Computational analysis of the evolutionary dynamics of proteins on a genomic scale

Timothy Hughes

PhD thesis

Department of Informatics
University of Bergen
2008

Preface

I entered the field of evolutionary bioinformatics with the naive belief that biological knowledge was fairly advanced. In particular, that higher level phenotypes (anatomy and physiology) could be traced back to lower level phenotypes and ultimately to the genome. I have come to realise that this is not at all the case. However, we are fortunate to live in an era where vast amounts of genomic and proteomic data are becoming available and, thus, the aquisition of such knowledge lies ahead of us, but with any luck, not in the distant future.

I sit at my desk on the fifth floor of the Høyteknologisenter in Bergen. Directly outside my window, a gang of seagulls is out surfing on the strong winds of a North Sea gale. These beautiful birds glide for hours, controlling altitude and position with minute adjustments of their wings. They are also truely amphibious and can swim, dive and run. One is really tempted to evoke design, but design would never have produced such an adapted animal. In fact, it is designers that plunder the organic world for good ideas and I have yet to see evidence of information flow in the opposite direction. I type away at my computer analysing the evolution of genomic sequences *in silico,* and I say to myself that hopefully I am contributing to understanding both how such a creature can evolve and how it functions.

The Norwegian people, via the Norwegian Reseach Council, has invested considerable financial resources by paying for my PhD studies. The direct output of this investment, apart from musings on "seagull surfing", is a thesis consisting of four papers. If we only consider my salary, then each paper cost approximately 300,000 kroner to produce. I often wonder whether these papers return to society something approaching what they cost, but an answer is, and will probably remain, elusive. However, biology as a field promises to make great advances in knowledge which, if used wisely, will be of immense practical benefit to mankind. My hope is that breakthroughs occur thick and fast, and that my work may have contributed in some infinitesimally small way.

Acknowlegements

I have many people to thank for helping me during my doctoral studies.

First, I would like to acknowledge the substantial financial contribution from the Norwegian Research Council (Norges Forskningsråd) which funded my PhD studies through the functional genomics program (FUGE).

Gratitude is also due to my supervisor, David Liberles, for taking me on as a PhD student and supervising me during the last four years. Since the beginning David has left me more or less free to explore the field and to define what I wanted to work on. Initially, this was rather daunting, but in retrospect I am glad that I was supervised in this way. I hope David feels that I made good use of this free rein.

Collegues at the Computational Biology Unit, the Bergen Center for Computational Science and the University of Bergen have provided a good work environment, in which I was able to pick up the biological and computational knowledge that has made my work possible. Everyone contributes to this good work environment, but I would like to single out Inge Jonassen for his open and relaxed way of leading CBU and also for having been my co-supervisor.

Everyone in my family (France, England and Norway) has shown interest in and been supportive of my work. This is much appreciated. A special thanks is due to my brother Jo for tipping me off on the exiting challenges that lie within bioinformatics and more specifically evolutionary biology.

Finally, thanks to Marit for everything, but in particular: proof-reading my papers and thesis (which means that any remaining typographical errors are not mine), cooking delicious meals which are the highlight of my day, and helping me remember that what really matters are the *non-biotic* aspects of *life*.

Contents

1. Summary **1**

2. Background **5**

2.1. From genotype to phenotype 5

2.1.1. Life 5

2.1.2. DNA - the genotype 5

2.1.3. RNA and protein - the basic level of the phenotype 6

2.2. Evolution 8

2.2.1. Genetic variation 8

2.2.2. Fitness 8

2.2.3. Genetic drift 9

2.2.4. Selection 11

2.2.5. Adaptation 13

2.3. Detecting deviations from neutrality 15

2.3.1. Principle of the d_n/d_s measure 15

2.3.2. Nei-Gojobori 16

2.3.3. Likelihood-based method 17

2.3.4. The McDonald-Kreitman test 18

2.3.5. Extensions of basic methods 20

2.3.6. Adaptive evolution 21

2.4. Gene duplication 22

2.4.1. Smaller-scale duplication (SSD) 23

2.4.2. Whole genome duplication (WGD) 25

2.4.3. Fixation of the duplication event 26

2.5. Gene duplicate retention 27

2.5.1. Pseudogenisation 27

2.5.2. Neofunctionalisation 27

2.5.3. Subfunctionalisation 28

2.5.4. Subfunctionalisation followed by neofunctionalisation 30

2.5.5. Dosage balance . . . 31
2.5.6. Robustness and increased dosage . . . 32
2.5.7. Summary . . . 33
2.6. Gene families . . . 36
2.6.1. Concept . . . 36
2.6.2. Similarity-based methods for building families . . . 38
2.6.3. Structure-based methods for building families . . . 38
2.6.4. Phylogenetic methods for building families . . . 39
2.6.5. Power-law distribution of gene family size . . . 39
2.6.6. Phylogenetic trees . . . 40

