van der Gaag, C., Minderaa, R. B., and Keysers, C. (2007), 'Facial expressions: What the mirror neuron system can and cannot tell us', *Social Neuroscience,* 2 (3-4), 179-222.

van Schaik, C. P. and Burkart, J. M. (2011), 'Social learning and evolution: the cultural intelligence hypothesis', *Philos Trans R Soc Lond B Biol Sci,* 366 (1567), 1008-16.

van Woerden, J. T., van Schaik, C. P., and Isler, K. (2010), 'Effects of seasonality on brain size evolution: evidence from strepsirrhine primates', *Am Nat,* 176 (6), 758-67.

van Woerden, J. T., et al. (2012), 'Large brains buffer energetic effects of seasonal habitats in catarrhine primates', *Evolution,* 66 (1), 191-9.

Verginelli, F., et al. (2009), 'Nutrigenetics in the light of human evolution', *J Nutrigenet Nutrigenomics,* 2 (2), 91-102.

Vina, J. and Borras, C. (2010), 'Women Live Longer than Men: Understanding Molecular Mechanisms Offers Opportunities to Intervene by Using Estrogenic Compounds', *Antioxidants & Redox Signaling,* 13 (3), 269-78.

Vina, J., et al. (2005), 'Why females live longer than males? Importance of the upregulation of longevity-associated genes by oestrogenic compounds', *Febs Letters,* 579 (12), 2541-45.

Visel, A., Bristow, J., and Pennacchio, L. A. (2007), 'Enhancer identification through comparative genomics', *Semin Cell Dev Biol,* 18 (1), 140-52.

Visscher, P. M., et al. (2012), 'Five years of GWAS discovery', *Am J Hum Genet,* 90 (1), 7-24.

Vite, C. H. and Head, E. (2014), 'Aging in the canine and feline brain', *Vet Clin North Am Small Anim Pract,* 44 (6), 1113-29.

Wang, X. (2011), 'Expression of Siglec-11 by human and chimpanzee ovarian stromal cells, with uniquely human ligands: implications for human ovarian physiology and pathology', *Glycobiology,* 21, 1038-48.

Wang, X., et al. (2009), 'Asymmetric centrosome inheritance maintains neural progenitors in the neocortex', *Nature,* 461 (7266), 947-55.

Wang, Y. Q., et al. (2005), 'Accelerated evolution of the pituitary adenylate cyclase-activating polypeptide precursor gene during human origin', *Genetics,* 170 (2), 801-6.

Ward, L. D. and Kellis, M. (2012), 'Interpreting noncoding genetic variation in complex traits and human disease', *Nat Biotechnol,* 30 (11), 1095-106.

Weinstein, G. D., McCullough, J. L., and Ross, P. (1984), 'Cell proliferation in normal epidermis', *J Invest Dermatol,* 82 (6), 623-8.

Weisbecker, V. and Goswami, A. (2010), 'Brain size, life history, and metabolism at the marsupial/placental dichotomy', *Proc Natl Acad Sci U S A,* 107 (37), 16216-21.

Wells, J. C. (2006), 'The evolution of human fatness and susceptibility to obesity: an ethological approach', *Biol Rev Camb Philos Soc,* 81 (2), 183-205.

Whitehouse, R. C., et al. (1982), 'Zinc in Plasma, Neutrophils, Lymphocytes, and Erythrocytes as Determined by Flameless Atomic-Absorption Spectrophotometry', *Clinical Chemistry,* 28 (3), 475-80.

Wicker, B., et al. (2003), 'Both of us disgusted in My Insula: The common neural basis of seeing and feeling disgust', *Neuron,* 40 (3), 655-64.

Wirz, K. (1950), 'Studien über die cerebralisation: Zur quantitativen bestimmung der rangordnung bei säugetieren', *Acta Anat,* 9 (1-2), 134-96.

Woods, C. G., Bond, J., and Enard, W. (2005), 'Autosomal recessive primary microcephaly (MCPH): a review of clinical, molecular, and evolutionary findings', *Am J Hum Genet,* 76 (5), 717-28.

Wrangham, R. W. (2009), *Catching fire : how cooking made us human* (New York: Basic Books) v, 309 p.

Yeaman, S. (2013), 'Genomic rearrangements and the evolution of clusters of locally adaptive loci', *Proc Natl Acad Sci U S A,* 110 (19), E1743-51.

Zhang, B. and Horvath, S. (2005), 'A general framework for weighted gene co-expression network analysis', *Stat Appl Genet Mol Biol,* 4, Article17.

Zhang, J., et al. (2012), 'Weighted frequent gene co-expression network mining to identify genes involved in genome stability', *PLoS Comput Biol,* 8 (8), e1002656.

Zhao, H., Michaelis, M. L., and Blagg, B. S. (2012), 'Hsp90 modulation for the treatment of Alzheimer's disease', *Adv Pharmacol,* 64, 1-25.

7. Appendices

Genes That Escape X-Inactivation in Humans Have High Intraspecific Variability in Expression, Are Associated with Mental Impairment but Are Not Slow Evolving

Yuchao Zhang,[1,2] Atahualpa Castillo-Morales,[3] Min Jiang,[1] Yufei Zhu,[1] Landian Hu,[1] Araxi O. Urrutia,[3] Xiangyin Kong,*[1] and Laurence D. Hurst*[3]

[1]State Key Laboratory of Medical Genomics, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences and Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, People's Republic of China

[2]Graduate School of the Chinese Academy of Sciences, Beijing, People's Republic of China

[3]Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

***Corresponding author:** E-mail: bssldh@bath.ac.uk; xykong@sibs.ac.cn.

Associate editor: Naoko Takezaki

Abstract

In female mammals most X-linked genes are subject to X-inactivation. However, in humans some X-linked genes escape silencing, these escapees being candidates for the phenotypic aberrations seen in polyX karyotypes. These escape genes have been reported to be under stronger purifying selection than other X-linked genes. Although it is known that escape from X-inactivation is much more common in humans than in mice, systematic assays of escape in humans have to date employed only interspecies somatic cell hybrids. Here we provide the first systematic next-generation sequencing analysis of escape in a human cell line. We analyzed RNA and genotype sequencing data obtained from B lymphocyte cell lines derived from Europeans (CEU) and Yorubans (YRI). By replicated detection of heterozygosis in the transcriptome, we identified 114 escaping genes, including 76 not previously known to be escapees. The newly described escape genes cluster on the X chromosome in the same chromosomal regions as the previously known escapees. There is an excess of escaping genes associated with mental retardation, consistent with this being a common phenotype of polyX phenotypes. We find both differences between populations and between individuals in the propensity to escape. Indeed, we provide the first evidence for there being both hyper- and hypo-escapee females in the human population, consistent with the highly variable phenotypic presentation of polyX karyotypes. Considering also prior data, we reclassify genes as being always, never, and sometimes escape genes. We fail to replicate the prior claim that genes that escape X-inactivation are under stronger purifying selection than others.

Key words: **X-inactivation, rate of evolution, expression evolution.**

Article

Introduction

Mammals have evolved a mechanism to inactivate one of the female X chromosomes. Although in humans the majority of X-linked genes are subject to X-inactivation, at least 15% (Carrel and Willard 2005) are thought to escape X-inactivation being expressed from both the active X (Xa) and inactive X (Xi) chromosomes. Escape genes in human are distributed in clusters (Tsuchiya et al. 2004; Carrel and Willard 2005) and probably controlled at the chromatin domain level. The majority of escape genes have been shown to be located on the short arm of the X chromosome (Disteche 1999). This may reflect a mechanistic constraint, these genes being too distant from the X-inactivation center (*Xic*) in the long arm to be affected. They may also be protected from the spreading of *XIST* RNA, coded for by the *XIST* gene within the *Xic*, by centromeric heterochromatin.

Given the strong conservation of gene content on the mammalian X chromosome, it has been possible to ask whether the ability to escape X-inactivation might be an evolvable trait. Principally, this has been addressed by comparing mice and humans (Disteche et al. 2002; Carrel and Willard 2005; Yang et al. 2010). For example, Yang et al. (2010) used RNA sequencing technology, in combination with single nucleotide polymorphism (SNP) identification, to infer the escape profile in mice and compared this with human data. The profiles of escape in mice and humans show significant differences in the number of genes and overall status of inactivation with escape being more prevalent in humans for reasons unknown.

It is likely that this prevalence of escape from X-inactivation in humans is related to the relative severity of polyX karyotypes in humans (Yang et al. 2010). PolyX karyotypes are associated with numerous phenotypes, including mental retardation and growth effects (Rooman et al. 2002). Typically, when more than one X is present, all X chromosomes but one are inactivated (Lyon 1961; Belmont et al. 1986). Genes that escape X-inactivation are hence good candidates for dosage-mediated phenotypic disruptions associated with polyX karyotypes (Linden et al. 1995; Tartaglia et al. 2010; Berletch et al. 2011). Determining which genes escape X-inactivation is thus of potential clinical relevance.

Analysis of polyX karyotypes has also suggested that there is variability in phenotypic presentation between individuals with the same karyotype (Rooman et al. 2002; Otter et al. 2010; Tartaglia et al. 2010). Indeed although many XXX females go undiagnosed (Gustavson 1999; Tartaglia et al. 2010), many have immediately evident phenotypes (Otter et al. 2010). This may reflect differing degrees of mosaicism (Tartaglia et al. 2010). It might also, however, reflect variability between individuals as regards which genes escape X-inactivation. Consistent with this expectation, in humans genes that escape X-inactivation can have different expression levels in different individuals (Brown and Greally 2003; Carrel and Willard 2005), these variably expressed genes estimated to comprise 10% or more of X-linked genes.

In addition to clinical relevance, knowing which genes escape X-inactivation is important for molecular evolutionary inference, as genes escaping X-inactivation have different mean dominance to those not escaping and may be under different selective pressures (Park et al. 2010). Indeed, Park et al. (2010) report that genes that always escape X-inactivation have a lower K_a/K_s than those that sometimes do or never do. This, they suggest, may reflect differences in dominance. However, at first sight one might think that a dominance argument would make the opposite prediction: if most new mutations are recessive, as genes that never escape are haploid expressed, new mutations should be under stronger purifying selection than those diploid expressed (i.e., those that escape X-inactivation). Moreover, the class with the highest K_a/K_s are those that sometimes escape. A priori, all else being equal, one would expect this class to sit between the extremes of those that never and those that always escape. With these two caveats, it is worth asking whether the prior result is robust to reclassification of genes on addition of new data. In addition, it is necessary to address whether any result is robust to quantitative control for differences in absolute expression level (Pal et al. 2001; Drummond et al. 2006), the strongest predictor of rates of evolution.

The largest prior effort to determine the status of X-inactivation on human genes in a human cell line employed a quantitative assay based on fluorescent, single-nucleotide primer extension (Carrel and Willard 2005). This study examined a limited number ($N = 94$) of X-linked genes in fibroblasts, finding evidence for some form of escape for 35% of them, with 15% showing escape in all samples (Carrel and Willard 2005). Given the limited scale of this cell line-based assay, the same authors used a more systematic somatic cell hybrid system for more than 600 X-linked transcripts. This identified 94 transcripts that always escape inactivation and a further 61 that are heterogeneous.

Although the somatic cell hybrid data appear relatively consistent with the fibroblast data (Carrel and Willard 2005), it is worthwhile asking whether cell line-based data on a high-throughput scale can confirm or discover genes that escape X-inactivation. We address this issue by examining the RNA-Seq data of immortalized B-cells looking for evidence of heterozygosity within the transcriptome at X-linked loci. We identify a further 76 genes sometimes subject to some degree of escape from X-inactivation. With the same data we can also address the question of the level of heterogeneity. Are some individuals hyper-escapees, permitting significantly more genes to escape than others? Do populations differ in their profile of escape? To address these issues, we study the profile of escape between two populations, US residents with northern and western European ancestry (CEU) and Yoruban individuals of Nigeria (YRI). We find strong evidence for heterogeneity in escape, finding both between-population and between-individual differences. We find no evidence that genes that always escape X-inactivation have an unusually low rate of protein evolution, before or after control for expression level. These results potentially have ramifications for pharmacogenomics, for the etiology of X chromosome ploidy disruption phenotypes, and for molecular evolutionary inference.

Results

Identification of 76 New X-Inactivation Escapees

We located the biallelic sites in annotated genes, for which the transcript information was extracted from UCSC reference genes. Because the expression from the inactive X chromosome should be no higher than that of the active X chromosome, we considered the version of the gene with a smaller number of reads in heterozygosis to be the "silenced" allele from the inactive X chromosome and those with larger numbers as the active alleles. Assuming that incidences where fewer than 10% of alleles are from the silenced allele are not trustworthy to call heterozygosity (Carrel and Willard 2005), we obtained a total of 103 genes displaying evidence of escape from X-inactivation among 37 CEU individuals and 113 genes among 40 YRI individuals.

We consider only genes with replicate evidence as "validated" escapees. Replication means either two or more individuals or two or more SNPs within one individual, providing evidence of escape (table 1) (for the set of 33 genes with *prima facie* evidence of escape but without replication, see supplementary table S1, Supplementary Material online). Allowing for overlap between the methods for replications, we find that we can replicate 38 of the previously reported escape genes based on the rodent/human somatic cell hybrids assay and the primary human cell line assay (Carrel and Willard 2005). In addition, we observed a further 76 validated genes that escape inactivation in B lymphocyte cell lines from normal individuals (table 1), giving a total of 114 robustly described escape genes. Of the newly validated escape genes, 62 were reported not to be escapees in the prior analysis (rather than simply not studied). Of these, 19 were doubly replicated in our sample, both by escape being detected in multiple individuals and through multiple SNPs within one individual.

Considering instances where we could in principle have provided additional support for escape (i.e., we have polymorphic markers passing transcriptome level quality control), there are 23 genes at a minimum 7× coverage for which

Table 1. The 114 Escape Genes and the Nature of the Replication Evidence.

Genes	SNPs	Persons	Reported	Genes	SNPs	Persons	Reported	Genes	SNPs	Persons	Reported
ABCB7	Yes	Yes	Heter	*HAUS7*	No	Yes		*SEPT6*	Yes	Yes	Inactive
AIFM1	No	Yes	Escape	*HCFC1*	No	Yes	Heter	*SH3BGRL*	No	Yes	Escape
ALG13	No	Yes	Heter	*HDHD1*	Yes	Yes	Escape	*SH3KBP1*	Yes	Yes	Heter
APEX2	Yes	Yes	Inactive	*HUWE1*	No	Yes	Inactive	*SLC25A43*	Yes	Yes	
APOO	No	Yes	Inactive	*IDS*	Yes	Yes	Inactive	*SLC25A5*	No	Yes	Inactive
ARHGAP4	No	Yes		*IGBP1*	No	Yes	Inactive	*SLC38A5*	No	Yes	Inactive
ARMCX3	No	Yes	Inactive	*IRAK1*	Yes	Yes	Inactive	*SMC1A*	No	Yes	Heter
ATP6AP1	No	Yes	Inactive	*LAMP2*	Yes	Yes	Inactive	*SNX12*	No	Yes	Inactive
ATP6AP2	Yes	Yes	Inactive	*LOC550643*	No	Yes		*STS*	Yes	Yes	Escape
ATP7A	No	Yes	Heter	*MAGED1*	No	Yes		*SUV39H1*	No	Yes	Inactive
BCOR	No	Yes	Heter	*MAGED2*	Yes	Yes	Inactive	*SYN1*	No	Yes	Inactive
BTK	Yes	Yes	Heter	*MAGEH1*	No	Yes	Inactive	*TAZ*	No	Yes	
CCDC22	No	Yes	Inactive	*MAP7D2*	No	Yes	Inactive	*TBC1D25*	No	Yes	Inactive
CD99L2	Yes	Yes	Inactive	*MAP7D3*	Yes	Yes	Inactive	*TBL1X*	Yes	Yes	Heter
CDK16	No	Yes	Escape	*MBNL3*	No	Yes	Inactive	*TCEAL4*	No	Yes	Heter
CTPS2	No	Yes	Escape	*MED12*	No	Yes	Inactive	*TLR7*	No	Yes	
CXORF21	Yes	Yes		*MED14*	No	Yes	Escape	*TMEM187*	Yes	Yes	Heter
CXORF38	Yes	Yes	Escape	*MID1IP1*	No	Yes	Inactive	*TRAPPC2*	No	Yes	Escape
CXORF40A	No	Yes	Inactive	*MORF4L2*	Yes	Yes	Heter	*TSIX*	Yes	Yes	
CYBB	No	Yes		*MPP1*	No	Yes	Inactive	*TSR2*	No	Yes	Inactive
DDX26B	No	Yes	Inactive	*MSL3*	Yes	Yes	Heter	*TXLNG*	Yes	Yes	Escape
DDX3X	No	Yes	Escape	*MTMR1*	No	Yes	Inactive	*UBA1*	Yes	Yes	Escape
DKC1	No	Yes	Inactive	*NSDHL*	No	Yes	Inactive	*UBL4A*	Yes	Yes	Inactive
DMD	Yes	Yes	Inactive	*P2RY10*	No	Yes		*USP9X*	Yes	Yes	Escape
DNASE1L1	Yes	Yes	Inactive	*PDHA1*	Yes	Yes	Inactive	*UTP14A*	No	Yes	Heter
DOCK11	No	Yes	Heter	*PDK3*	Yes	No	Inactive	*VBP1*	Yes	Yes	Inactive
EBP	No	Yes	Inactive	*PGK1*	No	Yes	Inactive	*WWC3*	Yes	Yes	Inactive
EDA2R	No	Yes	Heter	*PIM2*	No	Yes	Inactive	*XIAP*	No	Yes	Inactive
EIF1AX	No	Yes	Escape	*PIN4*	No	Yes	Heter	*XIST*	No	Yes	Escape
EIF2S3	Yes	Yes	Escape	*PIR*	Yes	Yes	Escape	*ZC4H2*	No	Yes	Inactive
ELF4	Yes	Yes		*PJA1*	No	Yes	Inactive	*ZFX*	Yes	Yes	Escape
ELK1	No	Yes	Inactive	*PLXNA3*	Yes	No	Inactive	*ZMYM3*	No	Yes	Inactive
FAM3A	No	Yes	Inactive	*PQBP1*	No	Yes	Inactive	*ZNF275*	Yes	Yes	Inactive
FLNA	Yes	Yes	Inactive	*PRKX*	Yes	Yes	Heter	*ZNF75D*	Yes	No	Inactive
FTSJ1	No	Yes	Inactive	*RBM3*	Yes	Yes	Inactive				
G6PD	Yes	Yes	Inactive	*RENBP*	Yes	Yes	Heter				
GDI1	No	Yes		*RNF113A*	No	Yes	Inactive				
GEMIN8	No	Yes	Escape	*RPL10*	Yes	No	Inactive				
GPR174	No	Yes		*SASH3*	No	Yes	Inactive				
GRIPAP1	No	Yes	Inactive	*SAT1*	No	Yes	Inactive				

Note.—The SNP column indicates whether genes have multi-SNPs within one individual that all support the hypothesis of X-inactivation escape. The Persons column indicates whether genes have replication by being identified as escaping in multiple individuals. The Reported column indicates the reported state in previously reported rodent/human somatic cells (Carrel and Willard 2005). Escape genes are those that escape X-inactivation in all females tested; Heter are heterogeneous genes, i.e., genes that exhibit XCI in some, but not all, females assayed. For the cases with "No" in persons column, all of them are able to attempt verification. So, here No indicates that these cases are potentially able to be replicated but actually not supported.

prior evidence (Carrel and Willard 2005) suggested escape from inactivation to some degree that we could not confirm (here we include our 33 nonreplicated escapees as providing support). Even if we permit a minimum of 20× coverage to consider a gene, we still find ten that we fail to replicate. As coverage increases, so decreasing false-negative calls of haploid expression, there is an approximately constant ratio of the number of genes whose escape we can confirm to the number we cannot confirm (with 7× coverage, the ratio of the number of those we cannot replicate to the number we can replicate is 0.47, whereas at 50× it is 0.45).

Of our 114 escapees, there are 110 incidences where two or more individuals across the whole sample show evidence that a given gene escapes X-inactivation. There are 60 that escape in at least two different individuals within the CEU population and 80 genes that escape in at least two different individuals within the YRI population (fig. 1), a total of 103 different genes with within-population replication. There are

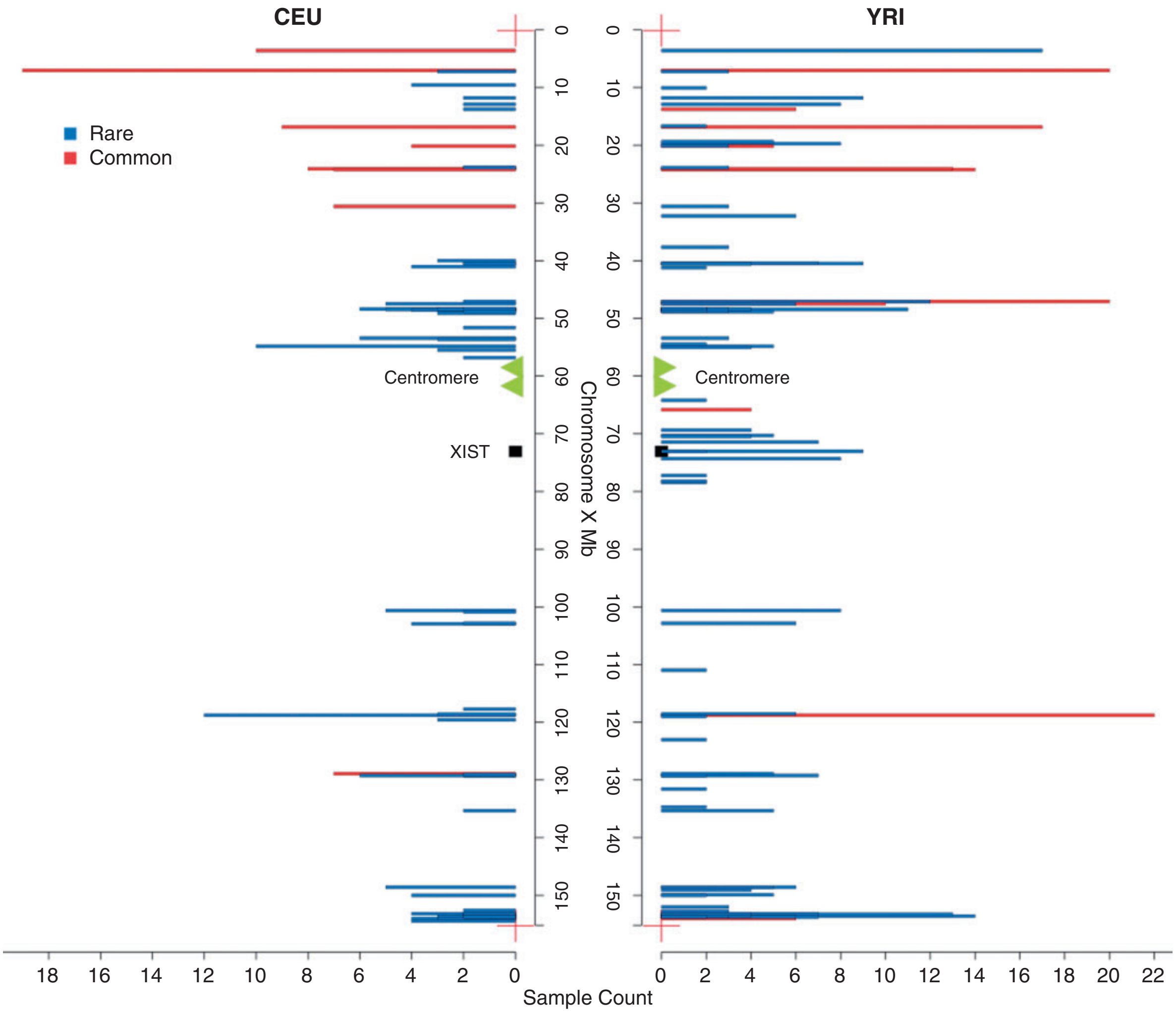

Fig. 1. The location of escape genes in CEU and YRI cluster in similar chromosomal locations. The genes found in more than three individuals and in greater than 50% of the potentially informative samples are considered to be common escape genes (red) in each population, whereas the others are rare escape genes (blue) in the populations. The genes solely replicated via more than one SNPs per gene are not included. Their inclusion makes no difference to qualitative trends. The *x* axis refers to the count of individuals with evidence for escape in the corresponding genes. Note *Xist* is within the *Xic* domain.

in total 45 genes confirmed heterozygous with multiple SNPs (see table 1) of which 27 were previously not known to escape X-inactivation and 41 of which are confirmed by between-individual replication as well. A few (three in CEU and four in YRI) of the replicated 114 genes contain heterozygosity with most, but not all, of the sites being consistent across several individuals. We considered these as replicated as 1) most sites were consistent and 2) as noted above, a lack of evidence for consistency of heterozygosity within a gene is not unexpected as it could reflect either SNPs having different coverage (in some individuals some potentially heterozygous sites could then not be interrogated) or owing to some falling below the 10% threshold that we set, in which case they would still be called homozygous.