3. Contributions **41**
3.1. Gene duplication and loss (paper I) . . . 42
3.1.1. Context . . . 42
3.1.2. Results . . . 42
3.1.3. Ideas for further work . . . 43
3.2. Models of duplicate retention (paper I) . . . 45
3.2.1. Context . . . 45
3.2.2. Results . . . 46
3.2.3. Ideas for further work . . . 47
3.3. The distribution of gene family size (paper II) . . . 48
3.3.1. Context . . . 48
3.3.2. Results . . . 48
3.3.3. Ideas for further work . . . 48
3.4. Hazard shift and WGD (paper III and IV) . . . 49
3.4.1. Context . . . 49
3.4.2. Results . . . 50
3.4.3. Ideas for further work . . . 50

Bibliography **53**

A. Paper I **65**

B. Paper II **67**

C. Paper III **69**

D. Paper IV **71**

1. Summary

Biology is primarily concerned with the study of all phenotypic aspects of living organisms and evolutionary biology is more specifically interested in elucidating how different phenotypes evolved. Proteins (and RNA molecules) are the most fundamental level of phenotype and are encoded by the genes in the organism's genome. Thus, at the most basic level, evolutionary biology seeks to understand how changes in the DNA sequence of genes affect protein functionality and how this modified functionality feeds back to shape the genome (and thus phenotype) of future generations.

Every nucleotide of the genome is constantly at risk of mutation and, if a mutation occurs in a gamete, it has a non-null probability of being passed on to the next generation. If the mutation has a negligeable effect on phenotype (neutral mutation) it may rise to fixation through genetic drift. If, however, the effect is non-negligeable and impacts on the oganism's fitness, it may either stand a higher chance of reaching fixation than a neutral mutation (positive selection) or it may stand a lower chance (negative or purifying selection). It is positive selection that drives the modified or new function which results in adaptation of the organism to its environment.

Because life has existed on earth for at least 3.5 billion years and because the state of the physical environment is relatively stable across time, the products of genes are usually well-adapted to a particular function. Most protein coding sequence is either evolving neutrally if the nucleotides encode amino acids that are functionally unimportant, or is under negative selective pressure if a change in the encoded amino acid would affect fitness. However, observation of the organic world both at the macro level (e.g. anatomy and physiology of organisms) and at the micro level (e.g. proteins) reveals what appear to be many cases of recent adaptation involving novel function. Of course, changes in an organism's physical and biotic environment may occur and would have the potential to drive adaptive changes in a gene's function. However, most genes, because they encode functions that are essential regardless of the organism's environment, are not free to evolve in this way.

The key process enabling a gene to escape the eye of selection is gene duplication. Through duplication of a gene, redundancy is introduced to the genome as it then contains two copies of the same gene, both of which encode the same functionality. Such

a duplication will generally be neutral and can reach fixation by drift. There are many fates for the gene duplicate pair, the most common of which is pseudogenisation (or gene death/loss) which involves one of the genes in the pair losing its protein encoding properties (fixation of a null mutation). The reason for this is that, in most cases, a null mutation to one of the genes in the pair does not have any fitness effect on the mutant individual as the other gene in the pair continues to fulfill the required function. However, some gene duplicates are retained. The process through which retention occurs is an intensively studied subject as differences in the gene content of genomes is one of the main drivers of phenotypic diversity among species. Several models of gene duplicate evolution have been formulated, the first and probably most intuitive model being the "neofunctionalisation" model [Ohno 1970]. The key idea of "neofunctionalisation" is that there is a small chance that one of the genes in the duplicate pair is subject to a mutation confering a new fitness enhancing function on the protein, thus ensuring the retention of both genes in the genome: one gene having the ancestral function and the other the new function (neofunctionalisation). This is one of the most obvious ways in which adaptive evolution can occur at the protein coding level.

Thus, gene duplication and the subsequent retention or loss are key processes shaping the evolution of genomes. They drive the actual number of genes in the genome and these genes functions. Moreover, they potentially produce neofunctionalisation.

In this thesis, using genomic data from mammalian species, I begin by estimating the rate at which genes duplicate, and the rate at which the sequence of the duplicates diverges and potentially pseudogenises (Paper I). These estimates are of interest in their own right as they represent a quantitative characterisation of an important evolutionary process, but they can also be used to investigate the predominant mode of gene duplicate evolution (Paper I). Further, these estimates can be used to investigate the evolution of the gene content of a genome and, more specifically, the distribution of gene family size (Paper II). Finally, although these estimates are for gene duplicates that are the result of small-scale duplication events (tandem and segmental duplication), the estimates can be applied to investigating some of the particularities of whole genome duplication (Paper III and IV).

The background knowledge required to understand the papers is presented in chapter 2. Hopefully, this background knowledge is sufficiently complete for the uninitiated reader to understand the essence of the findings of the papers. Readers familiar with the subject will probably find that they can skip large sections of this chapter. Each of the four papers is then introduced in chapter 3. Each introduction consists of more detailed background information that is relevant for the specific paper, a motivation of the work, a short summary of the results and some ideas for further work. Finally, the core of this

thesis, the actual papers together with their bibliographies and supplementary materials, are located in the appendix. This layout may seem somewhat unconventional, but it is made necessary by the guidelines for doctoral degrees at the University of Bergen which require the PhD candidate to produce papers which are later incorporated into the thesis.

2. Background

2.1. From genotype to phenotype

2.1.1. Life

Biology is the study of life. Life is the condition that distinguishes organic from inorganic objects. Although "life" has no formal definition, three of the most fundamental features of an organism are: first, a very high degree of chemical complexity compared to inorganic objects; second, the ability to extract, transform, store and use energy from their environment which enables the organism to generate and maintain its chemical complexity; and, third and foremost, the capacity for self-replication and self-assembly, what Schrödinger calls "architect's plan and builder's craft in one" [Schrödinger 1944]. This plan, the genetic information, is stored in the form of deoxyribonucleic acid (DNA) in the organism's cell(s) and, more specifically, in the cell's nucleus in the case of eukaryotic species.

2.1.2. DNA - the genotype

DNA is a long polymer of simple units called nucleotides, with a backbone made of sugars and phosphate moeities which are covalently linked by asymmetrical 5'-3' phosphodiester bonds. Attached to each sugar is one of four types of molecules called bases: adenine (A), guanine (G), cytosine (C) and thymine (T). It is the sequence of these four bases along the backbone (conventionally represented in the 5'-3' direction) that encodes the genetic information. Two of these polymeric strands are twisted about each other to form the DNA double helix in which each monomeric subunit in one strand forms hydrogen bonds specifically with a complementary subunit in the other strand (A with T, and G with C). The capacity of living cells to preserve their genetic material and to replicate it with high fidelity for the next generation derives directly from the structural complementarity between the two halves of the DNA molecule [Watson and Crick 1953].

The DNA double helix has several higher levels of organisation which fundamentally consist of further levels of coiling and super-coiling [Nelson and Cox 2000]. The highest level of organisation, which is visible in the light microscope during cell division, is the chromosome. The number of chromosomes in the organism's cell(s) depends on the species and on whether the cells are somatic or gametic. *Homo sapiens*, for example, which is a sexually reproducing diploid eukaryotic species, has two homologous sets of 23 chromosomes (diploid) in the nucleus of somatic cells and one set of 23 chromosomes (haploid) in gametes (spermatozoan and ovum). All descriptions of biological processes in this chapter will be for sexually reproducing diploid eukaryotic species (unless noted otherwise).

2.1.3. RNA and protein - the basic level of the phenotype

Through the process of transcription, information is transcribed from sections of one of the DNA strands known as genes (see Figure 2.1) to RNA (ribonucleic acid). The primary differences between DNA and RNA are that RNA contains the sugar ribose (rather than deoxyribose) and that the base thymine is replaced by uracil (U). In the case of eukaryotes, most protein-coding genes are encoded in sections (exons) which are interrupted by non-coding elements (introns). The introns are spliced out of the preliminary transcript and a mature messenger RNA is produced (mRNA). RNA has many roles: it has important functional properties e.g. enzymatic activity, but its main role is still considered to be as a transmitter (messenger) of genetic information. DNA is much more stable than RNA because of structural aspects of the deoxyribose versus the ribose sugar. This property of DNA makes it a more robust storage device for genetic information than RNA, and may be the cause of their divergent functions.

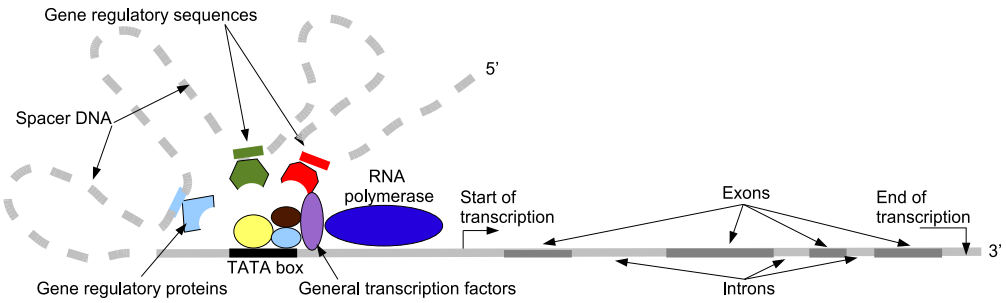

Figure 2.1.: Schematic representation of gene structure and transcription

Messenger RNA is converted to a chain of amino acids (or polypeptide) through the process of translation which converts consecutive triplets of nucleotides (codons) in the mRNA into a chain of amino acids according to the rules of the genetic code. Most

organisms use the so-called universal genetic code (see Figure 2.2) which is in fact not universal as it is not used by all species and the mitochondrial genome uses a slightly different code (mitochondria are small cellular organelles that contain DNA). Note that there are 4^3 possible codons, but only 20 different amino acids, thus there is redundancy in the genetic code with some amino acids encoded by multiple codons. For example, arginine (Arg) is encoded by six different codons. Thus, a change of base in the DNA (a mutation), if it occurs in protein-coding sequence, can be classified according to whether it changes the encoded amino acid: silent (or synonymous) mutations do not change the encoded amino acid, while replacement (or non-synonymous) mutations do. A section of protein-coding DNA (open reading frame) will always start with a methionine codon and end with one of the three stop codons.