Genes Newly Identified to Escaping X-Inactivation Cluster in Known Domains of Escape

The previously described escapees tend to be distant from the X-inactivation centre. The same is seen with the new inventory of escapees, in which we also see that escapees defined within each population locate to the same regions of the X chromosome (fig. 1), with the majority of escape genes being located in the short arm and the distal portion of the long arm of the X chromosome (region of PAR2). This is consistent with the previously reported control of chromatin domains in human X-inactivation (Tsuchiya et al. 2004; Carrel and Willard 2005; Yang et al. 2010) and with the related claim (Lahn and Page 1999) that escape genes are more common in the relatively recently added strata of the X chromosome (strata 3–5) compared with the more ancient strata (S1 and S2) (χ^2 test, $P = 0.02$; strata data from Kelkar et al. [2009]). S1, the most ancient stratum, dominates the long arm of the X and has the lowest proportion of genes escaping X-inactivation. For a frequency plot of escapees by strata, see supplementary figure S1, Supplementary Material online. The difference in escape regularity between the short and long arms of the X chromosome is not obviously explained as an artifact of expression level, read coverage, or density of heterozygous sites, as there is no significant difference between genes on two arms of the X, neither in expression level (Mann–Whitney U test, $P = 0.3779$), coverage at polymorphic sites (Mann–Whitney U test, $P = 0.120$), nor in density of heterozygous sites (229/391 in short arm vs. 363/619 in long arm).

There may also be a small cluster of genes escaping inactivation in the immediate vicinity of *Xic* (fig. 1). This cluster is visible only in the YRI population but this population has a greater extent of DNA heterozygosity, making it easier to identify escape genes. If this cluster is real, it is associated with few genes and, controlling for the degree of heterozygosis, there is no significant difference between the two populations. However, a similar cluster of escape was observed in prior analysis (Carrel and Willard 2005). The possibility that *Xist* may have weaker effects in the immediate vicinity of the X-inactivation center *Xic* (chrX: 65,000,000–80,000,000) (fig. 1) is, we suggest, worthy of deeper scrutiny. As regard the issues of sites in the vicinity of *Xic*, there is no significant expression difference between genes closer to *Xic* and those on the short arm (Mann–Whitney U test, $P = 0.6558$). Read coverage also shows no difference in this region (Mann–Whitney U test, $P = 0.676$).

Genes Escaping X-Inactivation Are Commonly Related to Mental Impairment

It is notable that many X chromosome ploidy alterations (including XXY and XXX, XXXX, XXXXX) are associated with learning impairments (Rooman et al. 2002). Indeed, this may be the only consistent feature of polyX karyotypes (Rooman et al. 2002). As typically all but one X is inactivated, the phenotype of X polysomies is often thought to reflect the action of genes that escape X-inactivation. Do we find any evidence that genes escaping inactivation are commonly associated with mental retardation?

We can define X-linked mental retardation (XLMR) or intellectual disability (ID) genes as those genes, mutation within which are associated with disturbance of normal intellectual functioning (Gecz et al. 2009; Stevenson and Schwartz 2009). A list of such genes is available from Greenwood Genetic Centre (Gecz et al. 2009). Among the 114 replicated escape genes, there are 22 genes (supplementary table S2, Supplementary Material online) involved in the diseases of XLMR or ID. There are 833 examinable genes covered with reads, including 91 XLMR/ID genes and 114 escapees. To determine whether 22 is significantly greater than expected, we randomly selected 114 of 833 and recorded how often the number of XLMR genes found is ≥22. The observed number is indeed more than expected by chance ($P = 0.0025$, from 10,000 simulants). These 22 genes would be good candidates for further analysis in this context, as impaired intellectual functioning may reflect higher dosage of these genes. There is considerable between-individual variation in the number of XLMR escape genes (supplementary fig. S2, Supplementary Material online). It would be instructive to know whether this variation correlates with any mental functioning parameters in XX females as well as polyX subjects.

A common, but not universal, phenotype of X polysomies is an effect on stature, typically manifested as rapid growth (Rooman et al. 2002). In no small part this is owing to overexpression of the pseudoautosomal gene *SHOX* (Rao et al. 1997). Linkage analysis has suggested, however, that Xq24 might also harbor such stature genes (Deng et al. 2002; Liu et al. 2004). This has been replicated in some (Liu et al. 2006) but not all (Visscher et al. 2007) studies. We find 6 of the 50 genes that reside within Xq24 escape X-inactivation (these being *LAMP2*, *SLC25A5*, *DOCK11*, *RNF113A*, *SEPT6*, and *SLC25A43*). This is no more than expected by chance (randomization test, as above, $P > 0.05$). Text mining for any association with growth phenotypes (via http://diseases.jensenlab.org/Search) suggests no evident connections.

The Profile of Escape Differs between CEU and YRI

In this study, we find a total of 66 genes that escape X-inactivation in both the CEU and YRI populations, including several well-known escape genes (*HDHD1*, *STS*, *ZFX*, *EIF2S3*, *CXorf38*, *DDX3X*). However, we are especially interested in the differences between populations rather than the common escape genes of the two populations. Table 2 presents all of the replicated escapee genes that are genetically polymorphic in both the populations and hence potentially identifiable as escapees, as well as the escape status of these genes.

We address whether there are differences between the populations by a randomization test (see Materials and Methods). The answer to the question as to whether the difference in the profiles is due to chance is unambiguous: the two groups of populations are considerably different (observed $\chi^2 = 196.56$, expected $= 119.94 \pm 16.07$ [SD]; from randomization $P < 0.0001$). This is unlikely to be an artifact of coverage differences between the two populations as the coverage is not significantly different between the two (Mann–Whitney U test, $P = 0.37$). Moreover, if we exclude from analysis any heterozygous sites with a less than 10× coverage, so giving more confidence in calling a lack of escape but also identifying fewer escaping genes, we still observe a significant difference between the populations (randomization test described in Materials and Methods, $P = 0.016$). P value is reduced not least because the sample size is reduced. At this cutoff, 82 genes in CEU and 90 genes in YRI are retained as escapees. Using only single-end data rather than single-end and paired-end data also makes no difference to the conclusion of between-population differences (randomization test described in Materials and Methods, $P = 0.006$). A difference in the profile of escape can be both because the genes escaping in the two populations are different and because the proportions of individuals showing escape for a given gene are different.

Analysis of Which Genes Are Variable in Their Propensity for Escape Is of Low Power

The analysis considering all genes *en mass* demonstrates striking variation between the populations. But can we identify which genes are different between the two populations?

Table 2. Escape Genes and Their Proportion of Escape among Individuals in CEU and YRI.

Population	Proportion of Individuals Showing Evidence of Escape				
	<20%	20–40%	40–60%	60–80%	>80%
Escape uniquely in CEU	*CYBB, TBC1D25, PQBP1, EDA2R, PJA1, MED12, PGK1, P2RY10, TCEAL4, SLC25A43, XIAP, AIFM1, MAP7D3, IDS, MTMR1, CD99L2, IRAK1, FLNA*	*TLR7, SAT1, APOO, MID1IP1, CXorf38, DDX3X, UBA1, SYN1, EBP, RBM3, SUV39H1, GRIPAP1, CCDC22, MAGED1, PIN4, SH3BGRL, BTK, LAMP2, UTP14A, ELF4, NSDHL, HCFC1, PLXNA3, UBL4A, FAM3A, DKC1, VBP1*	*PRKX, STS, MSL3, TRAPPC2, GEMIN8, TXLNG, EIF2S3, BCOR, ATP6AP2, SLC38A5, SMC1A, MAGED2, ZMYM3, SEPT6, SASH3, DNASE1L1*	*HDHD1, EIF1AX, ZFX, CXorf21, ATP6AP1*	—
Escape uniquely in YRI	*GEMIN8, SAT1, CYBB, BCOR, DDX3X, SLC38A5, PQBP1, CCDC22, MAGED1, SMC1A, MAGED2, MED12, PGK1, P2RY10, SH3BGRL, LAMP2, XIAP, SASH3, UTP14A, AIFM1, IDS, CD99L2, NSDHL, ATP6AP1, VBP1*	*STS, MSL3, TLR7, APOO, CXorf21, ATP6AP2, EBP, TBC1D25, SUV39H1, GRIPAP1, PJA1, ZMYM3, BTK, TCEAL4, SLC25A43, ELF4, MAP7D3, MTMR1, HCFC1, FLNA, DNASE1L1, PLXNA3, UBL4A, FAM3A*	*PRKX, TXLNG, EIF1AX, MID1IP1, CXorf38, UBA1, SYN1, RBM3, EDA2R, PIN4, SEPT6, IRAK1, DKC1*	*HDHD1, TRAPPC2, EIF2S3, ZFX*	—
Common	*AIFM1, CD99L2, CYBB, IDS, MED12, P2RY10, PGK1, PQBP1, XIAP*	*APOO, BTK, EBP, ELF4, FAM3A, GRIPAP1, HCFC1, PLXNA3, SUV39H1, TLR7, UBL4A*	*PRKX, SEPT6, TXLNG*	*HDHD1, ZFX, CXorf21, ATP6AP1*	—

Note.—The percentages indicate the proportion of informative individuals showing evidence of escape in the population (or populations) in which they escape in the population. Escaping genes that are common to both CEU and YRI within each range are colored in red.

Regarding the individual genes, we identified a number of cases with significant differences between the two groups (table 3 and supplementary fig. S3, Supplementary Material online). Owing to the different numbers of informative individuals for each gene, the *P* values for each gene are not strictly comparable. Importantly, with low sample sizes (few heterozygous individuals), *P* can never be very low. As the sample size varies between genes, the usual consideration of how to estimate the true number of significant instances, by examination of the form of the distribution of $1 - P$ versus rank order (Lai 2007), is not valid. As such we consider those that are significant as candidates for genes showing differences between populations, but this would require experimental confirmation, not least because no incidence passes multi-test correction.

Despite the above caveats, there is one potentially notable observation. In all examples of a potential difference between the two populations (at $P < 0.05$), the prevalence of escape is higher in CEU than in YRI. This remains true if we consider also incidences where *P* lies between 0.05 and 0.1. This, however, is likely to be an artifact of higher diversity in YRI which leads to more potentially informative samples (heterozygous at the DNA level). In the samples where we detect a difference, the mean sample size in YRI is around 20 and around 10 in CEU. If we consider a case where 4/10 are escapees in CEU and 2/20 are escapees in YRI (around the average that we observe in the cases of significant difference) and compare this with the symmetrical case (1/10 in CEU and 8/20 in YRI), it is indeed the case that the χ^2 values are higher for the former case ($\chi^2 = 3.75$) than in the symmetrical case ($\chi^2 = 2.85$). Thus, with the sorts of sample sizes and the sorts of ratios of escape to non-escape that we are looking at, we might expect to see more significant examples when the higher proportion of escapees is seen in the population with the lower number of informative examples.

Although we cannot be confident in having identified genes that show between-population within-species differences, it is worth asking whether there might be any commonality of those that are potentially different. On the X chromosome, the six genes that have significant escape variation ($P < 0.05$) are not clustered together (supplementary fig. S3, Supplementary Material online). Some of their neighboring genes with escape from X-inactivation do not have an escape profile showing significant differences between the two populations. This result might suggest that the between-population divergence, in regard to X-inactivation escape, is not owing to chromatin domain regulation. It could also mean, however, that owing to statistical limitations, we have incorrectly classified genes as to whether they differ in the escape propensity between different populations.

Evidence for Between-Individual Differences

The above data suggest that the two populations differ in their propensity to permit escape from X-inactivation. But might there also be females that are more or less prone to permitting escape? To address this, we can ask how often an

Table 3. Genes That Potentially Show Differences between the Two Populations in Escape Profile.

Gene	CEU	YRI	*P* Value
USP9X	4/14	0/34	0.00729
ATP6AP1	3/5	1/23	0.02046
MPP1	3/8	0/20	0.02228
SASH3	7/13	5/35	0.02229
TBL1X	4/17	0/21	0.03858
HUWE1	3/16	0/29	0.04549
MORF4L2	4/13	0/14	0.05364
CXorf21	7/9	3/14	0.05575
LOC550643	2/9	0/28	0.05874
LAMP2	3/12	1/33	0.06052
SMC1A	6/13	3/24	0.07425
VBP1	4/12	1/19	0.07787
SLC38A5	2/4	2/26	0.08774
BCOR	3/6	1/13	0.09626
SLC25A5	3/12	0/14	0.09837
CA5BP1	1/2	0/18	0.09976

Note.—The fractions in the CEU and YRI columns indicate the proportion of individuals with the gene escaping X-inactivation. The numerator is the number of escape samples, and the denominator is the number of heterozygous individuals at the DNA level. The differences between CEU and YRI were compared, and the *P* values were calculated (here, only genes with $P < 0.1$ are shown, and those with $P < 0.05$ are shown above the line). *P* values are from the randomization test as described in Materials and Methods.

individual has escape genes at potentially informative genes. To this end we calculated how many genes show escape and how many potential informative genes that could be heterozygous, but not show transcriptome level heterozygosity, and compared each individual with the total of others by the χ^2-like test through simulation (see Materials and Methods). Of the 77 individuals, 5 in CEU and 8 in YRI show more escape than expected by chance, what we term hyper-escapee females ($P < 0.05$) (table 4). After Holm's correction, four in CEU and one in YRI remain as hyper-escapees. This is consistent with the notion that even within populations individuals differ in their propensity to allow genes to escape inactivation (Carrel and Willard 2005). If we look only at these 13 individuals (significant before Holm's correction), we still detect the significant differences between the two populations (from randomization, $P = 0.028$). As before, this can be both because the genes escaping in the two populations are different and because the proportions of individuals showing escape for a given gene are different. In addition, we find evidence for five and six hypo-escapee females, in CEU and YRI, respectively, but only one (in YRI) is significant after multi-test correction (Holm's correction). Taken together, these results suggest that there are both between-individual and between-population differences in the propensity to escape.

Although these results *prima facie* suggest that 6–17% of females are hyper-escapees and 1–14% are hypo-escapees, this analysis comes with a caveat. As the individuals differ as regards which genes are potentially informative (heterozygous at the DNA level) and genes differ as regards their propensity to escape inactivation, some of the between-individual heterogeneity may reflect differences in the set of informative genes rather than escape tendencies per se. However, if for each person we consider only those genes that are informative in other individuals, five and one incidences of hyper- and hypo-escape are still evident after Holm's correction.

No Evidence That Permanently Escaping Genes Evolve Slowly

It has been reported that genes that always escape X-inactivation are under stronger purifying selection than either those that sometimes escape and those that never escape (Park et al. 2010), this being reflected in significantly lower K_a/K_s values. This was interpreted as possibly being due to differences in dominance. However, the group with the highest K_a/K_s were those that sometimes escape. A priori, all else being equal, from dominance arguments one would expect this class to sit between the extremes of those never and always escaping. Moreover, if most mutations are recessive, we might have expected that genes that never escape should be the ones under the stronger purifying selection as they are haploid expressed. With our new compendium of genes with replicated evidence for escape from X-inactivation, we can add to the prior data set to define new groupings of genes to examine the robustness of the prior claim.

The new merged data set comprises 446 genes (supplementary table S3, Supplementary Material online). We find evidence for heterogeneity between the three classes in K_a/K_s (Kruskall–Wallis test: $P = 0.016$). However, unlike what was previously described, when comparing between the different classes, the only robust result is that the heterogeneous group has a higher K_a/K_s than either those that always escape or those that never escape (fig. 2). Eliminating any genes for which $K_a/K_s > 1$ does not affect these conclusions and if anything makes the results more robust (Kruskal–Wallis test, $P = 0.010$; *P* for comparison of heterogeneous to inactive = 0.013, comparing heterogenous to escape = 0.019, and escape to inactive = 0.37). We thus cannot replicate the prior result that those genes that always escape have unusually low K_a/K_s. Genes that are heterogenous in expression appear to have higher K_a/K_s ratios.

Our data set requiring a minimum 7× coverage can legitimately report a new incidence of escape but may have a false-negative problem, i.e., genes that really do escape are categorized as not escaping just because coverage at the relevant heterozygous sites was not high enough to detect the rarely expressed allele. In this context, we would have forced some genes into the "sometimes escape" class when they should be in the "always escape" class. However, considering genes that ever escape X-inactivation as a single class (the union of sometimes and always, for which there should be no classification issue), there is no evidence that these evolve any slower than those that never escape (Mann–Whitney *U* test, $P = 0.11$) with those escaping having the higher median rate ($K_a/K_s = 0.15$ for genes that never escape and 0.22 for those that always or sometimes escape). Moreover, if the slow evolution of genes that always escape is real, then by miscalling

Table 4. Females Differ in Their Propensity to Allow Genes to Escape Inactivation.

CEU				YRI			
ID	Escape	*P* Value	Holm's	ID	Escape	*P* Value	Holm's
NA06985	3:32	0.018	0.522	NA18499	13:9	0.254	1
NA07000	0:26	0.006	0.192	NA18502	6:16	0.327	1
NA07037	3:13	0.495	1	NA18505	41:6	9.9e-6	3.7e-4
NA07055	6:19	0.598	1	NA18508	15:8	0.105	1
NA07056	7:20	0.728	1	NA18511	10:17	0.669	1
NA07345	39:5	9.9e-6	3.7e-4	NA18517	13:32	0.158	1
NA07346	1:11	0.194	1	NA18520	3:12	0.231	1
NA11830	31:16	9.9e-6	3.7e-4	NA18523	3:8	0.509	1
NA11832	8:16	0.852	1	NA18852	18:6	0.016	0.512
NA11840	12:19	0.511	1	NA18855	22:15	0.123	1
NA11882	2:26	0.023	0.644	NA18858	19:9	0.052	1
NA11894	1:3	1	1	NA18861	3:28	0.005	0.185
NA11918	1:6	0.54	1	NA18870	7:23	0.113	1
NA11920	2:17	0.138	1	NA18909	27:12	0.013	0.442
NA11931	4:28	0.071	1	NA18912	22:10	0.025	0.775
NA11993	5:30	0.088	1	NA18916	4:15	0.15	1
NA11995	6:9	0.621	1	NA19093	19:21	0.712	1
NA12004	18:3	9.9e-6	3.7e-4	NA19099	12:22	0.522	1
NA12006	3:33	0.014	0.42	NA19102	2:17	0.029	0.812
NA12044	4:19	0.27	1	NA19108	23:6	0.003	0.114
NA12057	6:27	0.201	1	NA19114	11:10	0.618	1
NA12145	9:24	0.75	1	NA19116	23:11	0.032	0.864
NA12156	11:4	0.008	0.248	NA19127	6:22	0.106	1
NA12234	9:12	0.305	1	NA19131	9:24	0.183	1
NA12249	4:16	0.421	1	NA19137	10:12	0.869	1
NA12287	3:15	0.298	1	NA19140	12:16	1	1
NA12489	0:13	0.064	1	NA19143	17:16	0.499	1
NA12717	6:15	1	1	NA19147	7:21	0.19	1
NA12751	5:20	0.372	1	NA19152	6:33	0.008	0.288
NA12761	4:8	1	1	NA19159	2:19	0.027	0.81
NA12763	2:13	0.252	1	NA19172	7:23	0.12	1
NA12776	7:19	0.858	1	NA19190	5:20	0.093	1
NA12813	31:3	9.9e-6	3.7e-4	NA19193	7:14	0.611	1
NA12815	13:25	0.767	1	NA19201	1:31	0.001	0.039
NA12828	5:14	0.837	1	NA19204	3:24	0.011	0.385
NA12873	2:15	0.19	1	NA19209	7:25	0.073	1
NA12892	0:23	0.005	0.165	NA19222	22:10	0.028	0.812
				NA19225	14:13	0.551	1
				NA19238	19:12	0.121	1
				NA19257	22:8	0.013	0.442

Note.—In the escape column, there are two numbers *N*:*M*. *N* is the number of escape genes and *M* is the number of the other potentially informative genes that show no evidence of escape. Significance after Holm's correction is marked in red and blue, red for hyper-escape and blue for hypo-escape.

some genes as being haploid expressed when before they were considered to be escapees, we would have moved slow evolving genes from the always escape class into the sometimes escape class. This bias would make it less likely that we would have obtained the result that the sometimes escape class are the fastest evolving. Were the sample of genes that were reclassified the faster evolving genes within the always escape class (possibly because they are low coverage hence lowly expressed and fast evolving), then this should have acted to exaggerate the slow evolution of the always escape class.

Our analysis and the prior one have a potential major artifact problem. While Park et al. (2010) compared genes that appear to show dosage compensation and those that do not, there was no quantitative control for differences in absolute expression level, the strongest predictor of rates of evolution (Pal et al. 2001; Drummond et al. 2006). If we allow for this covariate, can we recover any differences between the three classes? To address this, we reconsidered the merged data and obtained expression data from Su et al. (2004) where available (see Materials and Methods). This resulted in a data set of 262 genes (supplementary table S3,

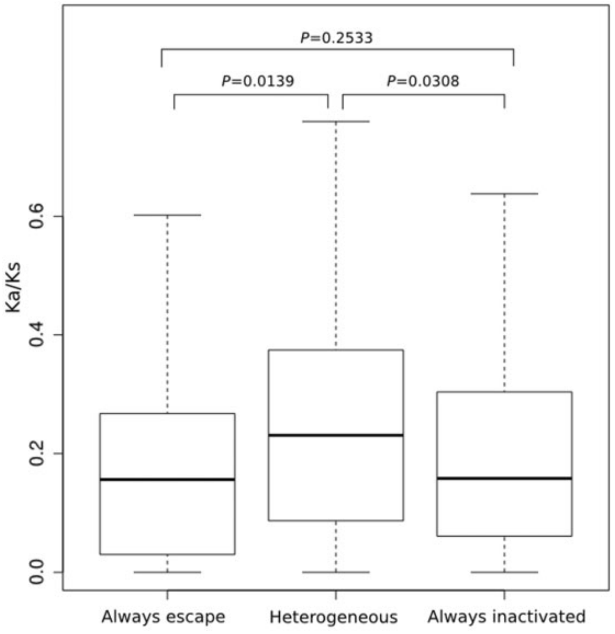

FIG. 2. K_a/K_s ratios of genes in the three X-inactivation classes in the merged data set. *P* values indicate significance on pairwise Mann–Whitney *U* tests. There are 35 that always escape, 206 always inactivated, and 205 heterogeneous ($N = 446$). Evolutionary rates are from the human–macaque orthologous genes with numbers taken from Ensembl or from Park et al. (2010). Outliners are not shown. Transverse lines indicate the median value.

Supplementary Material online). For each gene with expression data, we calculated the mean expression level across tissues. As expected, median expression rate is a predictor of K_a/K_s (Spearman correlation, $\rho = -0.15$, $P = 0.017$). However, the three gene classes show no evidence of differing in their median expression level (Kruskall Wallis test, $P = 0.57$). In the smaller sample set for which expression data are available, the Kruskall–Wallis statistic, comparing K_a/K_s values between genes belonging to different X-inactivation status classes, remains significant ($P = 0.03$). However, after control for expression level (by considering the residuals from the loess regression of K_a/K_s against log[expression level]), the Kuskall–Wallis test is marginally weaker and nonsignificant ($P = 0.071$). This result, however, is sensitive to the exclusion of genes with $K_a/K_s > 1$ ($P = 0.018$).

Discussion

Our analysis increases by approximately 50% the number of genes showing evidence of escaping X-inactivation in humans. These escapees cluster with others in the domains thought to be relatively protected from the spreading of *XIST*. Consistent with a common finding of mental impairment in polyX individuals, there is an excess of genes associated with mental impairment among the escapees. We also found evidence for between-individual and between-population differences in the propensity to permit escape. This is consistent with the observation that polyX karyotype bearers are highly heterogeneous in presentation (Rooman et al. 2002).

The true extent of variation in escape from X-inactivation is likely to be greater than that witnessed here. For example, while we examined one high-resolution high-quality data set from one cell lineage, variation between tissues/cells within an individual (Lopes et al. 2010; Berletch et al. 2011) may also be relevant. Assuming the variation to be real, it is not unexpected that we both find new candidates and fail to replicate a few prior instances (even though we had informative samples). Indeed, it is striking that we report 62 new examples of escape, where the prior effort had information but found no evidence of escape, and only 23 examples where we could not replicate escape.

Given the ability of RNA-Seq to falsely report haploid expression (DeVeale et al. 2012), false-negative calls of haploid expression must be considered an alternative explanation for our inability to replicate some instances of escape. Similarly, as false inference of haploid expression is increasing unlikely as coverage/expression level goes up, so too we might expect that genes with haploid expression might be skewed toward the low coverage end. Indeed, the coverage of genes whose escape we can replicate ($N = 38$) is higher than that of genes whose escape we could not replicate ($N = 23$) (Mann–Whitney *U* test, $P < 0.001$). Although consistent with some of the failure to replicate being an artifact of low coverage, the same result is consistent with lower expression level owing to haploid expression. Arguing against the latter is the evidence that the genes that appear to be haploid expressed are, when analyzed across multiple tissues, no different in median expression level than those presenting evidence of escape. Some of the inability to replicate prior evidence for escape appears relatively solid as many genes appear to be haploid expressed even with >50× coverage.

While RNA-Seq artifacts (DeVeale et al. 2012) are less likely to lead to false positives, can we be confident that we have not overinterpreted the data? Our method to infer escape from X-inactivation via heterozygosity could be misleading or detecting something other than escape from X-inactivation. We showed (see Materials and Methods) that mapping errors appear not to be a serious issue with very few cases of X-linked "heterozygosity" seen in males and few instances of there being more than three alleles detected in any given female-derived cell line (and these potentially misleading SNPs being removed from analysis). However, as the analysis is done *en mass* (not at the single cell level), it might be that our inference of escape from X-inactivation is wrong.

A key possibility is that each cell in a given cell culture is not uniformly inactivating the same X chromosome (intra-cell lineage heterogeneity). While eliminating SNPs at lower than 10% frequency will eliminate any instances where there is rare cell lineage heterogeneity, could it be that some higher proportion of cells, at least in some samples, are inactivating the paternally derived X but the remaining cells are inactivating the maternal X? In principle, this could lead us to misclassify intra-lineage heterogeneity for escape from X-inactivation. This is a priori unlikely, not least because the silencing of X-linked genes is achieved during early embryogenesis (Brown et al. 1991; Heard and Disteche 2006), so in a given cell line we would expect only one X to be active. More

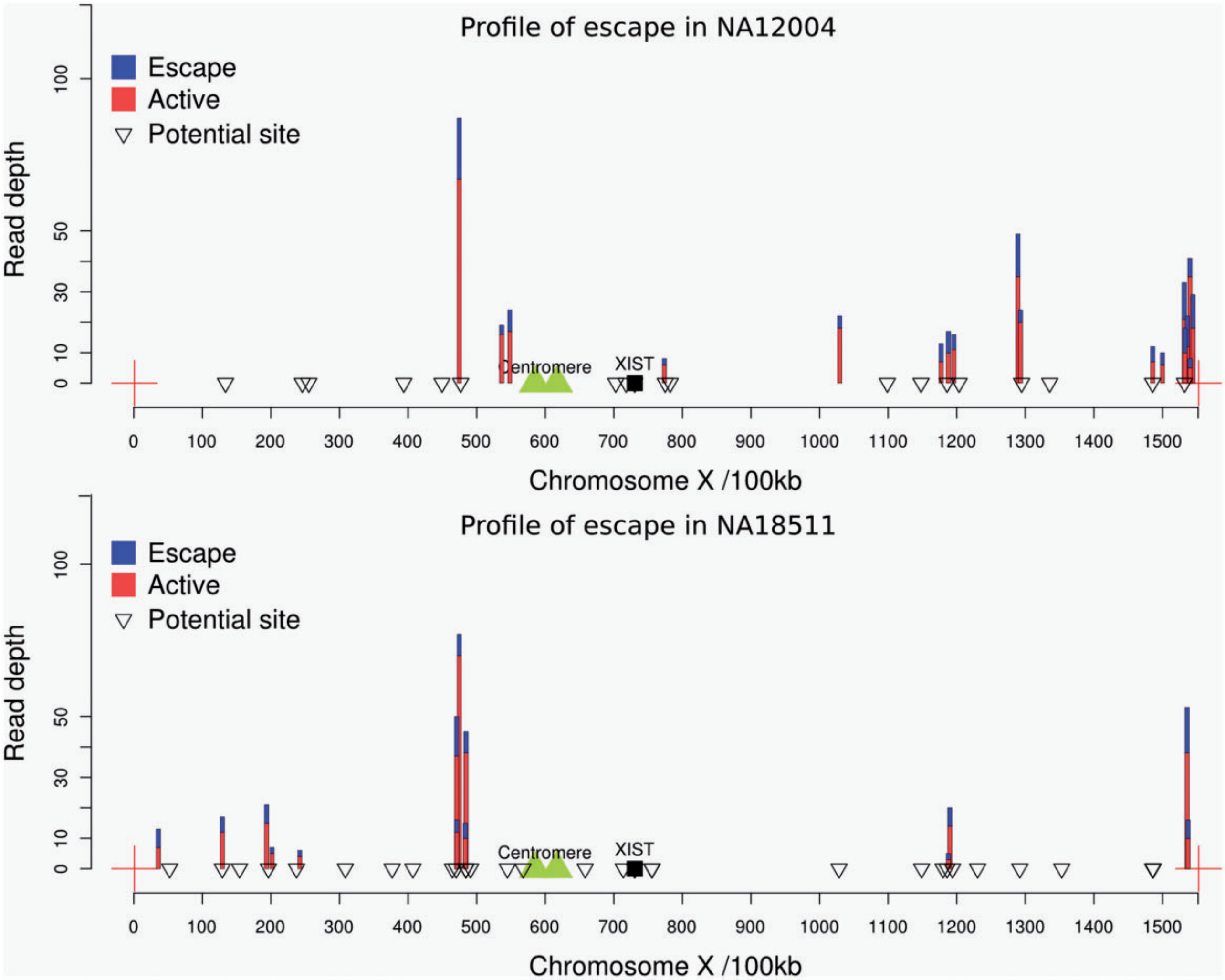

FIG. 3. Location of escape genes and haploid expressed genes on the X chromosome of one individual of CEU (NA12004) and YRI (NA18511). Genes marked as a "potential site" are those where there is exonic heterozygozity at the DNA level and transcripts that pass the coverage threshold but that do not show evidence of escape (i.e., no evidence of biallelic expression). Those marked in blue/red show evidence of escape. The sum height of the colored bar indicates the net read depth summing over both alleles. The proportion of blue to red indicates the proportion of expression from the inactive X chromosome (blue) and the active X chromosome (we always presume the minority allele is from the inactive X chromosome). The data for the pattern of escape from the remaining individuals are shown in supplementary figure S5, Supplementary Material online.

importantly, the possibility of intra-cell lineage heterogeneity in the filtered data is strongly rejected on three counts. First, if a cell line is heterogeneous for which X chromosome is inactivated, we should expect that all, or nearly all, genomically heterozygous genes should show evidence of escape by our method. This we never observe (see fig. 3). Second, and related, were any cell lines heterogeneous, we would not expect to be unable to "replicate" all prior examples of escape. Third, were there heterogeneity for which X to inactivate in the cell population, we should detect escape genes all along the chromosome and not in proximity to known escapees. In contrast to this expectation, the great majority of our escapees map to the same genomic locations, ones known previously to harbor escapees and in evolutionarily modern strata, where escape is expected (Lahn and Page 1999). As we noted, several more cluster around *Xic*, a cluster hinted at before. For the above reasons, we can confidently reject the possibility of false attribution of escape owing to intra-cell lineage heterogeneity.

We note that our evidence for escape does not preclude the possibility that the genes are haploid expressed in any given cell. It is possible that our escape genes are subject to allelic exclusion, permitting haploid expression in any given cell, but with the two alleles being each expressed in different cells within the cell lineage: some of the time the paternally derived allele is expressed, sometimes the maternally derived one, but not necessarily both in any given cell, at any given time. In this instance, the genes escape X-inactivation, in the sense that in some cells the genes are not subject to the usual inactivation that affects the rest of the chromosome. As these genes, although haploid expressed, are not subject to X-inactivation, we consider them a bona fide possible instance of escape. We note, however, that the inference of escape (in this and prior *en mass* analyses) need not imply diploid expression in any given cell. We suggest that single cell transcriptomics would be a sensible follow-up analysis, both to confirm our findings and to resolve whether escapees are subject to mono- or bi-allelic expression within any given cell.

That the prior finding of strong purifying selection on genes that always escape X-inactivation (Park et al. 2010) is not robust to addition of one set of extra data (and that from potentially more "natural" cell lines rather than inter-specific hybrids) leads us to suggest that it is better to withhold firm statements about the mode of evolution of genes in the three classes until more cell types are sampled. We do not wish to conclude that genes in the heterogeneous class are under weaker purifying selection, just that with the limited data available this is currently the best tentative conclusion. That the between-class heterogeneity is possibly sensitive to control for gene expression level provides further reason to be cautious in interpretation. We do wish to suggest that the prior claims (Park et al. 2010) for especially strong purifying selection on genes that always escape X-inactivation, and the concomitant interpretation of this in terms of dominance, should not be considered as robust. Given too that the sometimes escape class are not intermediate in their evolutionary rate between the always and never class (in neither the original nor this subsequent analysis) suggests that a simple interpretation in terms of dominance is not immediately attractive. The difference between the rate of evolution of genes that sometimes escape and those that always escape is unlikely to be owing to masking by Y-linked homologs as for both cases the presence of a Y-linked homolog is equally unlikely (supplementary table S4, Supplementary Material online). Y linked homologs are considerably more common for genes that always escape X-inactivation (supplementary table S4, Supplementary Material online).

If the genes that sometimes escape are those fastest evolving why might this be? Here we can only conjecture. Given that escape genes are strong candidates for sex-biased genes (Ellegren and Parsch 2007) and given faster evolution of sex-biased genes, differential strengths of purifying selection or positive selection associated with differential involvement in sex-biased expression would be a possibility worthy of future scrutiny. A further quandary is why it is that the X-inactivation status can be variable within a species but classes of gene appear to have characteristic evolutionary rates between species. One possibility is that the classificatory status (always escape, never escape, and sometimes escape) is relatively well conserved. Park et al. (2010) assert from unpublished work that X-inactivation status is conserved across primates. With cross-species data on X-inactivation status, this suggestion can be scrutinized further.

Materials and Methods

Data Collection

We used data generated by RNA sequencing of immortalized B-cells obtained from CEU and YRI individuals (Cheung et al. 2010). The RNA sequencing data were downloaded from the NCBI GEO database (Barrett et al. 2009) (CEU: GSE16921 and GSE25030, YRI: GSE19480). We used all of female individuals in the CEU and YRI data sets and randomly chose males as controls. Single-end sequencing data of GSE16921 and paired-end sequencing data of GSE25030 were aligned to the genome, and mapped files were combined to identify genes that escape inactivation. Samples NA10847 and NA12414 in GSE25030 were removed because the genotypes of these individuals were not available in the published version of dbSNP provided by the HapMap project. Gene and exon annotation data were obtained from the UCSC annotation database (hg19, GRCh37).

Coverage Analysis

We used the program BEDtools (Quinlan and Hall 2010) to calculate the genome-wide alignment coverage.

Mapping of the Reads to the Reference Genomes

Reads were mapped to the reference chromosomes sequence (build hg19) using Tophat (Trapnell et al. 2009). The retrieved reads were split so that they could be mapped against a collection of splice junctions, by which the RNA sequencing data can effectively be managed. We used the default settings of Tophat to analyze reads produced by the Illumina Genome Analyzer. These settings allowed no more than two mismatches on the high-quality (left) end of the reads with a sum of the Phred quality values at all mismatch positions not exceeding 70.

Heterozygous Allele Calling and Identification

We used the program SAMTOOLS (Li et al. 2009), which uses Bayesian inference to detect SNP sites in one individual. All possible bialleles at variant sites according to the reference genome were collected, whereas heterozygous sites with a QUAL value of 20 or less (Phred quality of sequencing) and a mapped depth of 6 or less were excluded from consideration.

Regarding the nonuniformity of single-end reads with different biases on the 5′- and 3′-end of fragments, we considered the regions in which the reads mapped to both the forward and reverse strands to improve the accuracy of the fragment tail determined by sequencing. The called biallelic sites that appeared only at the tail of reads and with reads mapped against only a forward or reverse strand were removed because this variant site may have been produced due to sequencing error. To improve the confidence of the heterozygosis identification, genotype data published by the International HapMap Project were used as a reference.

Strategy and Quality Control

As X-inactivation occurs early in embryogenesis (Brown et al. 1991; Heard and Disteche 2006), all cells from a given cell line derived from a postpartum subject should express only one of two alleles. This should be true regardless of whether the cell line has one or multiple founding cells, so long as all founding cells belong to the same lineage and the time to common ancestry of cells within that lineage is post the time of X-inactivation determination. Heterozygosity of X-linked markers in the transcriptome of a cell line is thus a possible indication of escape from X-inactivation. To identify genes that express both the maternal and paternal X chromosomes, we used high-throughput RNA sequencing data from normal female individuals in the CEU and YRI groups (see Materials and Methods). RNA sequencing reads were mapped against human reference genomes. The mapped reads reflected the status of expression (Wang et al. 2009). Expression from both alleles at an X-linked locus was evidenced from validated SNP sites in the mapped reads. Homozygosity in the transcriptome of genes heterozygous at the DNA level we define as genes lacking evidence for escape from X-inactivation. However, these genes could also be imprinted or subject to allelic exclusion, these both being forms of haploid expression that need not be mechanistically coupled to X-inactivation.

Although the approach is in principle straightforward, the sequencing fold-coverage and breadth-of-coverage can, however, influence the reliability and apparent extent of biallelic expression in our data. To minimize noise, information from regions with an insufficient coverage of mapped reads should be omitted. To this end, we calculated the coverage of the mapped reads based on the exons of all genes on the X chromosome (supplementary fig. S4, Supplementary Material online). The coverage of the mapped reads in YRI was slightly less than that in CEU (but not significantly so), which could impede observation of the most informative sites in YRI. However, the normalized abundance of the X chromosome and autosomes did not show a significant bias. The prior NGS study in mice (Yang et al. 2010) considered 5× coverage an adequate minimum to call escape from X-inactivation. We prefer that 7× or greater depth of coverage is the minimum level sufficient to find transcript level heterozygosis in our study, as, if a biallelic site is expressing equally from both alleles, then 7× coverage is adequate to incorrectly infer a lack of biallelic expression less than 5% of the time. Regions with lower coverage were excluded.

To avoid the identification of false-positive heterozygosis with low numbers of silenced alleles (potentially owing to cell line heterogeneity with a rare cell lineage having the opposite X-inactivation profile or owing to sequencing artifact), we required at least a 10% ratio of rare transcript variant versus common transcript variant, this being a previously employed threshold used to identify escape genes in humans (Carrel and Willard 2005). Note that rare/common here refers to the frequency of the alleles in the transcriptome of an individual not within the population. Although by this definition we exclude leaky or artifactual signals of heterozygosity, we may in turn incorrectly increase the number non-escapees (false negatives).

The variant sites in CEU and YRI obtained from dbSNP134 published by the International HapMap Project (Altshuler et al. 2010) were used as the validated variant sites to identify our heterozygous sites detected in mapped reads based on sequencing. A total of 73,792 and 89,732 X-linked SNP sites were detected in the CEU and YRI, respectively. Of these 21,087 SNPs and 26,413 are SNPs inside genes in CEU and YRI, respectively (31.24 and 37.41 SNPs per gene). However, most of these are intronic and hence of no utility for detection of escape from X-inactivation. Of the 1,001 X-linked genes (which include 823 known human protein-coding genes and 178 non-protein-coding genes [Hsu et al. 2006]), 675 and 706 X-linked genes identified in CEU and YRI, respectively, were considered to be potentially informative containing at least one well resolved exonic SNP in our sample.

Quality Control of Data: Mapping Errors are Rare

Before considering the derivation of genes potentially subject to escape from X-inactivation, as evidenced by heterozygosity in RNA-Seq samples, we investigate the quality of the data. Even with the quality control that we impose mapping errors may yet be an issue. This could be acute in the case of missing duplicate genes. Imagine we focus on an X-linked gene. Imagine too that this X-linked gene has, at least in some individuals, a paralog elsewhere in the genome but that this paralog does not feature in the reference genome. Under this circumstance, we would be forced to map the transcript from the non-focal gene back to the focal gene. If the two duplicates are allelically different, then we might incorrectly infer escape from X-inactivation. Ensuring that we employ only well-described SNPs from HapMap for the focal genes should mitigate this problem to a large degree (any random mutation in the non-focal gene we would not consider as evidence for heterozygosity) but need not necessarily eliminate it entirely. This could be considered one specific manifestation of the more general problem of incorrect mapping of RNA-Seq reads to the genome.

We can examine this problem by employing expression in male-derived cell lines as a negative control. If incorrect mapping is the issue and both the focal X-linked gene and the non-focal gene are expressed in males, then males too should appear "heterozygous" on the X chromosome. We detect very few instances (three polymorphic sites in CEU and two in YRI) of heterozygosity for X-linked genes in males suggesting that our female sample is largely free of mapping error. These sites are found in genes *STS*, *FTX*, *PLXNA3*, *CXorf4B*, and *MTMR1*. *STS PLXNA3* and *MTMR1* appeared in both of CEU and YRI and *CXorf40B* appeared only in YRI. Only one site shows heterozygosis in each of five males. This is most likely to be a mapping error possibly resulting from reads being derived from the undescribed areas or CNVs.

Note too that the presence of these heterozygous X-linked genes in males need not imply a mapping issue. It could be the case that there is one X-linked gene that within the cell culture has mutated and is polymorphic for a previously identified SNP (although this is unlikely to explain repeated heterozygosity). As the RNA-Seq data are from cell cultures *en mass* (not at the single cell), we therefore expect some low residual rate of mutationally derived heterozygosity. We removed from further analysis the sites that are heterozygous in males and could have misled analysis in females.

The robust nature of the evidence is confirmed by a further negative control. If mapping is a real problem, we should also detect X-linked loci in females with three or more alleles. We detect only 26 sites in 285 genes from 37 CEU females and only 14 sites in 510 genes from 40 YRI females with more than two alleles in a given female per X-linked gene. These sites too were removed from further analysis.

In principle, analysis of pseudoautosomal genes could provide a positive control. Unfortunately, the reference SNPs in HapMap used as the validated sites were not represented by any of 19 pseudoautosomal genes (Helena Mangs and Morris 2007), with the exception of *XG*; however, there was insufficient read coverage support for *XG*. Prior analysis of the same RNA-Seq data set has demonstrated its ability to detect autosomal heterozygosity (Cheung et al. 2010).

With the above quality controls we would, in addition, expect that signals of heterozygosity or homozygosity should be consistent between SNPs from the same gene. In both populations, we have several examples (32 and 44 genes in CEU and YRI, respectively) of instances where an individual has more than one polymorphic site in each population. Within the genes containing multiple informative sites, the majority (90.3% in CEU and 90.9% in YRI) of the RNA-Seq reads are consistent, i.e., the RNA-Seq reads were either all heterozygous or all homozygous at all potentially heterozygous sites. Many of the exceptions were instances where one site is heterozygous but the other site is not called heterozygous as the read coverage was not high enough. Considering instances where there are multiple potentially informative sites (read coverage high enough), there are 1,643 cases (genes in individuals) which have multiple potential heterozygous sites as well as sufficient read coverage. Of them, there are only 75 cases ($<$5%) where at least one site is not consistent with others.

Randomization to Determine Significance of Between-Population Variation in X-Inactivation

To determine whether there is between-population variation in escape tendency, in the two populations we calculated, for each gene, how many individuals have escaped inactivation (not necessarily replicated) and how many individuals could have been informative because they are heterozygous at the DNA level. The data from individuals whose genes lack coverage of sufficiently supported reads were excluded. We performed a χ^2-like test using *P* values derived from Monte Carlo simulations. The significance test was based on the null expectation that for any given gene the proportion of escapees is identical in CEU and YRI and dependent on the amassed proportion of escapees for that gene. To this end, we took the total observed number of escapees and randomly reallocated them to the two groups as a function of the relative number of potentially informative individuals within each group. For each gene we could then calculate a χ^2 value, which could be compared against the distribution from the simulations. With low sample sizes in some instances, this Monte Carlo method is preferable to derivation of *P* from χ^2 tables. For the overall difference between the two populations, we consider the sum χ^2 over all genes.

Molecular Evolutionary Rate Consideration and Merging of Data Sets

We downloaded from Ensembl a list of human macaque X-linked orthologs and associated K_a and K_s values. DAVID (http://david.abcc.ncifcrf.gov/conversion.jsp, last accessed September 20, 2013) was employed to convert Ensembl IDs to Refgene names. We then considered the genes that were informative in our sample (had SNPs and sufficient read coverage) and asked for how many we had rate estimation. We identified 291 such genes.

To consider the relationship between escape status and rate of evolution, we merge our data with that from the prior analysis (data from supplementary table S7, Supplementary Material online, of Park et al. [2010]). We apply the rule that if a gene has information from only one of the two data sets, then that data are preserved. If both sets agree on the status (always escape, heterogeneous, never escape), then the status is preserved. If the data sets disagree, then the gene is regarded as being in the heterogeneous class (i.e., sometimes escaping). Thus, some of the genes previously considered to always escape X-inactivation can now be considered in the sometimes escape class and some previously in the "never escape" class can also be reclassified as sometimes escape.

Rate of Gene Expression

The mean expression of 11,449 genes in 28 human tissues was derived from BioGPS, this corresponding to the data from the Affimetrix array analyzed by Su et al. (2004). We summarized GCRMA normalized probe intensity levels to Ensembl IDs corresponding to protein coding genes. All probes matching to more than one Ensembl gene ID were removed. We applied a mask to all expression values lower than the average of the expression of the negative controls in each tissue, transforming them to 0. Any gene that had expression values lower than the average of the negative controls in every tissue was removed. Expression values were then normalized against the total signal level in each tissue. Only after all the filtering did we extract only those genes that are X-linked.

Supplementary Material

Supplementary figures S1–S5 and tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

L.D.H. is a Wolfson Royal Society research merit Award Holder. A.O.U. is a Royal Society Dorothy Hodgkin Research Fellow. This work is supported by the National Basic Research Program of China (No. 2011CBA00400), the National Natural Science Foundation of China (No. 81030015), and the E-Institutes of Shanghai Municipal Education Commission.

References

Altshuler DM, Gibbs RA, Peltonen L, et al. (69 co-authors). 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58.

Barrett T, Troup DB, Wilhite SE, et al. (14 co-authors). 2009. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.* 37:D885–D890.

Belmont AS, Bignone F, Ts'o PO. 1986. The relative intranuclear positions of barr bodies in XXX non-transformed human fibroblasts. *Exp Cell Res.* 165:165–179.

Berletch J, Yang F, Xu J, Carrel L, Disteche C. 2011. Genes that escape from X inactivation. *Hum Genet.* 130:237–245.

Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, Willard HF. 1991. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* 349:38–44.

Brown CJ, Greally JM. 2003. A stain upon the silence: genes escaping X inactivation. *Trends Genet.* 19:432–438.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434:400–404.

Cheung VG, Nayak RR, Wang IX, Elwyn S, Cousins SM, Morley M, Spielman RS. 2010. Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol.* 8:e1000480.

Deng HW, Xu FH, Liu YZ, et al. (11 co-athors). 2002. A whole-genome linkage scan suggests several genomic regions potentially containing QTLs underlying the variation of stature. *Am J Med Genet.* 113: 29–39.