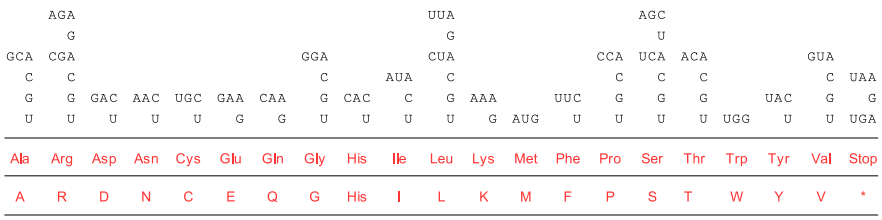

Figure 2.2.: The universal genetic code

Following translation, the polypeptide chain folds, first into secondary structures (alpha helices and beta sheets) and then into a three dimensional structure consisting of one or several domains (tertiary structure). Further, some proteins consist of multiple polypeptide chains (quaternary structure). It is the amino acid composition of the chain and the resulting structure that confer on proteins the functionality to extract, transform and use the energy from the organism's environment to generate and maintain the organism's chemical complexity, and ultimately enable the genetic information to be passed on to the next generation. The entire flow of genetic information from DNA, through RNA, to protein is known as the central dogma (see Figure 2.3).

Examples of the functionality of proteins are: enzymatic (catalysis of covalent bond breakage or formation), structural (mechanical support to cells and tissues), transport (spatial movement of small molecules and ions), motor (movement in cells and tissues), storage (of small molecules or ions), signalling (information transmission either within or between cells), receptor (detecting signals that are either internal or external to the organism), and regulation of gene expression. Proteins are thus the most fundamental level of phenotype, i.e. observable characteristics of an individual, as opposed to the underlying genetic features. Through their function, proteins are the primary determinants of higher levels of anatomy and physiology e.g. cells, organs, and organ systems.

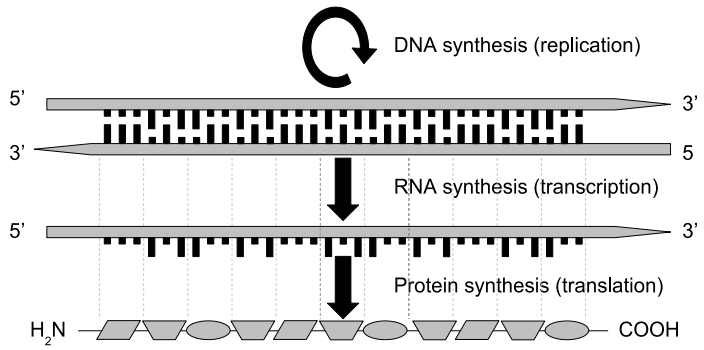

Figure 2.3.: The central dogma (adapted from [Alberts et al. 1997])

2.2. Evolution

2.2.1. Genetic variation

In summary, the DNA molecule has properties that make it a good storage and transmission device for genetic information, and the information it encodes are the blue prints for functionality that primarily aids in the preservation and faithful transmission of this information to the next generation. Nevertheless, the fidelity of DNA's information storage and transmission capacity is not perfect. Changes in the sequence may occur (mutations). At the level of a single nucleotide, there are three types of mutation: change in a base, deletion of a base, insertion of an additional base. Changes involving whole segments of a chromosome are also possible, including inversion, translocation, transposition and duplication of whole segments of DNA.

If such changes occur in the germ line, i.e. in cells that have the potential to go on to form a new organism (gametes), then all cells in a new organism generated from a mutated germ cell will carry the mutation. Thus, within a population of individuals there will always be a degree of genetic variation due to past mutations in the germ line that filtered down to the present population, plus new mutations that arose in the present population. An allele is a viable DNA coding at a given position (locus) on a chromosome, but the term may also refer to two allelic genes at a given locus. It is allelic variation initially caused by mutation that provides the raw material for evolution.

Many consider genetic variation caused by mutation as a defining characteristic of life to be added to the three presented in sub-section 2.1.1.

2.2.2. Fitness

A key property of a mutation is whether or not it affects the fitness of its bearer. Fitness is defined as an individual's propensity to contribute offspring to the next gener-

ation [Sober 1993]. If all individuals were phenotypically identical, then the expected number of offspring would be the same for all individuals. But, there is genetic variation and different genotypes in interaction with the environment produce different phenotypes, and these different phenotypes have different fitnesses. Thus, genetic mutations can be classified according to whether they have an advantageous, detrimental or neutral effect on fitness.

2.2.3. Genetic drift

When a mutation produces a new allele, there is initially only one occurence of the allele in the population. Subsequently, the allele may either increase in frequency or disappear (see Figure 2.4). These dynamics are affected by whether the allele in question has a fitness effect. This section summarises the results for a neutral allele (pure genetic drift) and, in the next section, the results for an advantageous allele are presented (selection).