DeVeale B, van der Kooy D, Babak T. 2012. Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. *PLoS Genet.* 8:e1002600.

Disteche CM. 1999. Escapees on the X chromosome. *Proc Natl Acad Sci U S A.* 96:14180–14182.

Disteche CM, Filippova GN, Tsuchiya KD. 2002. Escape from X inactivation. *Cytogenet. Genome Res.* 99:36–43.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol.* 23: 327–337.

Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet.* 8:689–698.

Gecz J, Shoubridge C, Corbett M. 2009. The genetic landscape of intellectual disability arising from chromosome X. *Trends Genet.* 25: 308–316.

Gustavson KH. 1999. Triple X syndrome deviation with mild symptoms. The majority goes undiagnosed. *Lakartidningen* 96:5646–5647.

Heard E, Disteche CM. 2006. Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes Dev.* 20: 1848–1867.

Helena Mangs A, Morris BJ. 2007. The human pseudoautosomal region (PAR): origin, function and future. *Curr Genomics.* 8:129–136.

Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC known genes. *Bioinformatics* 22:1036–1046.

Kelkar A, Thakur V, Ramaswamy R, Deobagkar D. 2009. Characterisation of inactivation domains and evolutionary strata in human X chromosome through Markov segmentation. *PLoS One* 4:e7885.

Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science* 286:964–967.

Lai Y. 2007. A moment-based method for estimating the proportion of true null hypotheses and its application to microarray gene expression data. *Biostatistics* 8:744–755.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

Linden M, Bender B, Robinson A. 1995. Sex chromosome tetrasomy and pentasomy. *Pediatrics* 96:672–682.

Liu Y-Z, Xiao P, Guo YF, et al. (13 co-athors). 2006. Genetic linkage of human height is confirmed to 9q22 and Xq24. *Hum Genet.* 119: 295–304.

Liu YZ, Xu FH, Shen H, et al. (15 co-authors). 2004. Genetic dissection of human stature in a large sample of multiplex pedigrees. *Ann Hum Genet.* 68:472–488.

Lopes A, Burgoyne P, Ojarikre A, Bauer J, Sargent C, Amorim A, Affara N. 2010. Transcriptional changes in response to X chromosome dosage in the mouse: implications for X inactivation and the molecular basis of Turner Syndrome. *BMC Genomics* 11:82.

Lyon MF. 1961. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* 190:372–373.

Otter M, Schrander-Stumpel CT, Curfs LM. 2010. Triple X syndrome: a review of the literature. *Eur J Hum Genet.* 18:265–271.

Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931.

Park C, Carrel L, Makova KD. 2010. Strong purifying selection at genes escaping X chromosome inactivation. *Mol Biol Evol.* 27:2446–2450.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

Rao E, Weiss B, Fukami M, et al. (17 co-athors). 1997. Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. *Nat Genet.* 16: 54–63.

Rooman RP, Van Driessche K, Du Caju MV. 2002. Growth and ovarian function in girls with 48,XXXX karyotype—patient report and review of the literature. *J Pediatr Endocrinol Metab.* 15:1051–1055.

Stevenson RE, Schwartz CE. 2009. X-linked intellectual disability: unique vulnerability of the male genome. *Dev Disabil Res Rev.* 15:361–368.

Su AI, Wiltshire T, Batalov S, et al. (13 co-authors). 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A.* 101:6062–6067.

Tartaglia N, Howell S, Sutherland A, Wilson R, Wilson L. 2010. A review of trisomy X (47,XXX). *Orphanet J Rare Dis.* 5:8.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105–1111.

Tsuchiya KD, Greally JM, Yi Y, Noel KP, Truong JP, Disteche CM. 2004. Comparative sequence and X-inactivation analyses of a domain of escape in human xp11.2 and the conserved segment in mouse. *Genome Res.* 14:1275–1284.

Visscher PM, Macgregor S, Benyamin B, et al. (14 co-authors). 2007. Genome partitioning of genetic variation for height from 11,214 sibling pairs. *Am J Hum Genet.* 81:1104–1110.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 10:57–63.

Yang F, Babak T, Shendure J, Disteche CM. 2010. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res.* 20: 614–622.

Presence–Absence Variation in *A. thaliana* Is Primarily Associated with Genomic Signatures Consistent with Relaxed Selective Constraints

Stephen J. Bush,[1] Atahualpa Castillo-Morales,[1] Jaime M. Tovar-Corona,[1] Lu Chen,[1,‡] Paula X. Kover,[1] and Araxi O. Urrutia*,[1]

[1]Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

[‡]Present address: Human Genetics, Wellcome Trust Sanger Institute, Genome Campus, Hinxton, UK

***Corresponding author:** E-mail: a.urrutia@bath.ac.uk.

Associate editor: Stephen Wright

Abstract

The sequencing of multiple genomes of the same plant species has revealed polymorphic gene and exon loss. Genes associated with disease resistance are overrepresented among those showing structural variations, suggesting an adaptive role for gene and exon presence–absence variation (PAV). To shed light on the possible functional relevance of polymorphic coding region loss and the mechanisms driving this process, we characterized genes that have lost entire exons or their whole coding regions in 17 fully sequenced *Arabidopsis thaliana* accessions. We found that although a significant enrichment in genes associated with certain functional categories is observed, PAV events are largely restricted to genes with signatures of reduced essentiality: PAV genes tend to be newer additions to the genome, tissue specific, and lowly expressed. In addition, PAV genes are located in regions of lower gene density and higher transposable element density. Partial coding region PAV events were associated with only a marginal reduction in gene expression level in the affected accession and occurred in genes with higher levels of alternative splicing in the Col-0 accession. Together, these results suggest that although adaptive scenarios cannot be ruled out, PAV events can be explained without invoking them.

***Key words:* exon deletion, presence–absence variation, whole genome evolution, transposable elements, adaptive evolution, *Arabidopsis*.**

Introduction

Intraspecies variation in gene content represents an important source of heterogeneity in the genome of a species and potentially contributes to an organism's adaptability in response to external pressures (Feuk et al. 2006). Cataloguing significant gains and losses in coding regions within or between species will allow a deeper understanding of the mechanisms underlying the molecular evolution of genomes and can assist in identifying functional variation in agronomically elite varieties of staple crops (Wang, You, et al. 2013). To this end, several studies have examined polymorphic full or partial gene loss in several plant species. For instance, after resequencing 50 rice genomes, up to 1,327 possible gene loss events (2.4% of the total gene set) were detected relative to the Nipponbare reference accession (Xu et al. 2012). Significant intraspecies variation in gene content has also been reported in maize (Swanson-Wagner et al. 2010), sorghum (Zheng et al. 2011), and soybean (McHale et al. 2012). Previous studies in the model plant *Arabidopsis thaliana*, using resequencing microarrays and Illumina sequencing-by-synthesis reads, have also shown significant variations in total nuclear genome sequence among naturally occurring strains (Clark et al. 2007; Ossowski et al. 2008). A more recent study using 18 fully sequenced *A. thaliana* genomes found that, relative to the reference accession Col-0, 93.4% of proteins had intraspecies variation in their genes, inclusive of large deletions (Gan et al. 2011) with around 775 genes per accession found to have deletions spanning 50% or more of their coding region sequence (Gan et al. 2011). A comparison of 80 *Arabidopsis* genomes found that 9% of the total genes in *A. thaliana* showed presence–absence variation (PAV) averaging 444 absent genes per accession (Tan et al. 2012).

Characterization of coding region PAV has shown certain gene categories to be significantly enriched. For instance, 52 of the 154 nucleotide-binding site leucine-rich repeat (NBS-LRR) *R* (resistance) genes were found to be deleted in at least 1 of 50 rice cultivars (Xu et al. 2012). Similar overrepresentation of the *R* genes in *A. thaliana* has also been observed (Bakker et al. 2006; Shen et al. 2006), while in the soybean, genes enriched in structural variation are more likely to be involved in nucleotide binding and biotic defense (McHale et al. 2012). Enrichment of particular functional gene categories among genes affected by structural polymorphism suggests these structural polymorphisms may have a functional role, allowing accessions to be better adapted to the environmental conditions they face.

However, this hypothesis has not been explicitly tested. If significant polymorphic deletions are adaptive, we would expect that affected genes should show multiple signatures of being under selection. On the other hand, if structural polymorphisms mostly affect genes evolving under relaxed constraints, then their adaptive significance should be questioned.

Here we characterize genes affected by PAV spanning whole exons in *A. thaliana*, to investigate which genomic features, if any, are associated with these polymorphisms. Our results provide insights into the likely functional impact of structural variation in protein-coding genes.

Results

In order to characterize PAV in *A. thaliana*, we examined previously identified polymorphic deletions in 17 fully sequenced *Arabidopsis* accessions for which transcriptome data were available (Gan et al. 2011) (see Materials and Methods). We compiled a set of deletions that spanned entire exons in any of 17 accessions relative to the Col-0 reference genome. A subset of the annotated deletions was experimentally validated (Gan et al. 2011). To further rule out the possibility of wrongly identifying deletions due to differences between assemblies, exons were confirmed as missing by searching for homology between the Col-0 exon on all other accessions (see Materials and Methods).

A total of 794 exons were classified as missing in at least one of 17 accessions, corresponding to 411 genes (~1.5% of the total gene set) including 81 genes where the full coding region was completely absent in at least one accession (supplementary table S1, Supplementary Material online). Exon losses are not uniformly distributed throughout the gene: missing exon sequences are more often found near the ends of each gene (supplementary fig. S1, Supplementary Material online).

Overall, ~0.3% of the genes in each accession have at least one missing exon, representing between 10 and 50 kb of missing sequence per accession (supplementary table S2, Supplementary Material online). A total of 200 genes had exon loss affecting more than one accession, consistent with a previous study reporting a "common history" to deletion events in *A. thaliana* (Santuari et al. 2010).

Because partial deletions spanning whole exons might have distinct functional implications compared with full coding region deletions, the 330 genes with partial coding region loss spanning at least one full exon in at least one accession (exon PAV [E-PAV]) and the 81 genes with full coding region polymorphic deletions affecting at least one accession (full coding DNA sequence PAV [CDS-PAV]) were examined separately.

Genes Involved in Signal Transduction and Both Nucleotide and Protein Binding Are Overrepresented among PAV Genes

In order to characterize PAV genes, we first assessed whether these genes were overrepresented in particular gene classes or gene ontology (GO) categories. To do so, we used four classification schemes: "GO," a condensed set of GO terms (GOslim), the Pfam protein domain database and the family classification scheme of (Gan et al. 2011) (see Materials and Methods). Of the 330 E-PAV genes, we found most to be poorly characterized with 50% of them having no associated GOslim term. The proportion of poorly characterized genes is greater among CDS-PAV genes, with more than 60% having no associated GOslim term for biological process. When examining genes with associated GOslim terms we found both E-PAV and CDS-PAV genes to be significantly enriched in genes associated with signal transduction and nucleotide binding (fig. 1 and supplementary fig. S2, Supplementary Material online). Furthermore, E-PAV genes also appear significantly enriched in genes associated with the GOslim term "other binding," which includes proteins that bind to lipids, metal ions, and ATP, among other cofactors (fig. 1). Significant overrepresentation of functional categories among PAV genes is consistent with a previous assessment of large coding region indels in the soybean genome (McHale et al. 2012) and of whole gene deletions in *A. thaliana* (Tan et al. 2012). This is also observed when classifying genes using a broader set of GO rather than "GOslim" terms (supplementary fig. S3, Supplementary Material online).

When classifying genes by family, we observe an overrepresentation of members of the NBS-LRR family—involved in pathogen detection (DeYoung and Innes 2006)—among E-PAV genes (families "NBS-LRR active TNL," adjusted P value = 8.57×10^{-35}, and "NBS-LRR active CNL," adjusted P value = 4.63×10^{-5}; fig. 1), consistent with previous findings (Shen et al. 2006). Furthermore, when examining the 3,753 Pfam ID gene associations (supplementary figs. S4 and S5, Supplementary Material online), we observe an overrepresentation of members of the NB-ARC (APAF-1, R proteins, and CED-4) and LRR domain containing families (note that "NBS-LRR" refers to a composite of the NBS and LRR domains and that the NBS domain is also known as "NB-ARC" [McHale et al. 2006]). No enrichment of any particular gene family was observed among CDS-PAV genes (data not shown).

These significant enrichments in gene functional and domain annotations are in line with previous findings in *Arabidopsis* (Tan et al. 2012) and other plant species (Swanson-Wagner et al. 2010; Zheng et al. 2011; McHale et al. 2012) and have been proposed to reflect the adaptive role of large polymorphic deletions.

Genes Affected by PAV Show Signatures Consistent with Relaxed Selective Constraints

To determine whether PAV genes are generally associated with fast evolving proteins potentially under positive selection, we examined the rates of nonsynonymous to synonymous changes per gene (d*N*/d*S*). Using a randomization test, E-PAV genes were found to have a significantly higher d*N*/d*S* ratio compared with genes with all exons present, but only eight genes have a d*N*/d*S* ratio above 1 (fig. 2 and supplementary tables S1 and S3, Supplementary Material online). CDS-PAV genes had a nonsignificant increase in d*N*/d*S* compared with intact genes (those not affected by deletions spanning at least one exon in any accession; fig. 2 and supplementary table S3, Supplementary Material online).

To further examine the selective pressures associated with PAV genes, we examined nucleotide diversity. We considered nucleotide diversity at both replacement sites and silent sites (defined as noncoding sites and the synonymous sites of

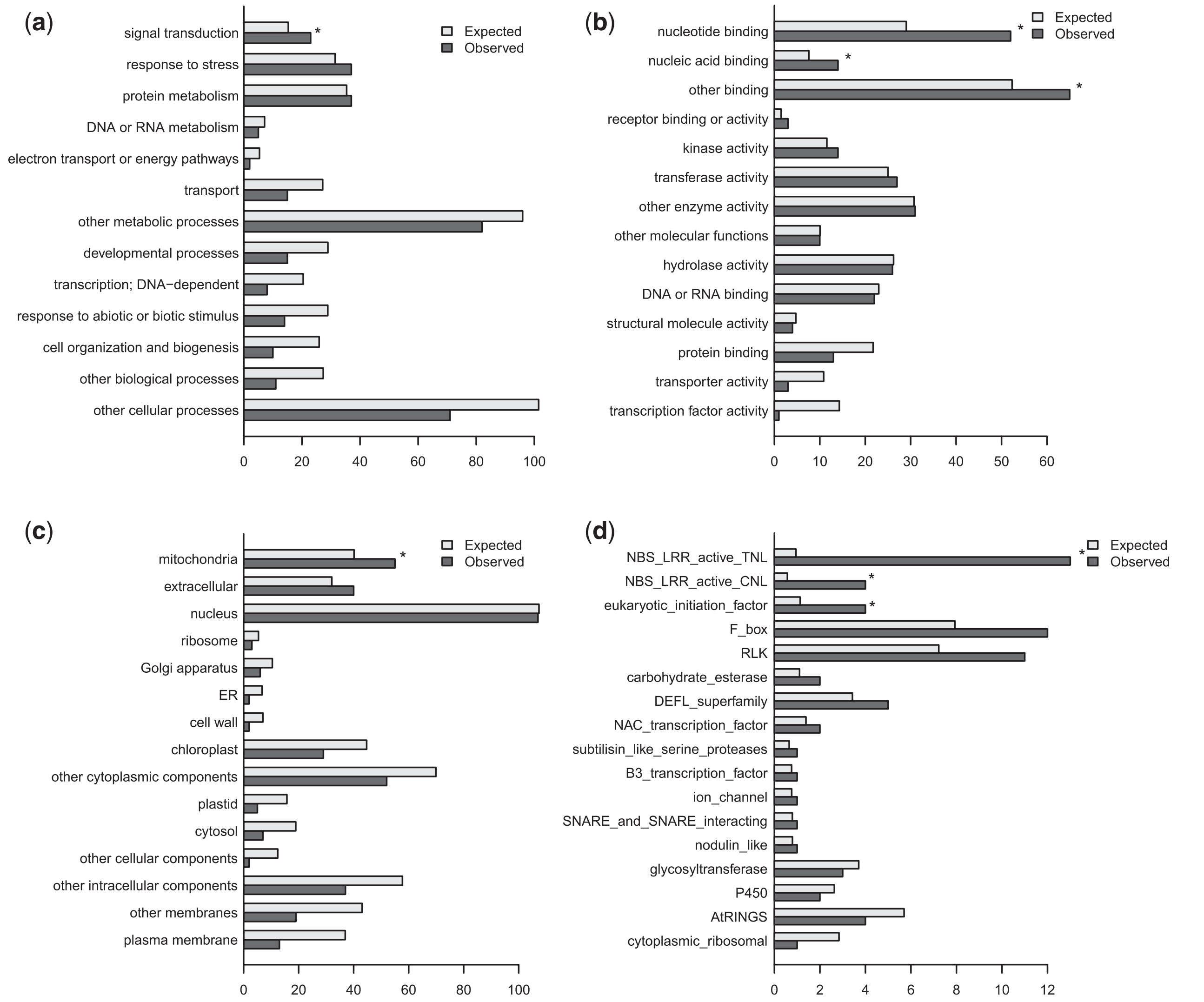

FIG. 1. Distribution of E-PAV genes ($n = 330$)—those with at least one, but not all, exons missing in at least one accession—by GOslim categories for molecular function (*a*), biological process (*b*) and cellular component (*c*), and by family (*d*). Both expected and observed number of E-PAV genes per category represented on each bar. Where there is a significant enrichment ($P \leq 0.05$) between the amount of observed and expected E-PAV genes for a particular category, an asterisk is shown over the bars. Only categories with at least one E-PAV gene are shown.

protein-coding regions) for each gene, according to Gan et al. (2011). PAV genes were found to be associated with higher nucleotide diversity in both silent and replacement sites (supplementary table S3, Supplementary Material online).

Although higher d*N*/dS and nucleotide diversity are suggestive of relaxed selective constraints, this pattern is also consistent with a scenario of positive and/or balancing selection. To differentiate between these possible scenarios, Tajima's *D* was calculated for each gene (see Materials and Methods). A threshold of ±2 was considered as the point at which *D* significantly departs from the null expectation of neutral evolution for any given gene. Of the 330 E-PAV genes with E-PAV, 24 have $D < -2$ and only 2 have $D > 2$ (AT1G12180, $D = 2.17$, and AT5G35460, $D = 2.05$, both of which are functionally uncharacterized). Among CDS-PAV genes, only seven have $D < -2$ and none have $D > 2$. Compared with the set of intact genes, there are no significant differences in the proportion of PAV genes either with $D < 2$ (randomization test $P = 1$ for both E- and CDS-PAV genes) or $D > 2$ (randomization test $P = 0.93$ and $P = 1$ for E- and CDS-PAV genes, respectively). As demographic characteristics of the *Arabidopsis* population may result in a shift in the average Tajima's *D* among the general pool of genes, it is possible that these hard thresholds may not be informative. Indeed, we find that intact genes in *Arabidopsis* have the average Tajima's *D* estimate shifted toward negative values. Thus, PAV genes could fall short of the hard threshold of + 2 and still have a higher *D* estimate than the general pool of genes, suggestive of balancing selection. However, E-PAV genes do not show significant differences in Tajima's *D* estimates compared with intact genes and CDS-PAV genes have; in fact, a significantly lower estimate of *D* (fig. 2 and supplementary tables S1 and S3, Supplementary Material online). It is possible that PAV genes may have a higher range of *D* values compared with intact genes, hiding a higher proportion of genes under positive and balancing selection that would not be reflected in overall changes in the mean. To test this, we compared the distributions of Tajima's *D*

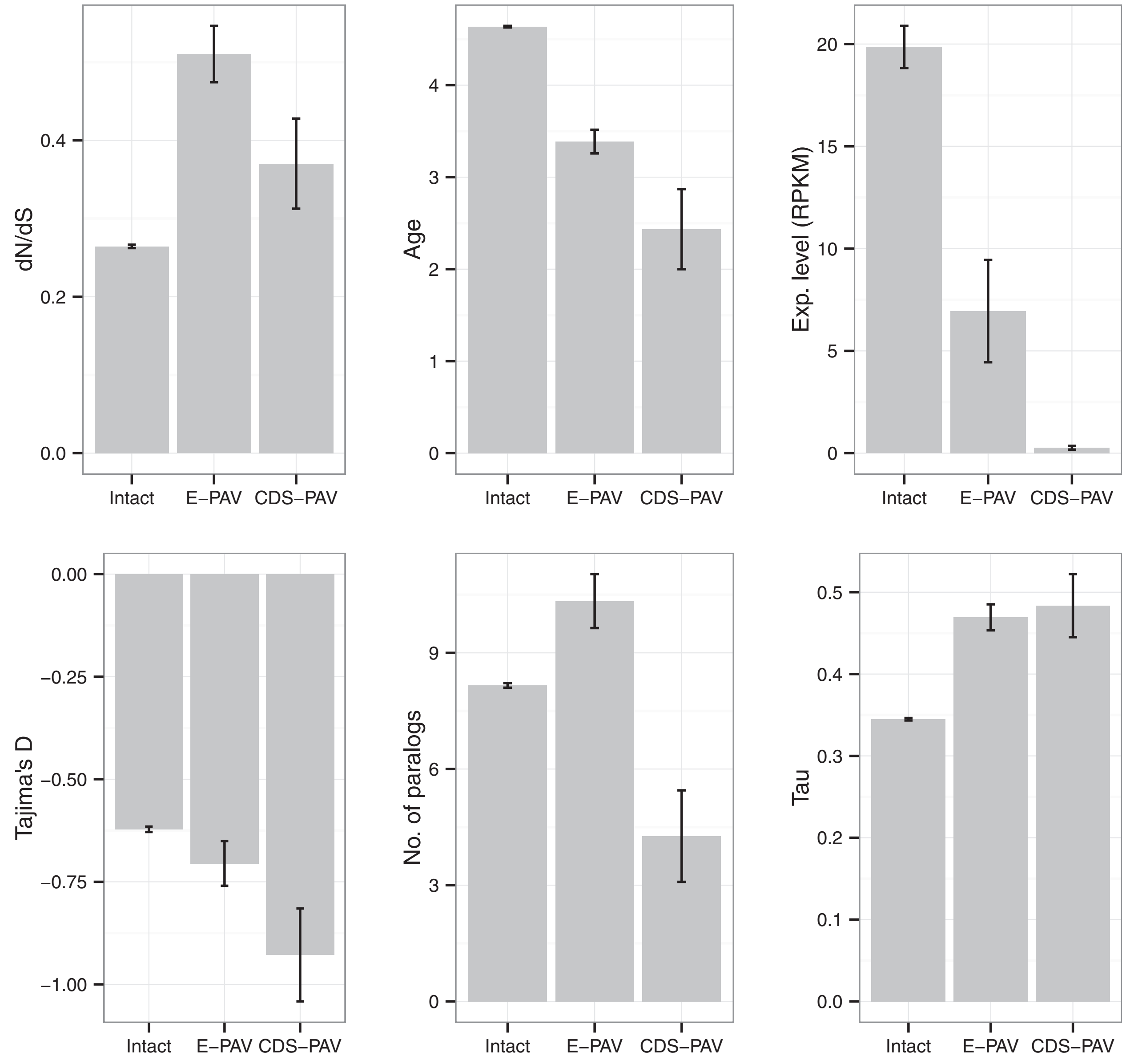

Fig. 2. Genetic features associated with intact (having no exons under P/A variation), E-PAV (having at least one, but not all, exons missing in at least one accession), and CDS-PAV (having the entire CDS missing in at least one accession) genes. From left to right, top to bottom: d*N*/d*S*, age, expression level, Tajima's *D*, number of paralogs, and *tau*. See supplementary table S3, Supplementary Material online, for values of means and statistical analysis.

estimates in the three sets of genes (intact, E-PAV, and CDS-PAV). However, we did not observe any evidence for increased dispersion in *D* among PAV genes (supplementary fig. S6, Supplementary Material online). To further examine this possibility, we examined the proportion of PAV genes below the fifth and above the 95th percentile of the "intact" distribution ($D = -2.05$ and 1.39, respectively). If a significantly higher proportion of E-PAV or CDS-PAV genes are found compared with the intact set at the positive end of the distribution, we can infer the existence of a detectable subset of PAV genes that may be undergoing balancing selection. However, such a pattern is clearly not observed—only 2.33% of E-PAV—and no CDS-PAV genes exceed the threshold value. At the opposite end of the distribution, we observe no overrepresentation in the proportion of E-PAV genes whose estimates of *D* are lower than the threshold (3.32%) although we do observe this for CDS-PAV genes (8.82%). This finding would suggest that a significant proportion of CDS-PAV genes might be undergoing stronger purifying or positive selection relative to intact genes. Together, these results suggest that although we cannot rule out the effect of balancing selection acting on a few individual PAV genes a general trend of balancing selection for PAV genes does not readily apply. The excess of negative *D* values among PAV genes coupled with the higher levels of nucleotide diversity and the significant increases in d*N*/d*S* ratios are consistent with a scenario of weaker purifying selection but could also be explained by positive selection.

We examined a number of parameters that have been previously associated by some studies with gene essentiality to further explore the functional importance of PAV genes, including a gene's age (Chen, Trachana, et al. 2012) and the number of paralogs it has (Hanada et al. 2009; Makino et al. 2009), along with weaker associations such as expression level (Cherry 2010) and tissue specificity (Wolf et al. 2006).

Compared with newer genes, older genes are more likely to be essential (Chen, Trachana, et al. 2012). After using the phylogenetic relationships of plant genomes to create a proxy for gene age, we observed that the 330 genes affected by E-PAV are more likely to be newer additions to the genome (fig. 2 and supplementary table S3, Supplementary Material online). It is also possible that E-PAV genes have a greater number of paralogous genes that might compensate for any loss of function. Consistent with this, we find that those genes with missing exons have higher number of paralogs compared with those genes with all exons present (fig. 2 and supplementary table S3, Supplementary Material online). However, the opposite result was observed when analyzing CDS-PAV genes—these have an average of 4.2 paralogs when compared with genes with no exon losses (fig. 2 and supplementary table S3, Supplementary Material online), suggesting their function is less essential. We then assessed the expression patterns of genes affected by exon presence–absence, because broadly and highly expressed genes are typically associated with higher levels of selection (Yang 2009). Using a randomization test, we found that genes with exon losses in one or more accessions, when compared with intact genes, had lower expression levels and higher tissue specificity (supplementary table S3, Supplementary Material online). In addition, we also observed that exons missing in at least one accession are, on average, shorter than exons present in all accessions (170 bp vs. 284 bp, randomization test $P = 9.9e^{-5}$; supplementary table S3, Supplementary Material online). However, although exons affected by polymorphic deletions are shorter on average compared with nondeleted exons, E-PAV genes are longer than unaffected genes (2,360 bp compared with 2,142 bp, respectively; randomization test $P = 0.008$; supplementary table S3, Supplementary Material online). By contrast, CDS-PAV genes—where polymorphic deletions encompass the gene's entire coding region—were found to be shorter than unaffected genes (640 bp compared with 2,142 bp, randomization test $P = 9.9e^{-5}$; supplementary table S3, Supplementary Material online).

Overall, these findings show that although certain functional categories are overrepresented among genes with exon loss, more generally significant coding region loss is prevalent among novel, lowly expressed and poorly functionally characterized genes. These genes seem to have evolved more recently in the *Arabidopsis* genome and are likely to be under reduced selective constraint.

PAV Genes Are Located in Genomic Regions That Are Gene-Poor and Transposable Element-Rich

When characterizing the genomic context of genes affected by PAV, we found that genes with both exon and full coding region loss are separated by longer intergenic distances (fig. 3 and supplementary table S3, Supplementary Material online). Transposable element density around PAV genes was then assessed as gene-poor areas have been associated with a higher transposable element (TE) density (Wright et al. 2003). To do this, we used the reference accession (Col-0) and calculated TE density for each gene in all intergenic sequence in 1- to 100-kb windows centered on each gene's midpoint by counting the number of bases found within TE annotations (see Materials and Methods). E-PAV genes were found to have an approximately 2-fold increase in the amount of bases annotated as a TE compared to genes that are intact in all accessions (e.g., TE sequence accounts for ~30% of the nongenic sequence within a 10-kb window surrounding an E-PAV gene; fig. 3 and supplementary table S4, Supplementary Material online). Significant enrichment of specific TE superfamilies was also observed, notably, DNA transposons and LTR retrotransposons (supplementary table S4, Supplementary Material online).

In addition, we found that genes with missing exons have, on average, a shorter distance from the gene boundary to the nearest TE than those genes with all exons present (2.5 kb compared to 5.7 kb; randomization test, $P = 9.9e^{-5}$; supplementary table S3, Supplementary Material online. When calculating the minimum distance to the nearest TE, classified by superfamily, E-PAV genes are significantly closer to every TE type: rolling circle TEs, DNA transposons, LTR retrotransposons, long interspersed elements (LINEs), and short interspersed elements (SINEs) (supplementary table S3, Supplementary Material online). Similar findings were obtained when analyzing TE content in the surrounding regions of CDS-PAV genes (supplementary table S3, Supplementary Material online).

Certain TE sequence motifs have been associated with recombination hotspots that could drive exon loss through promoting ectopic recombination events (Oliver and Greene 2009; Horton et al. 2012). To explore whether genes affected by PAV have a local enrichment for such hotspot motifs, we examined the density of these motifs both in and around genes (see Materials and Methods). However, we observed no significant differences in hotspot motif occupancy in the nongenic regions of windows surrounding E-PAV genes compared with intact genes (in window sizes of 1 to 100 kb centered on the gene's midpoint; supplementary table S4, Supplementary Material online). Nevertheless, a significant enrichment in hotspot motif occupancy was observed in the genic sequence of all windows centered on E-PAV genes compared with those centered on intact genes (fig. 3 and supplementary table S4, Supplementary Material online). When comparing CDS-PAV genes to the intact set, we observed no consistent pattern of higher hotspot motif density within genic regions and only a marginally higher proportion of hotspot motifs in the nongenic regions that surround them, in windows up to 3 kb in size ($P < 0.01$; supplementary table S5, Supplementary Material online).

Taken together, these results show that PAV genes are located in gene-poor and TE-rich regions of the genome, further supporting the hypothesis that PAV is associated with relaxed selective constraints. Enrichments of sequence motifs previously associated with recombination hotspots in or around PAV genes suggest that at least some exon deletion events may have resulted from recombination events involving these recombination hotspot motifs.

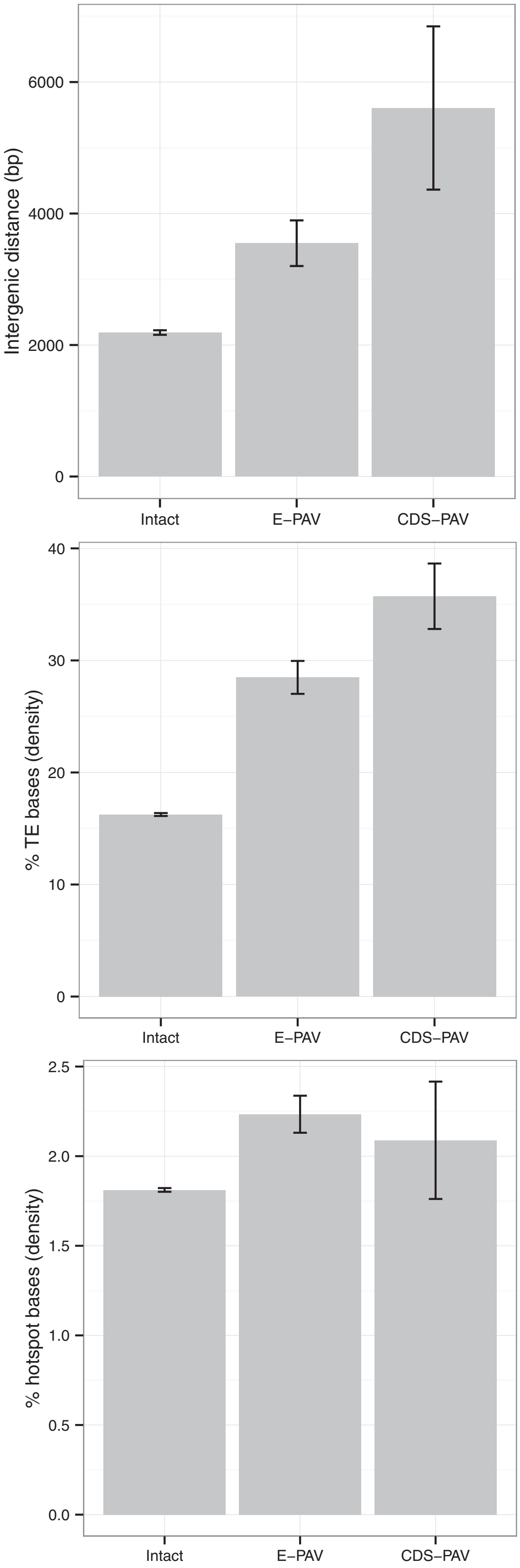

Fig. 3. Genomic context for intact (having no exons under P/A variation), E-PAV (having at least one, but not all, exons missing in at least one accession), and CDS-PAV (having the entire CDS missing in at least one accession) genes. Averaged values for the genes in each set are given for, from top to bottom, the intergenic distance, the percentage of TE bases in the nongenic sequence of a 10-kb window centered on that gene's midpoint, and the percentage of recombinogenic motifs in the genic sequence of a 1-kb window centered on that gene's midpoint. See also supplementary tables 3 and 4 (Supplementary Material online) for the values of specific TE families and other window sizes.

Exon Loss Is Associated with a Marginal Reduction in Expression Level

The aforementioned results suggest that E-PAV is associated with reduced selective constraints. To assess whether exon loss is likely to have resulted in reduced functionality for the genes affected, we compared expression levels for genes with and without missing exons across accessions. If exon loss causes or follows from diminished functionality by previous mutations, we would expect expression to be significantly reduced in those accessions affected by E-PAV. Using RNAseq transcription profiles for each *Arabidopsis* accession (Gan et al. 2011), we compared the expression patterns of individual genes in accessions affected by exon deletions with those accessions where the gene remained intact. To do this, we transformed expression data per accession to Z scores (Cheadle et al. 2003). We then looked only at those genes where exon loss had occurred in a single accession (210 genes). For each gene, we took 1) the expression level of that gene in the affected accession and 2) the mean expression level of that gene across the 17 unaffected accessions (the other 16 under study plus the reference genome, Col-0). We found that half of the genes examined had an expression level below this mean and 37% an expression level equal to it. However, on average, expression levels in the affected accession departed little from mean expression in unaffected accessions (0.15 standard deviations). In 27 genes (13% of cases), expression level in a gene affected by an exon deletion was higher than the mean expression across unaffected accessions with 14 cases showing a statistically significant difference (fig. 4 and supplementary table S6, Supplementary Material online). These 27 genes are generally poorly characterized with 12 having no functional category annotations. Most genes affected by exon deletions had low expression levels to begin with, although some exceptions are notable, such as rotamase CYP4 (AT3G62030; involved in a variety of cellular functions related to metabolism and response to several types of stress), which has an average expression level in the unaffected accessions of 400 rpkm, among the top 1% of genes with detectable expression in Col-0.

It is possible that the moderate effect of exon loss on gene expression levels is explained by an overrepresentation of alternatively spliced exons among the set of missing exons. This would allow for the production of viable protein products in their absence. In order to test this, we quantified alternative splicing in 15,540 *Arabidopsis* genes, including 103 of the 330 E-PAV associated genes using a "comparable alternative splicing index" (see Materials and Methods), which corrects for the distorting effect of variation in transcript coverage among genes (reviewed in Chen, Tovar-Corona, et al. [2012]). E-PAV genes were found to have a significantly higher number of alternative splicing events compared with

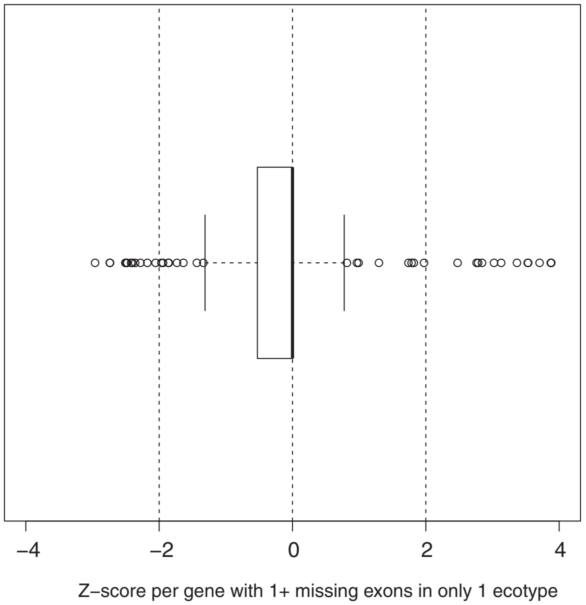

FIG. 4. Distribution of *Z* scores for standardized transcript abundance data in the affected accession. Data show that 210 genes that have one or more missing exons in only one of 17 *A. thaliana* accessions (relative to Col-0).

intact genes (3.35 and 1.13, respectively; randomization test $P = 0.046$).

Overall, these findings suggest that exon losses only have a marginal effect on the expression profile of genes in the accessions affected. The higher levels of alternative splicing among genes affected by exon loss raises the possibility that a significant proportion of lost exons are normally alternatively spliced, reducing selection pressure on these exons because a functional protein product would be produced in their absence anyway.

Discussion

Intraspecies structural variations in genes have been proposed to play an important role in the adaptation of particular populations to variation in environmental conditions (Feuk et al. 2006). Here we have characterized presence–absence coding sequence variation in 17 fully sequenced *A. thaliana* genomes, relative to the reference accession Col-0, affecting 411 genes including 81 instances of whole coding region deletions. We found a significant enrichment of genes associated with the GO terms for protein and nucleotide binding as well as signal transduction. Both gene family and Pfam annotation enrichment analysis revealed significant enrichments of gene members from the disease resistance associated NBS-LRR gene families. Significant deviations from random expectations have been observed in previous studies of PAV genes in plants, with similar overrepresentation of resistance-associated gene families among PAV genes. For instance, in sorghum (Zheng et al. 2011), PAV genes are enriched in nine Pfam categories, including the NB-ARC domain-containing family. In soybean (McHale et al. 2012), PAV-affected genes have also been found to be enriched for members of the NB-ARC family, and within the GO category of "defense response." CDS-PAV genes have also been shown to deviate from random expectations in *Arabidopsis* (Tan et al. 2012), with the greatest significant enrichment in PAV genes also reported for those with NB-ARC domains.

These functional and/or gene family enrichments can be suggestive of an adaptive role for PAV events by aiding specific ecotypes in adapting to their local environment. Our results—showing that genes associated with, for example, resistance are more likely to be affected by PAV—are, at first glance, consistent with this hypothesis. In addition, we were able to confirm a previous report of CDS-PAV for three members of the *R* gene family—the single-exon gene AT5G05400 and the multiexon genes AT5G18350 and AT5G49140 (Shen et al. 2006)—a family known to have signatures of positive selection in *A. thaliana* (Mondragon-Palomino et al. 2002). However, comprehensive analysis for evidence of selection does not support this as a general interpretation.

d*N*/d*S* ratios are one of the most widely used estimates of selective pressure acting on protein coding genes with d*N*/d*S* >> 1 indicative but not a definitive signature of positive selection (Hurst 2002). Although there are, on average, a higher number of substitutions in E-PAV genes compared with intact genes, this is not a clear signature of adaptation and can suggest comparatively relaxed negative, rather than stronger positive, selection.

We further found that PAV genes have significantly higher nucleotide diversity both at silent and replacement sites. Both observations are suggestive of weaker purifying selection; however, they can also be expected if PAV genes were under higher balancing selection. Indeed, there is evidence to suggest that the diversity of resistance-associated genes is maintained by balancing selection (Van der Hoorn et al. 2002), which are overrepresented among PAV genes. Balancing selection has been proposed to stably maintain both the intact gene and the absent allele (Tan et al. 2012).

So, is balancing selection the most parsimonious explanation for why PAV genes are associated with higher nucleotide diversity? A classic scenario of transspecies polymorphism, associated with balancing selection, cannot be assessed given the limited sequence variation data available for *A. lyrata*, *A. thaliana*'s closest sequenced relative. It is possible that the "gene/exon present" and the "gene/exon absent" alleles are under selection to be maintained in different *A. thaliana* populations, allowing them to better adapt to their local environment. This would be consistent with the increase in nucleotide diversity, but this scenario cannot be distinguished from alternative neutral models. Conditional neutrality at PAV loci, where the functional gene has ceased to be adaptive in some but not all environments, cannot be ruled out (e.g., in the case of resistance genes where the corresponding pathogen is absent [Gos and Wright 2008]). In this case, the absent allele would have no selective advantage at any point but rather result from relaxed constraints associated with PAV genes in some *Arabidopsis* populations. Moreover, a model of generalized relaxed constraints affecting the PAV loci would also lead to increased nucleotide diversity and slight increases in d*N*/d*S*.

Tajima's *D*, a comparison of two estimators of θ (the population mutation rate $4N_e\mu$)—the number of segregating sites and the average number of pairwise differences between sequences (Tajima 1989)—offers a more reliable estimate of selective pressures acting on a gene as it incorporates information about the distribution of segregating alleles in a species. This allows more accurate estimations of the degree and direction of departure of sequence evolution from a neutral expectation (although nonselectionist interpretations of *D* are also possible, such as recent population expansion or bottlenecking for negative and positive *D*, respectively) (Tajima 1989). Tajima's *D* values do not provide evidence for either E-PAV or CDS-PAV genes to be under balancing selection. Taking d*N*/d*S*, nucleotide diversity, and *D* estimates together, most PAV genes appear to be evolving under relaxed constraints.

A signature of relaxed selection associated with PAV genes is combined with a variety of features that have been associated with lower gene essentiality. We found that PAV genes have lower expression levels and higher tissue specificity; both of these features have been associated with higher rates of substitutions and reduced gene essentiality (Wolf et al. 2006; Cherry 2010). Older genes have been considered more essential (Chen, Trachana et al. 2012) and have been associated (in humans, flies and *Aspergillus*) with a higher expression level and stronger purifying selection (Wolf et al. 2009). We found that PAV genes are, on average, newer additions to the genome and that most exons affected by PAV do not have an orthologous exon in *A. lyrata* (663/794). We note that both E-PAV and CDS-PAV genes are enriched in reverse transcriptase domains (supplementary figs. S4 and S5, Supplementary Material online) and E-PAV genes for transposase domains (supplementary fig. S4, Supplementary Material online), suggesting exonization of TEs as the origin of some PAV-affected exons.

In addition, the fact that gene expression is only marginally reduced in accessions affected by exon deletion events suggests that the lost exons may only have had a limited impact on gene functionality. This is possibly explained in some cases by alternative splicing, which has already been associated with an increased frequency of exon loss in humans, mice, and rats—alternatively spliced forms are less likely to be conserved between species than constitutive exons (Modrek and Lee 2003). In *A. thaliana*, we found that genes with E-PAV are under weaker purifying selection and have a greater number of alternative splice events compared with intact genes. This observation suggests that alternatively spliced exons are likely to be under reduced selective constraints compared with constitutive exons, and thus whole exon deletions would have less of a detrimental effect than the loss of a constitutive exon. To the best of our knowledge, this is the first time that exon loss events have been associated with elevated alternative splicing levels within a species rather than between species.

The genomic context of genes has also been linked to both patterns of sequence evolution and features associated with gene essentiality. A recent study in *A. thaliana* has correlated the presence of TEs adjacent to genes with sequence variation within that gene (Wang, Weigel, et al. 2013), suggesting TEs tend to accumulate near genes under lower selective pressures located in regions with less efficient purging of TE sequence. Indeed, for our set of E-PAV genes, we find a higher density of TEs in the vicinity. In addition, we also find that genes undergoing PAV have an increased proportion of motifs associated with recombination hotspots within their sequence. Both findings are consistent with PAV events being associated with genes located in genomic regions evolving under reduced selective constraints. Moreover, higher TE content and hotspot motifs are consistent with the suggestion that unequal recombination between homologs may be a major mechanism for generating P/A polymorphisms (Tan et al. 2012). However, it should be noted that no recombinogenic motif is both necessary and sufficient for a recombination event to occur (Johnston and Cutler 2012), and as such, their connection, if any, to PAV remains speculative.

All of these features considered together suggest that although some individual deletions might have an adaptive value, overall coding region loss disproportionally affects genes under reduced selective pressures. So how are these results reconciled with the enrichment of certain gene families and GO functional terms? The enrichment of specific functional categories and gene families among PAV genes (fig. 1) leads to the implication of adaptive pressures favoring PAV on genes related to specific biological processes (Tan et al. 2012). However, as we have shown, PAV genes are associated with a variety of features suggestive of lower selective constraints. We argue that the enrichment of certain GO categories and/or gene families among genes associated with a particular genomic feature does not, by itself, allow us to draw conclusions about any adaptive processes these genes may be undergoing. Consistent with this, we find that intact genes associated with the gene categories in which PAV genes are enriched also show the same signatures of reduced selection (supplementary table S7, Supplementary Material online). This is notable for those sets of genes involved in, for example, signal transduction, nucleic acid binding, and the NBS-LRR family—categories enriched among PAV genes (fig. 1). For instance, if we compare the set of E-PAV genes to the set of genes with all exons present and the set of NBS-LRR genes to the set of genes belonging to other families, we find that both E-PAV and NBS-LRR genes are comparatively newer additions to the genome, have a higher d*N*/d*S* ratio, a higher number of alternative splicing events, a higher number of paralogs, a higher proportion of SNPs, and are found closer to TEs (supplementary table S7, Supplementary Material online). We note that the proportion of polymorphic sites is higher not only in PAV genes but in genes of that functional category. To demonstrate that PAV genes do not bias the comparison of, for example, the set of NBS-LRR genes to the set of genes belonging to other families, we repeat the analysis restricted to intact genes only and observe the same result (supplementary table S7, Supplementary Material online).

The fact that we observed fewer PAV genes than a previous study examining 80 fully sequenced *Arabidopsis* genomes ($n = 2{,}741$; Tan et al. 2012) is likely due to differences in methodology. First, our analysis uses 17 genomes assembled using a combination of read-to-reference genome (Col-0) alignment and de novo approaches, and—importantly—for which transcriptome data were available (Gan et al. 2011), rather than the 80 accessions reported by Cao et al. (2011). Second, we use a more conservative methodology for defining significant deletions while Tan et al. (2012) define PAV genes using what is referred to as the "broad definition": "one being found at a particular locus only in some genomes compared to the others." This allows a gene to be called as a PAV gene even if a copy exists at a different locus. To minimize the inclusion of rearrangement events as deletions, Tan et al. (2012) examined their predicted PAV genes using BlastN against a reference accession, excluding from the "absent" category any gene with a counterpart that matches >50% of its length. Our definition of PAV is more restrictive as we only deemed an exon or gene to be deleted if genome alignments showed that the deletion spanned at least a whole exon or whole gene with not a single identifiable base remaining. Finally, the Tan et al. (2012) study used genomes assembled according to the TAIR8 annotated positions, whereas our data are assembled according to TAIR10. There is a small risk, therefore, of having incorporated now-obsolete gene models into their findings. Regardless of the methodological differences and the resulting variation in sample size, it is worth noting that our results are not in contradiction to those of previous studies examining PAV both in *Arabidopsis* and other species, as we find similar deviations from random expectations in the functional annotations of genes. Our analysis of sequence evolution and other genic features of PAV genes do not rule out the possibilities of conditional neutrality at PAV loci or that balancing selection may be acting on PAV genes, allowing adaptation to the environmental conditions of specific ecotypes. Instead, the findings presented show that PAV events can be explained by a nonadaptive interpretation where genes under reduced constraints are more susceptible to the spread of allele variants containing significant deletions.

In summary, our results suggest that although significant enrichment in functional categories among PAV genes was observed, most exon loss events are observed in newer, poorly functionally characterized genes associated with signatures linked to less essential genes evolving under lower purifying or balancing selection. This may reduce the potential functional relevance of structural variations within these genes. We conclude that although an adaptive model for PAV cannot be ruled out, the observed functional enrichments among PAV genes and increased nucleotide diversity can also be interpreted without invoking selection.