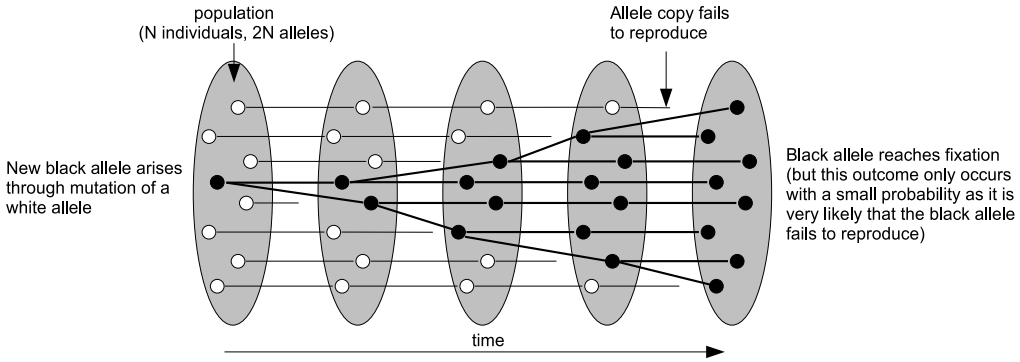

Figure 2.4.: Genetic drift (adapted from [Page and Holmes 1996])

Due to the stochastic nature of reproduction, an individual organism will not contribute all its DNA to the next generation. First, for all organisms there is the possibility that the individual does not reproduce in which case no DNA is contributed to the next generation. Second, for sexual reproducing species, the parent contributes only half of its DNA to any given descendent. And, third, recombination (exchange of sections of DNA between homologous chromosomes) occurs during the production of gametes through meiosis (see Figure 2.5). Thus, chance is a fundamental force driving the frequency of the four different nucleotides at a specific position of the genome of a population of individuals (allele frequencies).

If we consider a population of N diploid individuals, there will be $2N$ allelic copies of each gene. But, due to the random sampling of gametes that contribute to the next generation, some alleles will contribute no copies of themselves to the next generation

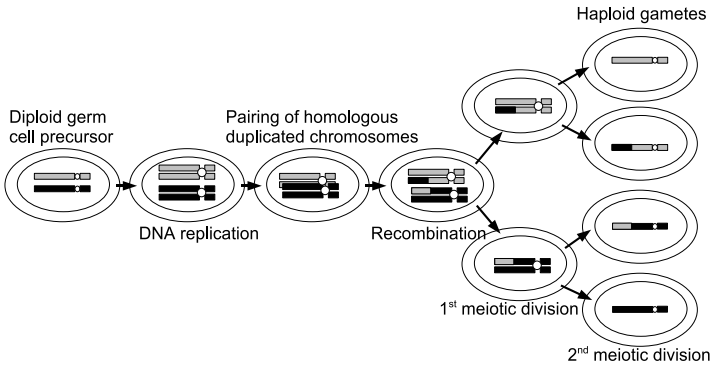

Figure 2.5.: Meiosis (adapted from [Alberts et al. 1997])

while others will contribute several. This causes the frequencies of the different alleles to change from generation to generation and given enough time all $2N$ alleles in the population will be descendent from one allele as all others will at some point fail to contribute copies to the next generation (see Figure 2.4). When an allele reaches a frequency greater than 99% in the population, it is said to be a fixed. A fixed mutation is referred to as a substitution.

The pure genetic drift model describes the dynamics of neutral alleles. In this model, one assumes an idealised population with constant population size, random mating, an equal number of each sex contributing to the gene pool, and non-overlapping generations. In real populations, one or more of the conditions is likely to be violated and the concept of effective population size (N_e) is used as a way of correcting for such violations. For a given real population, N_e is defined as the size of an idealised population having the same characteristics (with regard to genetic drift) as the real population (N) and it is usually the case that $N > N_e$.

The formulation of the model and its solution are mathematically advanced [Kimura 1983], but the results are simple and intuitive. First, the probability of fixation of an allele by random genetic drift is $1/2N$ which is its frequency in the population after it has arisen by mutation. Second, the expected time to fixation is $4N_e$ generations. Third, if we define K_0 to be the rate of substitution of neutral alleles (per generation), u to be the total mutation rate per generation, and f_0 to be the fraction of all mutations that are neutral, then $K_0 = uf_0$ [Hughes 1996].

Such a model is applicable to large sections of the genome. First, because the DNA molecule does not consist of a continuous string of genes, instead genes are separated by intergenic regions some parts of which may be functional, but current knowledge suggests that most of these regions are what is called "junk DNA" i.e. without phenotype. Second, because, even within a gene, not all nucleotides affect the 3D structure of

the polypeptide chain. Mutations within introns will not affect phenotype as these are excised from the mRNA before translation. Even if the mutation occurs within an exon, it may not affect phenotype: either because the mutation is synonymous, or, in the case of a non-synonymous mutation, because the affected amino acid plays an insignicant role in the 3D structure or the structure's function. However, it is important to note that the neutral model applies only as a first approximation as even synonymous mutations have been shown to have potential fitness effects [Chamary et al. 2006].

2.2.4. Selection

However, many mutations do have a fitness effect. The nature and magnitude of this effect is determined by the interaction of the resulting phenotype with the environment and feeds back to the genotype indirectly by affecting whether or not the genotype of the individual is represented in the next generation (see Figure 2.6), this is referred to as natural selection to distinguish it from the artificial selection applied by a breeder [Darwin 1859]. In this case, the evolution of allele frequencies is not influenced purely by chance and the genetic drift model needs to be enhanced to include selection. Broadly speaking, a selectively advantageous mutation has a higher probability of reaching fixation than a neutral mutation and, given that it does, it will do so more rapidly than a neutral mutation.