Materials and Methods

Genome Sequence and Annotations

Exon coordinates for *A. thaliana* strain Col-0 were obtained from The Arabidopsis Information Resource (TAIR) (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_gff3/TAIR10_GFF3_genes.gff, dated 20 March 2012 [last accessed October 8, 2013]). The genomes of 17 *A. thaliana* accessions (Bur-0, Can-0, Ct-1, Edi-0, Hi-0, Kn-0, Ler-0, Mt-0, No-0, Oy-0, Rsch-4, Sf-2, Tsu-0, Wil-2, Ws-0, Wu-0, and Zu-0) were obtained from Gan et al. (2011). We did not use data from Po-0 because it has an unusually high frequency of heterozygosity and high similarity to Oy-0 (Gan et al. 2011). Each genome has been fully sequenced and assembled, using a combination of de novo assembly and read mapping to the reference accession, Col-0.

Detecting Missing Exons Relative to Col-0

For this analysis, we selected a set of deletions spanning at least one full exon in at least one accession relative to the Col-0 reference genome from a wider set of deletion events described by Gan et al. (2011). Exons absent in the Col-0 reference genome but present in other accessions are not included in any analysis. Confirmation of these deletions is described by the original authors who analyzed deletion breakpoints (Gan et al. 2011). In this data set, deletion breakpoints were estimated to within ~30 bp, with left and right consensus sequences established by growing inward from these estimates using the read-mapping information. If there was a deletion, these two ends would overlap. Gan et al. (2011) confirmed this with alignments of the left and right consensus sequences, thus excluding errors of sequencing or misassembly. We further confirmed the presence or absence of each individual exon in each of 17 accessions relative to the Col-0 genome annotation using BlastN with default parameters (Altschul et al. 1990). Sequence alignments were obtained using the best hit homolog and the Smith–Waterman algorithm (fasta35 with parameters–a–A) (Pearson 2000). We confirmed an exon as missing if both 1) alignment could not be made and 2) if none of the nucleotide positions in the Col-0 exon mapped to any nucleotide in the accession.

Functional Category Enrichment Analysis

Four gene classification schemes were obtained. GOslim terms were obtained from TAIR (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.txt, dated 9 July 2013 [last accessed October 8, 2013]), excluding terms unsupported by experimental or computational analysis, that is, evidence codes ND, NR, and NAS. GO term annotations were obtained from Ensembl BioMart (17 July 2013) (Smedley et al. 2009). "Pfam" terms were obtained from Pfam v27.0 (17 July 2013) (Punta et al. 2012). In addition, 7,119 genes were classified into 49 distinct families as in Gan et al. (2011). Statistical significance of the enrichment of both GOslim, GO terms, of Pfam class and family membership among both E-PAV and CDS-PAV-affected genes was assessed using Monte Carlo random sampling (1000 randomizations), with the *P* value of the enrichment of each category obtained using a *Z* test. The significance of individual categories was corrected for multiple testing by the Benjamini–Hochberg procedure.

Sequence Evolution Analysis

To approximate selective constraint on a gene, we calculated d*N*/d*S*. For each gene, we obtained a local alignment of the Col-0 CDS against its *A. lyrata* ortholog, using the Smith–Waterman algorithm (fasta35 with parameters–a–A)

(Pearson 2000). d*N*/d*S* was calculated using the Yang and Nielson model, as implemented in the yn00 package of phylogenetic analysis by maximum likelihood (PAML) (Yang 2009). Using substitution estimates, as above, and SNP data from (Gan et al. 2011), we also estimated Tajima's *D* (Tajima 1989) per gene. Nucleotide diversity is calculated according to Gan et al. (2011).

Paralog Number and Gene Age Annotations

Ortholog and paralog data were obtained from BioMart (Vilella et al. 2009). A proxy for gene age was established using taxonomic classifications, based on the phylostratigraphic method of (Domazet-Lošo et al. 2007). If a candidate ortholog was identified for each *A. thaliana* gene in any of 15 plant and algal species at a minimum identity of 30%, the gene was considered to be as old as the "broadest" taxonomic category held in common (see supplementary table S8, Supplementary Material online). This allowed us to make use of ortholog data despite divergence times relative to *A. thaliana* being known for only its closest relatives—at ~5 million years for *A. lyrata* (Kuittinen et al. 2004) and 20 million years for *Brassica rapa* (Yang et al. 1999).

Gene Expression

Expression specificity was calculated as a tissue specificity index (*tau*) (Yanai et al. 2005), using the massively parallel signature sequencing (MPSS) database (Brenner et al. 2000; Meyers et al. 2004; Nakano et al. 2006). Expression levels were calculated using RNAseq transcript abundance data, as absolute read values corrected by sequence length in each accession (known as rpkm values: per gene, the number of reads per kilobase per million mapped reads) (Gan et al. 2011).

TE and Hotspot Motif Density

TE coordinates for *A. thaliana* strain Col-0 were obtained from TAIR (file "TAIR10_Transposable_Elements," dated 20 March 2012). For our analyses, we identified every instance of all 25 hotspot-associated motifs (of 5–9 bp) described by Horton et al. (2012) in the Col-0 reference genome. TE and hotspot motif density for each gene were calculated as the proportion of base pairs occupied by a TE or a hotspot motif within windows of size 1 to 100 kb centered on the nucleotide at the gene's midpoint. Windows consist of both coding and noncoding sequence within a region of length (window size)/2 up- and downstream of the midpoint base. Both TE and hotspot motif density were calculated as the number of TE or motif bases, respectively, relative to the number of intergenic or genic bases contained within the window, rather than the total number of bases in the window.

Alternative Splicing Events

Alternative splicing events were identified using the methods described in Chen et al. (2011). In brief, the number of alternative splicing events per gene was identified by aligning expressed sequence tag (EST) data obtained from dbEST (Boguski et al. 1993) to the genome sequence (ftp://ftp.ncbi.nih.gov/repository/dbEST, last accessed May 1, 2011). Those ESTs aligning to regions with no annotated gene were excluded from the analysis. EST alignments were then used to create an exon template. Alternative splicing events per gene were identified by comparing alignment coordinates for each individual EST to exon annotations. As a low EST coverage can increase the number of falsely positive claims that an exon is constitutive, rather than spliced, we excluded genes with 10 or fewer ESTs. ESTs were assigned to genes using gene annotation coordinates. A comparable alternative splicing index that avoids transcript coverage biases was obtained using the transcript normalization method described in Kim et al. (2007). Briefly, for each gene 100 random samples of 10 ESTs were selected. Finally, the number of alternative splicing events were calculated for each random sample (as detailed earlier), with an overall average calculated per gene.

Randomization Test

A randomization test was used to obtain numerical *P* values to assess the statistical significance of any variation in the characteristics of PAV-affected genes compared with intact genes. In brief, we contrasted genomic feature parameters in E-PAV ($n = 330$) or CDS-PAV genes ($n = 81$) to the distribution of means of the same genomic feature in $s = 10{,}000$ randomly generated subsets of an equal number of genes drawn from the complete gene set. The numerical *P* value was calculated as follows: let *q* be the number of times the mean value of the PAV set exceeded the mean value of the randomly generated subset. Letting $r = s - q$, then the *P* value of this test is $r + 1/s + 1$.

Supplementary Material

Supplementary tables S1–S8 and figures S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors acknowledge the valuable comments made by two anonymous reviewers. This work was supported by a University of Bath fee studentship to S.J.B., CONACyT scholarships to A.C.M. and J.M.T.C., a UK–China scholarship for excellence and University of Bath research studentship to L.C., a BBSRC grant (grant number BB/F022697/1) to P.X.K., and a Royal Society Dorothy Hodgkin Research Fellowship (DH071902), Royal Society research grant (grant number RG0870644), and a Royal Society research grant for fellows (grant number RG080272) to A.U.O.

References

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

Bakker EG, Toomajian C, Kreitman M, Bergelson J. 2006. A genome-wide survey of R gene polymorphisms in *Arabidopsis*. *Plant Cell* 18: 1803–1818.

Boguski MS, Lowe TMJ, Tolstoshev CM. 1993. dbEST—database for expressed sequence tags. *Nat Genet.* 4:332–333.

Brenner S, Johnson M, Bridgham J, et al. (24 co-authors). 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol.* 18:630–634.

Cao J, Schneeberger K, Ossowski S, et al. (17 co-authors). 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet.* 43:956–963.

Cheadle C, Vawter MP, Freed WJ, Becker KG. 2003. Analysis of microarray data using Z score transformation. *J Mol Diagn.* 5:73–81.

Chen L, Tovar-Corona JM, Urrutia AO. 2011. Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts. *Hum Mol Genet.* 20:4422–4429.

Chen L, Tovar-Corona JM, Urrutia AO. 2012. Alternative splicing: a potential source of functional innovation in the eukaryotic genome. *Int J Evol Biol.* 2012:10.

Chen W-H, Trachana K, Lercher MJ, Bork P. 2012. Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. *Mol Biol Evol.* 29:1703–1706.

Cherry JL. 2010. Expression level, evolutionary rate, and the cost of expression. *Genome Biol Evol.* 2:757–769.

Clark RM, Schweikert G, Toomajian C, et al. (18 co-authors). 2007. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317:338–342.

DeYoung BJ, Innes RW. 2006. Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat Immunol.* 7:1243–1249.

Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* 23:533–539.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nat Rev Genet.* 7:85–97.

Gan X, Stegle O, Behr J, et al. (23 co-authors). 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477: 419–423.

Gos G, Wright SI. 2008. Conditional neutrality at two adjacent NBS-LRR disease resistance loci in natural populations of *Arabidopsis lyrata*. *Mol Ecol.* 17:4953–4962.

Hanada K, Kuromori T, Myouga F, Toyoda T, Li WH, Shinozaki K. 2009. Evolutionary persistence of functional compensation by duplicate genes in *Arabidopsis*. *Genome Biol Evol.* 1:409–414.

Horton MW, Hancock AM, Huang YS, et al. (13 co-authors). 2012. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat Genet.* 44:212–216.

Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18:486.

Johnston HR, Cutler DJ. 2012. Population demographic history can cause the appearance of recombination hotspots. *Am J Hum Genet.* 90: 774–783.

Kim E, Magen A, Ast G. 2007. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.* 35:125–131.

Kuittinen H, de Haan AA, Vogl C, Oikarinen S, Leppala J, Koch M, Mitchell-Olds T, Langley CH, Savolainen O. 2004. Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. *Genetics* 168:1575–1584.

Makino T, Hokamp K, McLysaght A. 2009. The complex relationship of gene duplication and essentiality. *Trends Genet.* 25:152–155.

McHale L, Tan X, Koehl P, Michelmore R. 2006. Plant NBS-LRR proteins: adaptable guards. *Genome Biol.* 7:212.

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM. 2012. Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol.* 159:1295–1308.

Meyers BC, Tej SS, Vu TH, Haudenschild CD, Agrawal V, Edberg SB, Ghazal H, Decola S. 2004. The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis*. *Genome Res.* 14:1641–1653.

Modrek B, Lee CJ. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet.* 34:177–180.

Mondragon-Palomino M, Meyers BC, Michelmore RW, Gaut BS. 2002. Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. *Genome Res.* 12:1305–1315.

Nakano M, Nobuta K, Vemaraju K, Tej SS, Skogen JW, Meyers BC. 2006. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Res.* 34: D731–D735.

Oliver KR, Greene WK. 2009. Transposable elements: powerful facilitators of evolution. *Bioessays* 31:703–714.

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. 2008. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* 18:2024–2033.

Pearson WR. 2000. Flexible sequence similarity searching with the FASTA3 Program Package. *Methods Mol Biol.* 132:185–219.

Punta M, Coggill PC, Eberhardt RY, et al. (16 co-authors). 2012. The Pfam protein families database. *Nucleic Acids Res.* 40:D290–D301.

Santuari L, Pradervand S, Amiguet-Vercher A-M, Thomas J, Dorcey E, Harshman K, Xenarios I, Juenger T, Hardtke C. 2010. Substantial deletion overlap among divergent *Arabidopsis* genomes revealed by intersection of short reads and tiling arrays. *Genome Biol.* 11:R4.

Shen J, Araki H, Chen L, Chen JQ, Tian D. 2006. Unique evolutionary mechanism in R-genes under the presence/absence polymorphism in *Arabidopsis thaliana*. *Genetics* 172:1243–1250.

Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. 2009. BioMart—biological queries made easy. *BMC Genomics* 10:22.

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM. 2010. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* 20:1689–1699.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Tan S, Zhong Y, Hou H, Yang S, Tian D. 2012. Variation of presence/absence genes among *Arabidopsis* populations. *BMC Evol Biol.* 12:86.

Van der Hoorn RAL, De Wit PJGM, Joosten MHAJ. 2002. Balancing selection favors guarding resistance proteins. *Trends Plant Sci.* 7: 67–71.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19:327–335.

Wang X, Weigel D, Smith LM. 2013. Transposon variants and their effects on gene expression in *Arabidopsis*. *PLoS Genet.* 9: e1003255.

Wang Y, You FM, Lazo GR, Luo M-C, Thilmony R, Gordon S, Kianian SF, Gu YQ. 2013. PIECE: a database for plant gene structure comparison and evolution. *Nucleic Acids Res.* 41:D1159–D1166.

Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. *Proc R Soc B.* 273:1507–1515.

Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A.* 106:7273–7280.

Wright SI, Agrawal N, Bureau TE. 2003. Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. *Genome Res.* 13:1897–1903.

Xu X, Liu X, Ge S, et al. (25 co-authors). 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol.* 30:105–111.

Yanai I, Benjamin H, Shmoish M, et al. (12 co-authors). 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21:650–659.

Yang H. 2009. In plants, expression breadth and expression level distinctly and non-linearly correlate with gene structure. *Biol Direct.* 4:45.

Yang YW, Lai KN, Tai PY, Li WH. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *J Mol Evol.* 48:597–604.

Zheng L-Y, Guo X-S, He B, et al. (11 co-authors). 2011. Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). *Genome Biol.* 12:R114.

Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity

Lu Chen,[†,‡,1] Stephen J. Bush,[†,1] Jaime M. Tovar-Corona,[1] Atahualpa Castillo-Morales,[1] and Araxi O. Urrutia*[,1]

[1]Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

[†]These authors contributed equally to this work.

[‡]Present address: Human Genetics, Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridgeshire, United Kingdom

Associate editor: Csaba Pal

***Corresponding author:** E-mail: a.urrutia@bath.ac.uk.

Abstract

What at the genomic level underlies organism complexity? Although several genomic features have been associated with organism complexity, in the case of alternative splicing, which has long been proposed to explain the variation in complexity, no such link has been established. Here, we analyzed over 39 million expressed sequence tags available for 47 eukaryotic species with fully sequenced genomes to obtain a comparable index of alternative splicing estimates, which corrects for the distorting effect of a variable number of transcripts per species—an important obstacle for comparative studies of alternative splicing. We find that alternative splicing has steadily increased over the last 1,400 My of eukaryotic evolution and is strongly associated with organism complexity, assayed as the number of cell types. Importantly, this association is not explained as a by-product of covariance between alternative splicing with other variables previously linked to complexity including gene content, protein length, proteome disorder, and protein interactivity. In addition, we found no evidence to suggest that the relationship of alternative splicing to cell type number is explained by drift due to reduced N_e in more complex species. Taken together, our results firmly establish alternative splicing as a significant predictor of organism complexity and are, in principle, consistent with an important role of transcript diversification through alternative splicing as a means of determining a genome's functional information capacity.

Key words: **organism complexity, alternative splicing, genome evolution, transcriptome evolution, expressed sequence tags.**

Introduction

Prior to widespread genome sequencing, it was assumed that organism complexity was proportional to gene content—that more complex organisms encode a greater amount of genetic information (Taft and Mattick 2003), the unit of which is the gene (Bird 1995). However, the sequencing of the human genome, revealing a lower than expected number of genes (Fields et al. 1994), initiated a hunt to uncover the genomic basis of organism complexity (Nilsen and Graveley 2010) as, despite two rounds of whole genome duplication at the base of the vertebrate lineage (Ohno 1970; Dehal and Boore 2005), the human genome contains almost as many genes as that of a worm (Lander et al. 2001). Several genomic features have been shown to have a significant association with organism complexity, measured as the number of distinct cell types per species (cell type number [CTN]). These variables include various measures of the potential number of molecular interactions per protein: the number and proportion of protein–protein interaction (PPI) domains in each protein (Xia et al. 2008; Schad et al. 2011) and protein disorder (flexibility in a protein's 3D structure to adopt a variety of conformations) (Romero et al. 2006; Dunker et al. 2008; Schad et al. 2011). More recently, total coding region length in a genome was shown to be positively associated with organism complexity (Schad et al. 2011). This same study also showed that when restricting the analysis to metazoans, gene number becomes a significant predictor of organism complexity.

Alternative splicing, a posttranscriptional process in eukaryotes by which multiple distinct transcripts are produced from a single gene, has the potential to boost the total number of distinct proteins encoded in a genome in the absence of increases in gene number (Nilsen and Graveley 2010). As such, an association between alternative splicing and organism complexity has long been proposed. Under an "adaptive" model, an increase in alternative splicing could facilitate the evolution of higher organismal complexity, by increasing proteome diversity (and thus, diversifying functionality) at a level disproportionate to increases in the number of protein-coding genes (Graveley 2001; Xing and Lee 2007; Chen et al. 2012). Indeed, over the last decade, alternative splicing prevalence (ASP; the proportion of multi-exon genes that have at least one alternative splicing event) has been successively revised upward for humans, with recent deep sequencing transcriptome analyses estimating that up to 94% of multiexon genes undergo alternative splicing (Pan et al. 2008; Wang et al. 2008). However, assessing the expansion of ASP through evolutionary time and establishing a link between alternative splicing and organism complexity have proved difficult (Nilsen and Graveley 2010). The main barrier to comparative studies of ASP arises from the fact that differences in transcript sequence coverage across species can distort both the proportion of genes classified as undergoing alternative splicing and the number of alternative splicing events detected (Brett et al. 2002; Kim et al. 2004; Kim et al. 2007; Takeda et al. 2008; Mollet et al. 2010; Nilsen and Graveley 2010; Schad et al. 2011). Kim et al. (2007) devised a method of transcript number normalization to obtain comparable ASP indices involving the identification of alternative splicing events from a random sample of 10 transcripts per gene. Importantly, they showed that alternative splicing in vertebrate species was higher than among invertebrates and that this was not explained by the higher abundance of transcripts available for vertebrate species. Although not directly tested, these findings were suggestive of a link between alternative splicing and complexity as vertebrates are generally considered to have a higher CTN compared with invertebrates. Surprisingly, there are still no current data sets for comparable alternative splicing indices, and controlling for transcript abundance in comparative analyses of ASP is the exception rather than the rule. The resulting lack of comparable estimates for the number of alternative splicing events per gene has hampered efforts to quantify ASP across taxa (Harrison et al. 2002), the accumulation of splicing events over time (Warnefors and Eyre-Walker 2011), and the link between alternative splicing rates and organism complexity (Nilsen and Graveley 2010; Xue et al. 2012). The only attempt to directly assess the relationship between alternative splicing variation and CTN (Schad et al. 2011) was considered inconclusive by the authors because of the lack of comparable alternative splicing measures.

Here, we assess the prevalence of alternative splicing in 47 eukaryotic genomes by calculating a comparable index of alternative splicing, which corrects for differences in transcript coverage (adapted from Kim et al. [2007]; see Materials and Methods). The species examined include metazoans, plants, fungi, and protists. We then examined how these alternative splicing indices relate to organism complexity and compared the strength of alternative splicing as a predictor of CTN to previously described correlates, including the number of protein-interacting domains encoded per gene (Xia et al. 2008), protein disorder (Romero et al. 2006; Dunker et al. 2008; Schad et al. 2011; Xue et al. 2012), the number of PPIs, gene number, and various measures of coding region length (Schad et al. 2011).

We find that alternative splicing has steadily increased over the last 1,400 My of eukaryotic evolution. We also find that alternative splicing is strongly associated with CTN and that this relationship is not a by-product of the relationship between various genomic features and complexity.

It is important to note that if increases in the proportion of alternatively spliced genes or the level of alternative splicing these genes undergo are linked with CTN, such an association would not constitute proof of causality. Under a "nonadaptive" model, the association of alternative splicing and organism complexity could be a by-product of the link between complexity and a lower effective population size (N_e). The passive emergence of "genomic complexity" and even organismal complexity itself is suggested by the work of Lynch and coworkers, who argue that nonadaptive processes explain the majority of the variance in organism complexity as "more complex" organisms have a smaller N_e (Lynch and Conery 2003; Lynch 2007). As documented consequences of a comparatively small N_e include the accumulation of slightly deleterious mutations, both in coding (Nikolaev et al. 2007; Popadin et al. 2007; Gayral et al. 2013) and regulatory (Keightley et al. 2005) sequences, as well as an increase in average intron and coding region lengths (Lynch and Conery 2003), it is reasonable to expect that mutations impairing splicing regulation will accumulate more rapidly in more complex organisms resulting in higher (but not necessarily functional) transcript diversity. Consistent with this, single species studies have shown that a significant proportion of alternative splicing events are probably the result of noncoding "noise" and not biologically meaningful (Pickrell et al. 2010; Leoni et al. 2011).

Using a limited sample size, we do not find any evidence to suggest that the association of alternative splicing and CTN is explained by differences in N_e. To the best of our knowledge, this is the most comprehensive assessment of alternative splicing levels (ASLs) covering all major eukaryotic taxa, and the first time in which the link between alternative splicing and CTN has been assessed using a comparative index of alternative splicing which corrects for differential transcript coverage.

Results

ASP Has Increased throughout Evolutionary Time

To assess whether ASLs have changed over time, over 39 million publicly available partial transcripts, representing 112 eukaryotes (20 protists, 18 plants, 23 fungi and 51 metazoans including 23 chordates), were aligned to their corresponding genomes to identify alternative splicing events (see Materials and Methods). To minimize the strong dependence of alternative splicing event detection on transcript coverage per gene (Brett et al. 2002; Kim et al. 2004; Kim et al. 2007; Takeda et al. 2008; Mollet et al. 2010; Nilsen and Graveley 2010; Schad et al. 2011), we used a transcript normalization protocol (Kim et al. 2007) where alternative splicing events are identified in randomly selected samples of 10 expressed sequence tags (ESTs) per gene. We obtained a comparable alternative splicing index per gene by averaging the number of alternative splicing events in 100 samples (Kim et al. 2007) (supplementary fig. S1, Supplementary Material online).