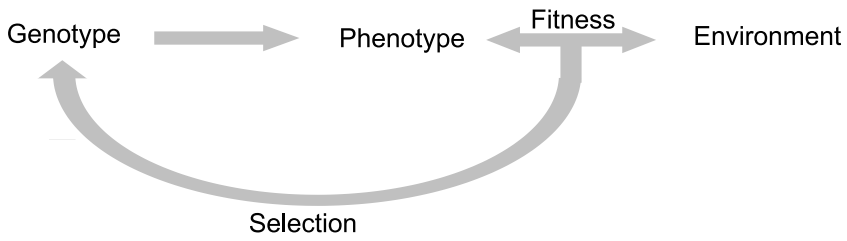

Figure 2.6.: Fitness as the driver of selection

If we define the selective advantage of a heterozygote for the mutant to be s and assume additivity, so that the selective advantage of a homozygote for the mutant is $2s$, then, although the mathematical formulation of the model is advanced, it is tractable and the results are intuitive [Hughes 1996]. A mutant with selective advantage s and initial frequency q will have the following probability of fixation (Pf):

$$Pf = \frac{1 - \exp(-4N_e sq)}{1 - \exp(-4N_e s)} \tag{2.1}$$

The result for a neutral mutation is obtained by evaluating the limit as $s \to 0$:

$$\lim_{s \to 0} Pf(s) = q$$

In the case of a new mutation occurring on one chromosome in a diploid population $q = 1/2N$. Thus, for a neutral mutation in such a population $Pf = 1/2N$, as already mentioned.

Assuming the effective size of the population is equal to the actual size, N may be substituted for N_e and the probability of fixation of an individual mutant gene is obtained from equation 2.1 by setting $q = 1/2N$. If s is small:

$$Pf \simeq \frac{2s}{1 - \exp(-4Ns)}$$

For a positive s and a very large N, $Pf \simeq 2s$. If $N \neq N_e$, this value should be modified by a factor of N_e/N [Kimura 1964]. So that

$$Pf = 2s(N_e/N)$$

This probability will be quite low if s is low. In fact, an advantageous mutant will behave essentially like a neutral mutant if $s < 1/2N_e$ [Kimura 1983].

The rate of substitution of selectively advantageous mutants K_a is given by $K_a = 4N_e s f_a u$ where s is the average selective advantage of these mutants and f_a is the frequency of mutants that are advantageous. The main difference with the neutral case is that both effective population and average selective advantage play a role.

It is important to note that we have only described one form of selection here. The kind of selection that the Kimura model describes is additive advantage which means that fitness of the heterozygote has intermediary fitness between the two homozygotes, this results in directional selection whereby the selectively advantageous allele rises to fixation. However, it is, for example, possible that the heterozygote has superior fitness to the homozygotes (heterozygote advantage), in this case, allelic diversity will be maintained. A classic example of this is a locus in the human genome coding for a protein which affects the shape of red blood cells [Page and Holmes 1996]. Individuals that are homozygous for the wild type have normal red blood cells and are susceptible to malaria; those that are homozygous for the mutant allele have grossly mishaped red blood cells which detrimentally affects their oxygen carrying capacity; heterozygotes, however, have only slightly irregular blood cells which does not significantly affect oxygen carrying capacity at the same time as it confers resistence to malaria. In this

way, natural selection preserves both alleles in the population. Another way in which allelic diversity can be maintained is frequency-dependent selection in which the frequency of an allele is inversely correlated with the selective advantage is confers. In such a case, an allele will fail to reach fixation because as the frequency rises the selective advantage disappears.

2.2.5. Adaptation

Selection produces a pressure on the genome which ensures that detrimental mutations have a low probability of rising to high frequency while beneficial mutations are retained and fixed in the population. This results in a genotype that encodes a phenotype which is adapted to its environment. However, detecting and locating adaptive features in molecular data is not trivial.

The first challenge in the study of adaptation is to find features that share a common ancestry (homologous features). In the case of sequence data, this involves first finding sequences that are homologous and then locating within the sequences the residues that are homologous. When sequences are separated by a short divergence time, there are few mutations between the sequences and both tasks can be relatively simple. However, when the divergence time is greater, insertions, deletions and substitutions accumulate, resulting in sequences of different length and composition, and these tasks become more complicated. Orthologous genes, which are genes that occur in two different species and have diverged from the sequence in the common ancestor due to the speciation event that separates them, are an example of the kind of sequence for which the assignment of homology might be more problematic, in particular if divergence times are great. In this case, given a query sequence, homologous sequences are usually identified by searching for similar sequences using tools such as BLAST [Altschul et al. 1997] and by then assuming that statistically significant similarity implies homology. Following identification of homologous sequences, one typically employs a multiple sequence alignment algorithm to locate homologous residues [Thompson et al. 1994, Notredame et al. 2000, Edgar 2004]. The inputs to such an algorithm are a set of homologous sequences and the output is a matrix in which each row corresponds to a sequence and homologous residues in the sequences are placed in the same column (see Figure 2.7).