Using the comparable alternative splicing index, we calculated for each species both ASP, defined as the proportion of alternatively spliced genes in the sample of genes analyzed, and ASL, defined as the average number of alternative splicing events per gene. Genomes with comparable alternative splicing estimates available for fewer than 500 genes were

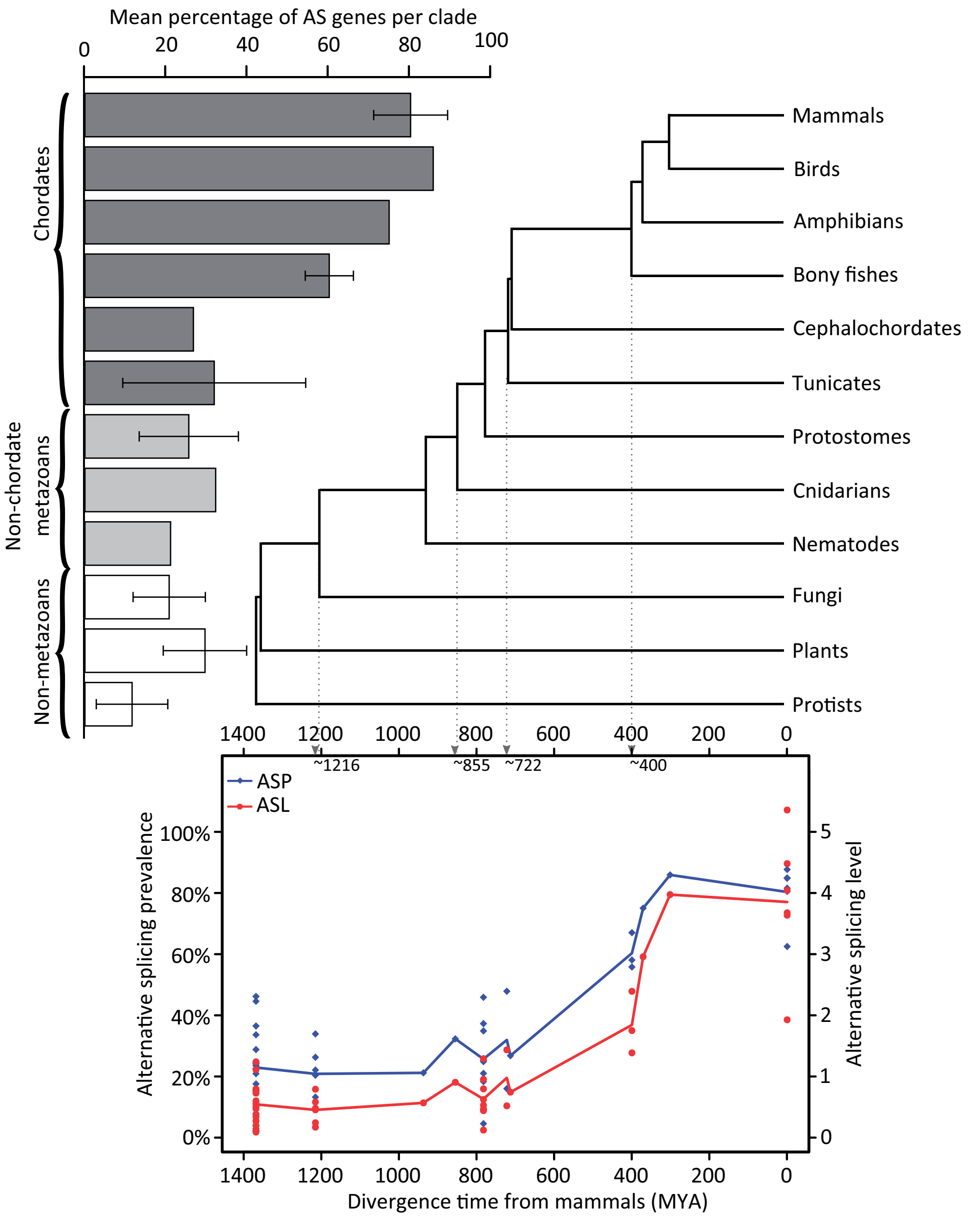

Fig. 1. Variance in alternative splicing over evolutionary time. Bars show the average percentage of alternatively spliced genes per species grouped according to their divergence from humans, as shown in the adjacent phylogenetic tree (data from Hedges et al. 2006), and their taxonomic category (chordate, nonchordate metazoan, or nonmetazoan). The scatter plot shows changes in alternative splicing prevalance, that is, the percentage of alternatively spliced genes per genome (blue) and in alternative splicing level, that is, the average number of alternative splicing events per gene for each species (red). Trend lines represent the mean of all values at each divergence time. Although the relative positions of cephalochordates and tunicates on this tree are disputed (Delsuc et al. 2006), this does not significantly alter the trend.

excluded from further analyses leaving, in total, 47 species (6 protists, 10 plants, 6 fungi, and 25 metazoans; supplementary table S1, Supplementary Material online). We found that both ASP and ASL vary among eukaryotic clades with chordates having both the highest ASP and ASL compared with nonchordate metazoans, fungi, plants, and protists (fig. 1 and supplementary table S1, Supplementary Material online). Although our ASP estimates are higher in most clades compared with a previous study based on eight species using comparable alternative splicing indices, the relative differences among clades are consistent (Kim et al. 2007).

An increase in alternative splicing through evolutionary time (fig. 1) is consistent with observations reporting links between ASP and evolutionary time restricted to metazoan species (Warnefors and Eyre-Walker 2011) and show that it is not an artifact of differential transcript coverage among species (Nilsen and Graveley 2010; Schad et al. 2011). The higher prevalence and levels of alternative splicing in plant species compared with fungi and protists suggest that AS levels have independently increased in this lineage.

Overall, by using comparable alternative splicing estimates from species covering all major eukaryotic clades and correcting for differential transcript coverage, we show that

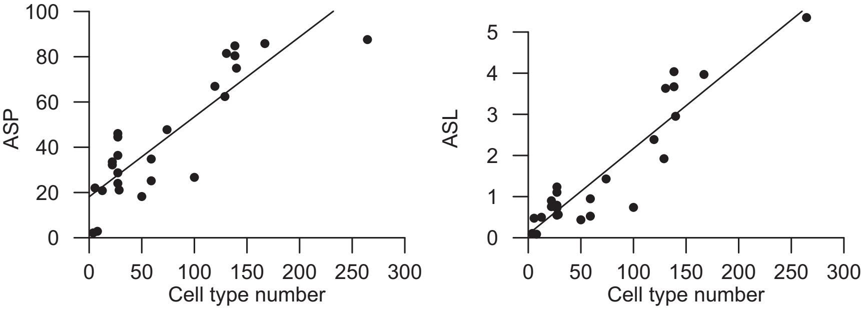

FIG. 2. Relationship between alternative splicing and organism complexity, assayed as CTN. Graphs show the relationship between CTN and ASP ($r^2 = 0.76$; $P = 9.36 \times 10^{-9}$) and ASL ($r^2 = 0.83$; $P = 1.77 \times 10^{-10}$).

alternative splicing has increased over the last 1,400 My of eukaryotic evolution in the metazoan lineage with a more moderate and potentially independent rise in alternative splicing in plants.

Alternative Splicing Is a Strong Predictor of Organism Complexity, Assayed as Cell Type Diversity

A previous attempt to assess the link between alternative splicing and organism complexity, assayed as the number of distinct cell types (Schad et al. 2011), was rendered inconclusive because of the known bias caused by differential transcript sequence coverage among genes and species (Brett et al. 2002; Kim et al. 2004; Kim et al. 2007; Takeda et al. 2008; Mollet et al. 2010; Nilsen and Graveley 2010; Schad et al. 2011). As such, we assessed the relationship of ASP and ASL with the number of distinct cell types per species (CTN) as a proxy of organism complexity using the comparable AS index (see Materials and Methods). We found that both ASL and ASP are strongly associated with CTN (ASP: $r^2 = 0.76$, $P = 9.36 \times 10^{-9}$; ASL: $r^2 = 0.83$, $P = 1.77 \times 10^{-10}$; supplementary table S2, Supplementary Material online, and fig. 2). This association remains strong when restricting the analyses to the metazoan-fungi lineage (for ASP, $r^2 = 0.71$, $P = 2.45 \times 10^{-5}$, and for ASL, $r^2 = 0.81$, $P = 1.28 \times 10^{-6}$; supplementary table S3, Supplementary Material online).

Several genomic and functional parameters have previously been associated with organism complexity (using CTN as a proxy). Xia et al. (2008) reported a strong link between CTN and PPI domain coverage. Other genomic variables found to have a more moderate association with CTN include protein disorder (Romero et al. 2006; Dunker et al. 2008; Schad et al. 2011; Xue et al. 2012) and proteome size (assayed as concatenated protein length) (Schad et al. 2011). Gene number, previously found to be unrelated to CTN, has recently been reconsidered as a significant predictor but only after plant genomes are excluded from the analyses (Schad et al. 2011).

How does alternative splicing compare to these previously reported predictors of CTN? To address this, we compared the relationship between CTN and alternative splicing with that of 12 additional genomic measures of protein interactivity as well as proteome disorder, gene length, and number, all previously linked to CTN (see Materials and Methods for descriptions and sources of each variable assessed). Of all parameters tested, ASL was found to have the strongest association with CTN ($r^2 = 0.83$, $P = 1.77 \times 10^{-10}$) followed by ASP and the average number of PPI domains per protein ($r^2 = 0.76$, $P = 9.36 \times 10^{-9}$ and $r^2 = 0.64$, $P = 8.19 \times 10^{-11}$ respectively; supplementary table S2, Supplementary Material online). We then re-examined the relationship between each parameter with CTN restricting the analyses to a set of 24 species for which data in all variables tested were available. The mean number of interactions per protein was not included in this or subsequent analyses due to the small number of species for which data were available ($n = 10$). ASL remained the top predictor of CTN ($r^2 = 0.87$, $P = 2.80 \times 10^{-11}$) with ASP showing an increased ($r^2 = 0.80$, $P = 2.66 \times 10^{-9}$) and the average number of PPI domains per protein a decreased association with CTN ($r^2 = 0.59$, $P = 6.42 \times 10^{-6}$; table 1).

As the relationship between genomic parameters and CTN has been shown to increase after the removal of plant genomes (Schad et al. 2011), we reassessed the predictive power of all parameters after restricting the analyses to the metazoan-fungi lineage. This resulted in a stronger association between CTN and many parameters with the two alternative splicing indices remaining the best predictors of CTN (supplementary table S3, Supplementary Material online). Consistent with previous findings (Schad et al. 2011), when plant genomes are excluded, gene number was found to be significantly associated with CTN ($r^2 = 0.34$, $P = 1.74 \times 10^{-3}$; supplementary table S3, Supplementary Material online).

Because of the tendency of related species to resemble one another, it is also necessary to control for this nonindependence in a comparative analysis of patterns across species. Pagel's λ measures the extent to which observed correlations between traits reflect their shared evolutionary history assuming an evolutionary model under Brownian motion (Pagel 1999). For the 24 species for which data in all variables tested were available, we obtained estimates of λ and restricted log likelihood for the correlations between CTN and each genomic variables, recalculating each correlation to account for phylogenetic nonindependence of the variables by fitting a phylogenetic generalized least squares (PGLS) model (see Materials and Methods). ASL remained the top predictor of CTN even after taking into account the strength of the phylogenetic signal ($r^2 = 0.87$, $P = 1.59 \times 10^{-13}$, $\lambda = 0$), followed by ASP ($r^2 = 0.77$, $P = 8.38 \times 10^{-11}$, $\lambda = 0.052$) and the percentage of PPI

Table 1. Association between CTN and Genomic Features Before and After Phylogenetic Signal Correction in 24 Eukaryotic Species.

Category	Variable	Linear Regression		PGLS Regression		
		r^2	P	r^2	P	λ
Alternative splicing	ASL	0.87	2.80×10^{-11}	0.87	1.59×10^{-13}	0
	ASP	0.80	2.66×10^{-9}	0.77	8.38×10^{-11}	0.05
Sizes and lengths	Number of genes	−0.01	0.40	0.26	1.23×10^{-3}	0.76
	Average protein length	−0.05	0.97	0.12	0.03	0.79
	Proteome information content	3.25×10^{-3}	0.31	0.09	0.05	0.65
	Proteome size	0.31	2.59×10^{-3}	0.49	4.08×10^{-6}	0.75
Disorder	Mean % of disordered binding sites	−0.03	0.59	0.02	0.26	0.71
	Mean number of disordered binding sites	−0.04	0.78	−0.04	0.99	0.68
	Total number of disordered binding sites	0.04	0.18	0.21	3.97×10^{-3}	0.69
	Mean proteome disorder	−0.03	0.64	6.45×10^{-3}	0.34	0.71
Interactivity	% PPI domain seq per protein	0.60	5.36×10^{-6}	0.60	1.30×10^{-7}	0
	Average number of PPI domains per protein	0.59	6.42×10^{-6}	0.59	1.61×10^{-7}	0
	Proportion of proteins with 1 + PPI domains	0.54	2.33×10^{-5}	0.54	7.80×10^{-7}	0

domain sequence per protein ($r^2 = 0.60$, $P = 1.3 \times 10^{-7}$, $\lambda = 0$; table 1). This pattern holds true if we only take into account metazoan and fungal species (supplementary table S3, Supplementary Material online).

As most of the assessed parameters covary among themselves (supplementary tables S4 and S5, Supplementary Material online), the association between some variables with CTN may be secondary to their covariance with another genomic feature which is in turn linked to CTN. To identify the genomic parameters that significantly contribute to CTN, we carried out a stepwise analysis (see Materials and Methods). In the optimal stepwise regression model, the majority of the variance in CTN is explained by ASL, supported by proteome size (supplementary table S6, Supplementary Material online). Similar results are obtained when constraining the data to the metazoan-fungal lineage (supplementary table S6, Supplementary Material online). In fact, contrasting each variable directly against AS by including ASL/ASP in multiple regression models with each additional variable revealed that in all cases, only the AS parameter remained significantly associated with CTN (supplementary table S2, Supplementary Material online). The only exception was proteome size that remained significantly associated with CTN after correcting for either ASP or ASL, but only when fungi and metazoans were included in the analysis (supplementary table S3, Supplementary Material online).

To best capture the predictive value of sets of covarying variables, we used a principal component analysis to reduce the dimensionality among the 13 predictors of complexity. This analysis was performed on a subset of species where data were available for all predictors ($n = 24$). Interestingly, PC1 and PC2 (which explain 35.2% and 31.4% of the variance in the matrix, respectively) allow chordates to be differentiated from all other species (fig. 3). Of all resulting principal components, we found that PC1 is the only significant predictor of CTN ($r^2 = 0.66$, $P = 8.58 \times 10^{-7}$). The two alternative splicing variables (ASP and ASL) and the three protein interactivity variables (average number of PPI domains per protein, PPI domain coverage, and the proportion of proteins with at least one PPI domain) were found to be the main contributors to PC1. Similar results were obtained when restricting the analyses to the metazoan-fungi lineage (data not shown). It is worth noting, however, that the value of r^2 when regressing PC1 against CTN, when including either all species or only metazoans and fungi, is lower than that of ASL ($r^2 = 0.83$, $P = 1.77 \times 10^{-10}$), suggesting that collapsing the dimensionality of the variables does not improve the prediction of CTN beyond the variance explained by ASL alone.

The above results show that AS is significantly associated with CTN and that this association is not explained as a by-product of the relationship between AS and other genomic features also related to CTN. However, it is possible that some of these associations might be explained by ascertainment bias resulting from the fact that humans and other closely related species have been disproportionately studied. With the exceptions of *Caenorhabditis elegans* and *Drosophila melanogaster*, larger amounts of data exist for vertebrates than other species. It is possible that the higher estimates of AS and other genomic features, and even higher CTN among vertebrates, might partly result from the greater availability of data for these species. To address this possibility, we used the total number of ESTs per species as a proxy for interest in a species as higher transcript availability has a direct impact on the quality of genome annotation. Compared with other proxies of "research interest" such as "number of publications per species," the number of ESTs approximates how much data have accumulated rather than how many interpretations of it there have been.

We established that the number of ESTs per species is significantly associated with various genomic characteristics (supplementary table S7, Supplementary Material online). Notably, ASL and ASP, as well as CTN, were found to be significantly related with transcript number per species (ASL: $r^2 = 0.45$, $P = 7.29 \times 10^{-7}$; ASP: $r^2 = 0.39$, $P = 8.01 \times 10^{-6}$; complexity $r^2 = 0.41$, $P = 5.01 \times 10^{-5}$). Thus, we re-examined the relationship of CTN with AS and other gene features using the residuals of a quadratic polynomial regression with EST number. This correction resulted in a marked reduction in the variance in CTN explained by ASL and ASP ($r^2 = 0.47$, $P = 9.84 \times 10^{-5}$ and $r^2 = 0.57$,

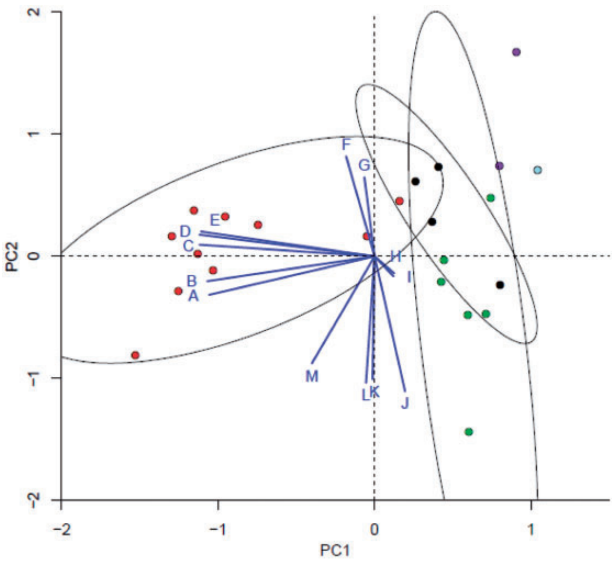

Fig. 3. Biplot of the first two principal components built from 13 functional genomic variables available for 24 species (see supplementary table S1, Supplementary Material online). Graph shows the distribution of species along PC1, which explains 35.2% of the variance in this data set, and PC2, which accounts for 31.4%. Points represent each of 24 species for which data were available for all variables and are colored by taxonomic category: chordates (red), nonchordate metazoans (black), plants (green), fungi (blue), and protists (purple). Ellipses show the clustering of species according to their dispersion along PC1 and PC2 (with confidence limit 0.95). Blue lines radiating from (0,0) represent each variable included in the analysis. The direction of each line represents the highest correlation coefficient between the PC scores and the variable, with the length of each line proportional to the strength of this correlation. Letter codes for each variable: ASL (A), ASP (B), % PPI domain sequence per protein (C), proportion of proteins with at least one PPI domain (D), average number of PPI domains per protein (E), average protein length (F), mean number of disordered binding sites per protein (G), mean proteome disorder (H), mean % of disordered binding sites per protein (I), number of genes (J), total number of disordered binding sites per proteome (K), proteome information content (L), and proteome size (M).

$P = 8.82 \times 10^{-6}$, respectively; supplementary table S8, Supplementary Material online). Correcting all variables by transcript coverage also reduced the predictive value of other gene features for CTN (supplementary table S8, Supplementary Material online). However, the relative order of gene feature parameters as predictors of CTN remained unaltered with splicing and, to a lesser extent, the degree of protein–protein interactivity the most strongly associated with CTN (supplementary table S8, Supplementary Material online). Furthermore, if considering all 13 variables, the optimal stepwise regression model (see Materials and Methods) explained 90% of the variance in CTN ($P = 1.81 \times 10^{-5}$), with the strongest of five predictors being ASP (supplementary table S9, Supplementary Material online). When restricting the analyses to the fungi-metazoan lineage, we found that the optimal regression model contained only two regressors, ASP and the mean percentage of disordered binding sites per protein (see Materials and Methods for a description of this variable) (supplementary table S9, Supplementary Material online). In fact, only three parameters (average protein length, the number of genes, and the total number of disordered binding sites per protein) remained significantly associated with CTN in a regression model directly comparing each variable with either ASP or ASL (supplementary table S8, Supplementary Material online). An alternative transformation of the data, taking the natural log of EST number, resulted in lower correlation coefficients, but the relative strength of each variable in a regression against complexity remained unchanged (supplementary table S10, Supplementary Material online).

Our data span a diverse range of species with associated variations in the number of available ESTs per species (supplementary table S1, Supplementary Material online). For genomes with lower EST numbers (often those that also have a lower CTN), highly expressed genes will make a disproportionate contribution to each species' comparative alternative splicing index as the number of genes with the minimum required number of ESTs will be smaller. As such, we expect lowly expressed genes to primarily contribute data for genomes with a higher number of available ESTs.

Under the nonadaptive model, a reduced N_e among more complex organisms (assayed as those with higher CTN) would result in an accumulation of mutations detrimental to splicing regulation, potentially resulting in the proliferation of "noisy" alternative splicing events. Such neutral increases in alternative splicing should be particularly pronounced among lowly expressed genes, which, on average, are under lower selective pressures compared with highly expressed genes. Importantly for this study, if lowly expressed genes are more highly spliced, then our data would overestimate ASL for species with high EST numbers, artificially inflating the correlation strength with CTN.

Using microarray data for four model species (human, mouse, *Caenorhabditis elegans*, and *Arabidopsis thaliana*; see Materials and Methods), we find that, as expected, there is a strong correlation between the number of ESTs per gene and gene expression level. However, contrary to the prediction of the nonadaptive model, we found that the more highly expressed genes are also more highly spliced (supplementary figs. S2–S5, Supplementary Material online). Therefore, our data might be underestimating ASP and ASL in genomes with a higher number of available ESTs, as more lowly expressed genes—with lower ASLs—disproportionately contribute to the species' alternative splicing indices. By extension, the relationship of AS with CTN might also be underestimated.

Discussion

Here, we have assessed ASLs in 47 eukaryotic species and showed that alternative splicing has increased over the last 1,400 My of evolution. Our data range from *Plasmodium falciparum*, in which 3% of genes are spliced with an average of 0.09 splice events per gene, to humans, where 88% of genes are spliced with an average of 5.35 splice events per gene. Consistent with the findings of Kim et al. (2007), we find that chordates have higher levels of alternative splicing than any other taxonomic group with mammals and birds having both proportionately more genes that are alternatively spliced (ASP) and a higher number of alternative splicing events per gene (ASL). We observed significant increases over time in ASP and ASL for the opisthokonts and show that past claims for an increased level of alternative splicing along the evolution of metazoans are not explained by differential transcript coverage (Warnefors and Eyre-Walker 2011). Our data do not support a previous claim for lower ASLs among birds compared with mammalian species (Chacko and Ranganathan 2009), and in fact, ASLs in the chicken genome were found to be among the highest of all species tested.

Plant genomes were found to have higher levels of alternative splicing than both protist and fungal species, comparable to those found among invertebrate species. This is consistent with relatively low levels of alternative splicing in the eukaryotic ancestor with independent rises in the plant and metazoan lineages. None of the plant genomes we examined, however, match the levels of alternative splicing observed in the vertebrate lineage.

Our results demonstrate a strong association between alternative splicing and organism complexity providing, to the best of our knowledge, the first systematic evidence for a link between these two variables. In this study, we have used the number of cell types as a proxy for organism complexity. CTN has been proposed as an indicator of an organism complexity as the higher number of components or cell types in more complex organisms should reflect, to some degree, their higher number of functions (McShea 2000). We acknowledge, however, that complexity is difficult to define and even more difficult to measure and that all operational definitions for "complexity" are, to various degrees, contentious (Adami 2002). Several proxies of organismal complexity have been proposed; however, these measures are either relevant to some taxonomic groups, such as encephalization coefficient, or no measurements are available for a large number of species, such as phenotypic complexity (Tenaillon et al. 2007). Although accepting that "organism complexity" is likely to be a multidimensional variable encompassing many other features, we chose this measure as, compared with other proxies, cell types are more easily quantifiable for organisms from distant taxonomic groups. It is important to note that, as CTN data are drawn from a diverse range of studies (see Materials and Methods), more detailed characterizations of CTN can appear anomalous. For example, we expect chimpanzees to have a similar CTN to humans, but currently, humans are the better characterized species and as such the human CTN appears higher (supplementary table S1, Supplementary Material online). To address whether this type of outlier confounds our results, we repeat our analyses using the average CTN for the order each species belongs to. This makes the assumption that any variation in CTN between species of a given order reflects measurement noise, rather than relevant biological information. Our results do not significantly differ when using these alternate values of CTN (supplementary tables S11 and S12, Supplementary Material online).

Importantly, as most past studies analyzing the relationship between various genomic features and organism complexity have adopted CTN as a proxy (Xia et al. 2008; Chen et al. 2011; Schad et al. 2011; Xue et al. 2012), its use allowed us to contrast our results with those of others. Such comparisons showed that the relationship of alternative splicing and CTN is not secondary to other genomic features previously associated with CTN, including proteome size (measured as total protein coding sequence length [Schad et al. 2011]), protein disorder (Schad et al. 2011; Xue et al. 2012), and protein interactivity.

Before the full sequencing of nuclear eukaryotic genomes became widespread, gene number was expected to have a direct relationship with organism complexity as more genes would encode a higher number of proteins boosting the number of potential molecular interactions (Romero et al. 2006; Dunker et al. 2008). The sequencing of the human genome, however, found no evidence for such an association (Fields et al. 1994). The discrepancy between organism complexity and gene content became known as the G-paradox (Claverie 2001; Betran and Long 2002; Hahn and Wray 2002; Taft and Mattick 2003). However, a recent study concluded that gene number and organism complexity are related after all, albeit only when plant species are removed from the analyses (Schad et al. 2011).

Our findings also support a significant association between gene number and CTN in the absence of plant genomes ($r^2 = 0.34$, $P = 1.74 \times 10^{-3}$; supplementary table S3, Supplementary Material online). However, ASL has a stronger association with CTN ($r^2 = 0.77$, $P = 1.09 \times 10^{-8}$) and is sufficient to explain the relationship between CTN and gene number.

Unlike alternative splicing and gene number, which directly impact on the number of interacting proteins, additional gene features linked to CTN can boost the interactivity potential of individual proteins without expanding their number. One of the simplest measures of the functional potential of the proteome, total coding region length, has been found to be significantly associated with CTN (Schad et al. 2011). Although we observed a similar association between proteome size and CTN, this relationship is entirely explained as a by-product of both variables' covariance with alternative splicing. Proteome size remains a marginal, albeit significant, predictor of CTN in a stepwise regression model restricted to the metazoan and fungi lineage where ASL was the strongest variable (table 1). Moreover, proteome size was not a significant contributor to the only principal component found to be significantly associated with CTN.

Protein disorder—the lack of equilibrium in a protein's 3D structure under physiological conditions (Romero et al. 2006)—has been proposed as a candidate predictor of organism complexity as higher intrinsic disorder allows individual proteins to adopt a greater variety of conformations, increasing the average number of interacting partners per protein and potentially boosting functional diversification of the proteome (Romero et al. 2006; Dunker et al. 2008). Nevertheless, subsequent findings show the association between disorder and CTN only explains any substantial amount of variance when bacterial species are included (Schad et al. 2011; Xue et al. 2012). Our analyses of protein disorder using both stepwise regressions and principal component analysis do not provide any evidence of hidden covariance between protein disorder and CTN. Moreover, despite the fact that past studies have found a higher than expected number of disordered motifs in alternatively spliced areas at the gene level (Romero et al. 2006; Buljan et al. 2012), we do not find a significant association between protein disorder and alternative splicing per species (supplementary tables S4 and S5, Supplementary Material online).

Finally, a third measure of potential molecular interactions per protein, the presence of PPI domains, has been shown to be strongly associated with CTN (Xia et al. 2008). We found three protein interactivity parameters—PPI domain coverage, the average number of PPI domains per protein, and the proportion of proteins with at least one PPI domain—to be significantly associated with CTN regardless of the set of species examined (supplementary tables S2 and S3, Supplementary Material online). A head-to-head comparison between predictors of CTN showed that protein interactivity measures are better predictors of CTN than any other variable with the exception of alternative splicing. After controlling for alternative splicing, however, no protein interactivity parameter was found to be significantly associated with CTN (supplementary tables S2 and S3, Supplementary Material online). An additional measure of protein interactivity previously associated with CTN, the mean number of PPIs (Schad et al. 2011), was not included in most of our analyses as data were limited to only 10 species in our set. These comparisons show that although protein interactivity is significantly associated with CTN, there is a great overlap between the variance in CTN explained by protein interactivity and that explained by alternative splicing.

Several studies have proposed an association between alternative splicing and protein domain content, suggesting that alternative splicing could act as a buffer against reduced functionality because of "domain overload"—too many protein domains or domains in the wrong combination (Kriventseva et al. 2003; Resch et al. 2004; Floris et al. 2008). A large-scale analysis has shown that protein domains are nonrandomly combined in functional proteins with fewer protein domain co-occurrences observed than expected, suggesting that certain protein domains "avoid" each other (Parikesit et al. 2011), whereas other domains—including PPI domains—are "promiscuous" and tend to coexist within individual transcripts (Basu et al. 2008). Our analyses of covariance among functional gene variables showed that alternative splicing and PPI measures are positively correlated—genomes with higher levels of alternative splicing also have a higher PPI domain presence. We further examined the association between ASL and PPI domain coverage within species but found only a marginal association between the two variables constrained to a few species (supplementary table S13, Supplementary Material online). This finding suggests that although genomes with a high level of alternative splicing also tend to have a higher PPI domain coverage, there is no support for a role for alternative splicing acting as a buffer of PPI domain overload.

Overall, our results are consistent with a direct association between alternative splicing and CTN, one which is not explained by other genomic features previously associated with organism complexity. This finding is, in principle, consistent with previous suggestions that alternative splicing may underlie the rise in complexity during eukaryotic evolution thanks to its potential to expand transcript diversity and thereby increase the number of potential molecular interactions and functions (reviewed in Xing and Lee 2007; Nilsen and Graveley 2010; Chen et al. 2012).

Nevertheless, it is important to note that the rise in CTN has been accompanied by a reduction in effective population size (Lynch and Conery 2003). Classical nearly neutral theory proposes that as effective population sizes diminish so too does the efficiency of purifying selection, resulting in the accumulation of slightly deleterious mutations, both in coding (Nikolaev et al. 2007; Popadin et al. 2007; Gayral et al. 2013) and regulatory (Keightley et al. 2005) sequences. The increased role of drift relative to selection has also been invoked to explain the proliferation of a number of genomic features among increasingly complex species (Lynch and Conery 2003; Lynch 2007). Although more recent studies have disputed this conclusion (Kuo et al. 2009; Whitney and Garland 2010; Whitney et al. 2011), a significant proportion of alternative splicing events have nevertheless been suggested to result from noisy alternative splicing (Sorek et al. 2004; Pickrell et al. 2010; Leoni et al. 2011). Thus, it is possible that the observed increase in alternative splicing among more complex species might be the result of increased genetic drift as a result of reductions in effective population size, rather than being directly associated with organism complexity. Using estimates of effective population size for the 12 species represented in this study (Lynch and Conery 2003), we found that a genome's capacity for alternative splicing remains strongly correlated with CTN even after controlling for effective population size (partial Spearman's correlation coefficients: ASL = 0.71, $P = 2.37 \times 10^{-3}$; ASP = 0.70, $P = 3.35 \times 10^{-3}$). Although based on a small sample of species, this finding suggests that the association between CTN and alternative splicing is not a by-product of reduced effective population sizes among more complex species. Future studies should be able to assess the functional contribution of increases in alternative splicing in the eukaryotic lineages we report here.

In addition, it is worth noting that a significant correlation of any genomic feature with CTN does not necessarily demonstrate a causal role on the evolution of organism complexity, that is, a higher CTN. It is beyond the scope of this study to address this directly. Nevertheless, network theory provides some clues, which allows us to speculate as to the likelihood that increases in transcript diversification, facilitated by alternative splicing, have affected the evolution of organism complexity. Boolean networks have been proposed as models for genetic networks as the attractors, representing different stable patterns of gene expression, correspond to different cell types (Kauffman 1969; Serra et al. 2010). In Boolean networks, increases in the number of nodes leads to a higher number of attractors within the network at a rate equal to or exceeding the square root of the number of nodes in the network (Samuelsson and Troein 2003). If we imagine each distinct transcript as a node in the genetic network, we can speculate that alternative splicing, by increasing the number of nodes (transcripts), would lead to an increased number of attractors (cell types). Indeed, a previous study that generated relational networks for seven species associated the number of functions in a proteome with the number of polyform transcriptional units in the genome, those that produce protein isoforms with different functional assignments (which are strongly associated with the levels of splicing). Various properties of these networks (such as the number of nodes) were found to be strongly associated with organism complexity, suggesting a link between splicing and both multifunctionality and multicellularity (Kanapin et al. 2010).

We conclude that alternative splicing increases over the last 1,400 My of eukaryotic evolution are strongly associated with CTN. Furthermore, this association is stronger and more robust than other parameters previously associated with CTN, although we cannot rule out the contributions of other genomic features as many covary. Our findings are consistent with an adaptive scenario whereby a genome's capacity for alternative splicing—with its resulting expansion of the transcript pool—could constitute a critical component of the underlying mechanisms explaining the diversification of cell types and the rise in organism complexity over time. Nevertheless, the data here presented do not allow us to reach a conclusion on the functional relevance of increases in alternative splicing or to establish causality regarding the association of alternative splicing and organism complexity; thus, it is possible that a "nonadaptive model" may account for it.

To the best of our knowledge, our results represent the first systematic assessment of the relationship between alternative splicing, evolutionary time, and CTN and provide evidence for a strong association of alternative splicing and CTN. Our results further constitute the most comprehensive head-to-head comparison, to date, of variables associated with CTN.

Materials and Methods

Organism Complexity

The number of unique cell types was used as a proxy of organism complexity. Estimates of CTN per species were compiled from previous studies (Valentine et al. 1994; Bell and Mooers 1997; Hedges et al. 2004; Haygood and Investigators 2006; Lang et al. 2010; Schad et al. 2011); data in graph form from Valentine et al. (1994) as interpreted by both Erwin (2009) and Vogel and Chothia (2006) were also included. Following the methodology of Vogel and Chothia (2006), where more than one estimate of CTN was available for a species, the average of the minimum and maximum number was used. In addition, we included a revised CTN estimate for humans (Vickaryous and Hall 2006). Supplementary table S1, Supplementary Material online, provides averaged complexity estimates for both pro- and eukaryotic species, whereas supplementary table S14, Supplementary Material online, shows the sources.

Identification of Alternative Splicing Events

Comparable alternative splicing events were obtained using the following approach. Over 39 million EST sequences, accounting for 112 species, were downloaded from dbEST (Boguski et al. 1993) and matched to their corresponding genome using GMAP (Wu and Watanabe 2005) (these species are identified in supplementary table S1, Supplementary Material online, by a positive value in the column titled "total number of ESTs"). Genome sequences and annotations were obtained from sources contained in supplementary table S1, Supplementary Material online. Cancer-derived EST libraries from human and mouse were removed from all analyses presented. To ensure high-quality alignments, we only retained those ESTs with 95% identity. ESTs were assigned to genes using gene annotation coordinates. EST alignments were then used to create an exon template. These templates were generally in agreement with existing exon annotations and also identify a small number of nonannotated exons and discard orphan exons likely to be nested genes. Alternative splicing events per gene were identified by comparing alignment coordinates for each individual EST to exon annotations. A comparable alternative splicing index that avoids transcript coverage biases was obtained using the transcript normalization method described by Kim et al. (2007). Briefly, for each gene with greater than 10 ESTs, 100 random samples of 10 ESTs were selected. The number of alternative splicing events were calculated for each random sample (as detailed earlier), with an overall average calculated per gene. The ability of this method to correct for transcript coverage bias and calculate an accurate number of alternative splicing events is shown in supplementary figure S1, Supplementary Material online. To estimate ASP, a gene was considered to be alternatively spliced if it had at least an average of one alternative splicing event identified in each of the 100 random samples.

Additional Functional Genomic Parameters

Gene number per species was obtained from Ensembl BioMart version 0.8 (March 2013) (Kinsella et al. 2011). Proteome size (total amino acids encoded by all peptides), proteome information content (total amino acids encoded by primary transcripts only), and average protein length were calculated from mRNA transcripts obtained from Ensembl BioMart version 0.8 (March 2013) (Kinsella et al. 2011). The exception is the lancelet, *Branchiostoma floridae*, where transcripts were obtained from Putnam et al. (2008). PPI domains per protein were identified using HMMER3 with default parameters (Finn et al. 2011) and the Pfam-A database (Finn et al. 2008), with results parsed to consider matches to the 642 confirmed PPI domains as described by Xia et al. (2008). Protein disorder data were obtained from Schad et al. (2011). "Disordered sites" are those which are not at equilibrium in the protein's 3D structure under physiological conditions and can thus adopt a greater variety of conformations. We obtained the mean number of disordered binding sites per protein, the total number of disordered binding sites across all annotated proteins per species, and the mean percentage of disordered binding sites per protein (i.e., the mean number of disordered sites per protein as a percentage of the protein's length). The latter term is considered the disorder of the protein. Mean proteome disorder is taken as the mean disorder per protein. The average number of PPIs per protein for each species was also obtained from Schad et al. (2011). Data on effective population size were obtained from Lynch and Conery (2003).

Statistical Analysis

All statistical tests were performed in R, version 2.15.2 (Team 2012). For stepwise regression analysis, new regressors are included at each step according to the Akaike Information Criterion (Akaike 1974), estimated using the package "MASS" (Venables and Ripley 2002). Principal component analysis was performed using the R packages "FactoMineR" (Lê et al. 2008) and "Vegan."

Correction for Phylogenetic Autocorrelation

To assess and control for the strength of the phylogenetic signal on the correlation between CTN and the different genomic variables in this study, we computed Pagel's λ (Pagel 1999) based on maximization of the restricted log-likelihood using the gls subroutine from the R-package nlme (Pinheiro et al. 2013). Optimum negative values of Pagel's λ are reported as $\lambda = 0$. We used the subroutine PGLS in the R-package Caper (Orme et al. 2012) to examine the "true" associations between the different genomic variables and CTN after using the optimal λ values to control for the strength of the phylogenetic signal. This method implements generalized least squares models, which account for phylogenetic nonindependence by incorporating the covariance between taxa into comparisons that determine the correlation between dependent and independent variables. PGLS is an extension of the independent contrasts methods proposed by Felsenstein (1985) that provides a more general and flexible approach for assessing correlations between traits while accounting for phylogenetic divergence. An ultrametric phylogenetic tree for the analyzed species was created by obtaining the divergence time between each pair of species from Hedges et al. (2006).

Expression Level

Microarray data for four species (*Homo sapiens*, *Mus musculus*, *A. thaliana*, and *C. elegans*) were obtained from the following sources. For *H. sapiens* and *M. musculus*, Affymetrix array data analyzed by Su et al. (2004) was obtained from BioGPS (http://biogps.org, last accessed November 21, 2013). For *H. sapiens*, we obtained the expression of 11,449 genes across 28 tissues. We summarized gcRMA (GC robust multiarray average) normalized probe intensity levels to Ensembl IDs corresponding to protein coding genes. All probes matching to more than one Ensembl gene ID were removed. Expression values were then normalized against the total signal level in each tissue. For *M. musculus*, we obtained 9,825 genes with one-to-one orthologs in the human across 79 different tissues and cell types. Where more than one array exists for a given tissue, data were averaged. The per probe expression signal was summarized to Ensembl gene IDs using the average expression of all the probe sets matching a single Ensembl ID. All probes matching to more than one Ensembl gene ID were removed. Expression values were then normalized against the total signal level in each tissue. For *A. thaliana*, data were obtained from the Arabidopsis Development Atlas, as generated by the AtGenExpress Consortium (Schmid et al. 2005) (NASCARRAYS reference numbers 149–154, together representing 79 tissues, were downloaded from NASC AffyWatch [http://affymetrix.arabidopsis.info/, last accessed November 7, 2011]). Expression level was then quantified as the average gcRMA across all 79 tissues (with each value itself the mean of triplets) (Yang and Gaut 2011). For *C. elegans*, tissue-specific expression for 13 tissues (germline, hypodermis, intestine, muscle, neurons, pharynx, coelomocytes, distal tip, excretory cells, spermatheca, spermatheca uterine valve, uterus, and vulva) was obtained from Chikina et al. (2009) (http://worm-tissue.princeton.edu, last accessed November 28, 2013), who analyzed a compendium of 916 microarray experiments from 53 data sets. Expression values in this data set are already normalized to have mean 0 and variance 1. Expression level is taken as the mean across all tissues.

Supplementary Material

Supplementary tables S1–S14 and figures S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Humberto Gutierrez, Laurence Hurst, Nicolas Galtier, and an anonymous reviewer for comments. This work was supported by a UK-China scholarship for excellence and University of Bath research studentship to L.C., University of Bath fee studentship to S.J.B., CONACyT scholarships to J.M.T.-C. and A.C.M., a Royal Society Dorothy Hodgkin Research Fellowship (DH071902), Royal Society research grant (grant number RG0870644), and a Royal Society research grant for fellows (grant number RG080272) to A.O.U.

References

Adami C. 2002. What is complexity? *Bioessays* 24:1085–1094.

Akaike H. 1974. A new look at the statistical model identification. *Automatic Control IEEE Trans.* 19:716–723.

Basu MK, Carmel L, Rogozin IB, Koonin EV. 2008. Evolution of protein domain promiscuity in eukaryotes. *Genome Res.* 18:449–461.

Bell G, Mooers AO. 1997. Size and complexity among multicellular organisms. *Biol J Linn Soc.* 60:345–363.

Betran E, Long M. 2002. Expansion of genome coding regions by acquisition of new genes. *Genetica* 115:65–80.

Bird AP. 1995. Gene number, noise reduction and biological complexity. *Trends Genet.* 11:94–100.

Boguski MS, Lowe TMJ, Tolstoshev CM. 1993. dbEST—database for expressed sequence tags. *Nat Genet.* 4:332–333.

Brett D, Pospisil H, Valcarcel J, Reich J, Bork P. 2002. Alternative splicing and genome complexity. *Nat Genet.* 30:29–30.

Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM. 2012. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell.* 46:871–883.

Chacko E, Ranganathan S. 2009. Comprehensive splicing graph analysis of alternative splicing patterns in chicken, compared to human and mouse. *BMC Genomics* 10:S5.

Chen CH, Lin HY, Pan CL, Chen FC. 2011. The plausible reason why the length of 5′ untranslated region is unrelated to organismal complexity. *BMC Res Notes.* 4:312.

Chen L, Tovar-Corona JM, Urrutia AO. 2012. Alternative splicing: a potential source of functional innovation in the eukaryotic genome. *Int J Evol Biol.* 2012:10.

Chikina MD, Huttenhower C, Murphy CT, Troyanskaya OG. 2009. Global prediction of tissue-specific gene expression and context-dependent gene networks in *Caenorhabditis elegans*. *PLoS Comput Biol.* 5:e1000417.

Claverie J-M. 2001. What if there are only 30,000 human genes? *Science* 291:1255–1257.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* 3:e314.

Delsuc F, Brinkmann H, Chourrout D, Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* 439:965–968.

Dunker AK, Oldfield CJ, Meng J, Romero P, Yang JY, Chen JW, Vacic V, Obradovic Z, Uversky VN. 2008. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* 9(Suppl 2):S1.

Erwin DH. 2009. Early origin of the bilaterian developmental toolkit. *Philos Trans R Soc Lond B Biol Sci.* 364:2253–2261.

Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat.* 125:1–15.

Fields C, Adams MD, White O, Venter JC. 1994. How many genes in the human genome? *Nat Genet.* 7:345–346.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39:W29–W37.

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz H-R, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, et al. 2008. The Pfam protein families database. *Nucleic Acids Res.* 36:D281–D288.

Floris M, Orsini M, Thanaraj T. 2008. Splice-mediated variants of proteins (SpliVaP)—data and characterization of changes in signatures among protein isoforms due to alternative splicing. *BMC Genomics* 9:453.

Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, Nabholz B, Lourenco JM, Alves PC, Ballenghien M, Faivre N, et al. 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. *PLoS Genet.* 9:e1003457.

Graveley BR. 2001. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17:100–107.

Hahn MW, Wray GA. 2002. The g-value paradox. *Evol Dev.* 4:73–75.

Harrison PM, Kumar A, Lang N, Snyder M, Gerstein M. 2002. A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.* 30:1083–1090.

Haygood R. 2006. Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. Mutation rate and the cost of complexity. *Mol Biol Evol.* 23:957–963.

Hedges S, Blair J, Venturi M, Shoe J. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol Biol.* 4:2.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22: 2971–2972.

Kanapin AA, Mulder N, Kuznetsov VA. 2010. Projection of gene-protein networks to the functional space of the proteome and its application to analysis of organism complexity. *BMC Genomics* 11(Suppl 1):S4.

Kauffman SA. 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol.* 22:437–467.

Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* 3:e42.

Kim E, Magen A, Ast G. 2007. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.* 35:125–131.

Kim H, Klein R, Majewski J, Ott J. 2004. Estimating rates of alternative splicing in mammals and invertebrates. *Nat Genet.* 36:915–916.

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011:bar030.

Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S. 2003. Increase of functional diversity by alternative splicing. *Trends Genet.* 19:124–128.

Kuo CH, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* 19: 1450–1454.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

Lang D, Weiche B, Timmerhaus G, Richardt S, Riano-Pachon DM, Correa LG, Reski R, Mueller-Roeber B, Rensing SA. 2010. Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol Evol.* 2:488–503.

Lê S, Josse J, Husson F. 2008. FactoMineR: an R package for multivariate analysis. *J Stat Softw.* 25:1–18.

Leoni G, Le Pera L, Ferre F, Raimondo D, Tramontano A. 2011. Coding potential of the products of alternative splicing in human. *Genome Biol.* 12:R9.

Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A.* 104:8597–8604.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* 302:1401–1404.

McShea DW. 2000. Functional complexity in organisms: parts as proxies. *Biol Philos.* 15:641–668.

Mollet IG, Ben-Dov C, Felício-Silva D, Grosso AR, Eleutério P, Alves R, Staller R, Silva TS, Carmo-Fonseca M. 2010. Unconstrained mining of transcript data reveals increased alternative splicing complexity in the human transcriptome. *Nucleic Acids Res.* 38:4740–4754.

Nikolaev SI, Montoya-Burgos JI, Popadin K, Parand L, Margulies EH, National Institutes of Health Intramural Sequencing Center Comparative Sequencing P, Antonarakis SE. 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proc Natl Acad Sci U S A.* 104: 20443–20448.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463:457–463.

Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W. 2012. caper: comparative analyses of phylogenetics and evolution in R. R package version 0.5. [cited 2014 Mar 25]. Available from: http://CRAN.R-project.org/package=caper.

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–884.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet.* 40:1413–1415.

Parikesit AA, Stadler PF, Prohaska SJ. 2011. Evolution and quantitative comparison of genome-wide protein domain distributions. *Genes* 2: 912–924.

Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.* 6:e1001236.

Pinheiro J, Bates D, DebRoy S, Sarkar D 2013. nlme: linear and nonlinear mixed effects models. R package version 3-1.113. [cited 2014 Mar 25]. Available from: http://CRAN.R-project.org/package=nlme.

Popadin K, Polishchuk LV, Mamirova L, Knorre D, Gunbin K. 2007. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc Natl Acad Sci U S A.* 104:13390–13395.

Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453:1064–1071.

R Development Core Team. 2012. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Resch A, Xing Y, Modrek B, Gorlick M, Riley R, Lee C. 2004. Assessing the impact of alternative splicing on domain interactions in the human proteome. *J Proteome Res.* 3:76–83.

Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z, et al. 2006. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A.* 103:8390–8395.

Samuelsson B, Troein C. 2003. Superpolynomial growth in the number of attractors in kauffman networks. *Phys Rev Lett.* 90:098701.

Schad E, Tompa P, Hegyi H. 2011. The relationship between proteome size, structural disorder and organism complexity. *Genome Biol.* 12: R120.

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU. 2005. A gene expression map of *Arabidopsis thaliana* development. *Nat Genet.* 37:501–506.

Serra R, Villani M, Barbieri A, Kauffman SA, Colacci A. 2010. On the dynamics of random Boolean networks subject to noise: attractors, ergodic sets and cell types. *J Theor Biol.* 265:185–193.

Sorek R, Shamir R, Ast G. 2004. How prevalent is functional alternative splicing in the human genome? *Trends Genet.* 20:68–71.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A.* 101:6062–6067.

Taft R, Mattick J. 2003. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *Genome Biol.* 5:P1.

Takeda J-i, Suzuki Y, Sakate R, Sato Y, Seki M, Irie T, Takeuchi N, Ueda T, Nakao M, Sugano S, et al. 2008. Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs. *Nucleic Acids Res.* 36:6386–6395.

Tenaillon O, Silander OK, Uzan J-P, Chao L. 2007. Quantifying organismal complexity using a population genetic approach. *PLoS One* 2:e217.

Valentine JW, Collins AG, Meyer CP. 1994. Morphological complexity increase in metazoans. *Paleobiology* 20:131–142.

Venables WN, Ripley BD. 2002. Modern applied statistics with S. 4th ed. New York: Springer.

Vickaryous MK, Hall BK. 2006. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol Rev.* 81:425–455.

Vogel C, Chothia C. 2006. Protein family expansions and biological complexity. *PLoS Comput Biol.* 2:e48.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470–476.

Warnefors M, Eyre-Walker A. 2011. The accumulation of gene regulation through time. *Genome Biol Evol.* 3:667–673.

Whitney KD, Boussau B, Baack EJ, Garland T Jr. 2011. Drift and genome complexity revisited. *PLoS Genet.* 7:e1002092.

Whitney KD, Garland T Jr. 2010. Did Genetic Drift Drive Increases in Genome Complexity? *PLoS Genet.* 6:e1001080.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21: 1859–1875.

Xia K, Fu Z, Hou L, Han J-DJ. 2008. Impacts of protein–protein interaction domains on organism and network complexity. *Genome Res.* 18:1500–1508.

Xing Y, Lee C. 2007. Relating alternative splicing to proteome complexity and genome evolution. *Adv Exp Med Biol.* 623:36–49.

Xue B, Dunker AK, Uversky VN. 2012. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J Biomol Struct Dyn.* 30:137–149.

Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. *Mol Biol Evol.* 28: 2359–2369.