Once the homologous sequences have been aligned, the task of determining whether some of the features are adaptive can begin. In the multiple sequence alignment, a large proportion of sites are either identical or occupied by amino-acids with similar physio-chemical properties indicating that the site is under negative selective pressure.

Figure 2.7.: Section of a multiple sequence alignment

However, there are also large numbers of sites that are occupied by physio-chemically different amino-acids, and one cannot immediately tell whether these differences are due to the fact that the changes have a neutral effect on fitness and have been fixed by drift, or whether natural selection may have played a role in their fixation. During the first half of the 20th century, sequence data was scarce and differences between sequences were thought to be rare. It was widely believed that the differences that did exist were the result of adaptation. However, in 1966, two studies, one on *Homo sapiens* [Harris 1966] and one on *Drosophila pseudoobscura* [Lewontin and Hubby 1966], revealed high levels of genetic variation. Kimura showed that such levels of genetic diversity are only consistent with a significant fraction of mutations being neutral and genetic drift playing a major role in their fixation [Kimura 1968]. The large amounts of genomic data, which have become available since, have further confirmed that many substitutions are neutral or nearly-neutral [Ohta 2002]. It is only recently that positive selection in sequence evolution has been detected on a significant scale.

This raises the question of why, at least until very recently, so little positive selection had been detected. There are several possible reasons. First, it could be that what is considered to be clear and plentiful examples of adaptation at the morphological and physiological level is only the result of a small amount of adaptation at the molecular level. Second, we could be looking in the wrong sections of DNA sequences: adaptation might be occurring mostly in gene regulation or alternative splicing rather than in protein coding regions. Third, it could be that the methods used for testing for positive selection were simply not powerful enough to reject the null hypothesis of neutral evolution. We review some of these tests and the results of their application in the next section.

The dichotomy between the amount of adaptation observed at the macro level and at the molecular level make the “hunt” for adaptive evolution an exciting pursuit. Are the adaptive physiology and morphology only driven by a small number of molecular changes? If a large number of substitutions are adaptive, in what sections of the DNA are they located? Detecting adaptation is also extremely important from a practical

point of view, as identification of adaptive sequence can provide important information on sequence function and change in function.

2.3. Detecting deviations from neutrality

There are two main sources of genetic variation: within populations (of a given species) and between species. Both of these types of variation can be used to detect deviations from neutrality. This section presents a short overview of some of the most widely used methods at the DNA level.

2.3.1. Principle of the d_n/d_s measure

A very intuitive measure requiring only two sequences from different species is the d_n/d_s ratio. The measure builds on the assumption that synonymous mutations are neutral as they do not lead to a change in the encoded amino acid, while non-synonymous mutations change the encoded amino acid and may or may not affect fitness depending on the nature of the replacement and the role of the affected amino acid in the folded protein. We have previously seen that beneficial mutations stand a higher chance of rising to fixation and will do so more rapidly than neutral mutations, whereas the opposite is true for deleterious mutations. Thus, if we observe an equal number of non-synonymous substitutions per non-synonymous site (d_n) and synonymous substitutions per synonymous site (d_s) when comparing two aligned protein-coding sections of DNA, then non-synonymous substitutions are accumulating at the same rate as synonymous substitutions and the sequence is likely to be evolving neutrally (as long as it is reasonable to assume that the underlying mutation rate is the same for synonymous and non-synonymous mutations). On the other hand, excess of non-synonymous substitutions per non-synonymous site indicates positive selection and deficit indicates negative selection. The ratio d_n/d_s is also referred to as K_a/K_s (where the "a" stands for asynonymous):

$$d_n/d_s > 1 \text{ positive selection}$$
$$d_n/d_s \simeq 1 \text{ neutral evolution}$$
$$d_n/d_s < 1 \text{ negative selection}$$

It has been shown that there is codon bias in protein-coding genes and that this bias may be due to selection for translational efficiency [Eyre-Walker 1996], however, this is not thought to be strong enough to invalidate the use of tests that rely on the assumption that synonymous mutations are neutral. Further, codon bias and other processes that

render synonymous mutations non-neutral can be incorporated into the model for d_n/d_s estimation [Anisimova and Liberles 2007].

2.3.2. Nei-Gojobori

There are a number of methods for computing the d_n/d_s ratio. They can be divided into counting methods and maximum likelihood methods. Here, the Nei-Gojobori method is explained [Nei and Gojobori 1986] as it is a very intuitive method (the next section is devoted to the more advanced maximum likelihood method). The key assumption of the Nei-Gojobori method is that all nucleotide substitutions are equally likely. In order to compute the d_n/d_s ratio, we need to estimate the number of synonymous and non-synonymous sites and the number of synonymous and non-synonymous substitutions between two aligned protein coding sequences.

The amino acid alignment is first reverse-translated to the encoding nucleotide sequence. We denote f_i, the proportion of potential synonymous mutations at the i^{th} nucleotide position of a codon, and we define this as the ratio of the number of synonymous changes to the sum of synonymous and non-synonymous mutations excluding stop mutations. Then, the number of potential synonymous sites for a codon is given by $f_1 + f_2 + f_3$ and the number of potential non-synonymous sites is $3 - (f_1 + f_2 + f_3)$. For example, UUU has only one synonymous substitution (to UUC), thus the number of synonymous sites for the codon is 1/3 and the number of non-synonymous sites is $3 - 1/3$. To obtain the total number of synonymous and non-synonymous sites for the whole sequence, we sum over the codons. Note that we are comparing two sequences, so we compute the total number of sites of each type separately for both sequences and then take the average.

In order to compute the number of substitutions, we compare the two sequences codon by codon and count the number of nucleotide differences for each pair of codons. If there is one nucleotide difference, then we know whether it is synonymous or not. If there are two differences, there are two possible pathways that explain the differences. For example, between UUU and GUA:

UUU (Phe) → GUU (Val) → GUA (Val) i.e. 1 syn. and 1 non-syn. substitution
UUU (Phe) → UUA (Leu) → GUA (Val) i.e. 2 non-syn. substitutions

Assuming both pathways occur with equal probability, the number of synonymous differences is 0.5 and the number of replacement differences is 1.5. In some comparisons of codons, there are pathways with termination codons, these pathways are eliminated from the computation. This calculation is performed for all codons and we sum over the codons.

We then compute the number of synonymous substitutions per synonymous site (d_s) and similarly for non-synonymous substitutions (d_n). In some cases, there may be more substitutions between the sequences than observed when comparing them because the same site may have undergone multiple substitutions. These ratios are therefore corrected for these multiple hits. This method for computing the ratio is intuitive and useful for explaining the concept, but it builds on the assumption that all nucleotide substitutions are equally likely and this is rarely the case.

For example, it is usually the case that the transition rate is much higher than the transversion rate. In this case, the number of potential sites that can produce synonymous substitutions is expected to be higher than the number estimated by the Nei-Gojobori method, because transitional changes at third codon positions are mainly synonymous. Thus, the Nei-Gojobori method will overestimate d_s and underestimate d_n, leading to a downward biased ratio.

A number of improvements to this basic counting method have been implemented [Li 1993, Pamilo and Bianchi 1993, Ina 1995], but the most conceptually simple way of incorporating more realistic models of evolution is by using a maximum likelihood estimation of a Markov chain model of codon substitution.

2.3.3. Likelihood-based method

Markov chain models of codon substitution were proposed by Goldman and Yang [Goldman and Yang 1994]. In these models, the codon triplet is considered the unit of evolution and a Markov chain is used to describe substitutions from one codon to another. The state space of the chain are the sense codons in the genetic code. Stop codons are not allowed inside a functional protein and are not considered in the chain. The Markov model is constructed by specifying the substitution rate matrix, $Q = \{q_{ij}\}$ where q_{ij} is the instantaneous rate from codons i to j ($i \neq j$). The model in common use is a simplified version of the model of Goldman and Yang [Yang 2006]:

$$q_{ij} = \begin{cases} 0, & \text{if and j differ at two or three codon positions} \\ \pi_j, & \text{if i and j differ by a synonymous transversion} \\ \kappa\pi_j, & \text{if i and j differ by a synonymous transition} \\ \omega\pi_j, & \text{if i and j differ by a nonsynonymous transversion} \\ \omega\kappa\pi_j, & \text{if i and j differ by a nonsynonymous transition} \end{cases}$$

where κ is the transition/transversion rate ratio, ω is the non-synonymous/synonymous rate ratio, and π_j is the equilibrium frequency of the codon j. Mutations are assumed to

occur independently at the three codon positions so that simultaneous changes at two or three positions are considered negligeable and are given a rate of zero.

From this model, it is possible to calculate a transition probability matrix $P(t) = \{p_{ij}(t)\}$ where $p_{ij}(t)$ is the probability that a given codon i will become j time t later. One can then use a maximum likelihood method to fit the Markov model to data of two sequences to estimate parameters in the model. The log likelihood function is:

$$l(t) = \sum_i \sum_j n_{ij} \ln\{\pi_i p_{ij}(t)\}$$

where n_{ij} is the number of sites occupied by codons i and j in the two sequences. The codon frequencies are usually estimated by using the observed frequencies in the data, while parameters t, κ and ω are estimated by numerical maximization of the log likelihood. Then d_s and d_n are calculated from the estimates of t, κ, ω, and π_j according to their definition (see [Yang 2006] for full details). By estimating two models, one where ω is free to vary and one were ω is fixed to 1, one can perform a likelihood ratio test to determine whether the null hypothesis of neutral evolution can be rejected in favour of positive or negative selection.

The main advantages of the likelihood method are its conceptual simplicity and the ease with which more realistic models of codon substitution can be accomodated.

2.3.4. The McDonald-Kreitman test

Several tests have been developed to test for deviation from neutrality in population genetic data [Tajima 1989, Fay and Wu 2000], the most commonly used is the McDonald-Kreitman test [McDonald and Kreitman 1991]. In the MK test, variable sites in protein coding genes from closely related species are classified into a 2x2 contingency table, whether a site has a polymorphism or a fixed difference, and whether the difference is synonymous or non-synonymous. For example, suppose we sample five sequences from species 1 and four from species 2. A site with data AAAAA in species 1 and GGGG in species 2 is called a fixed difference. A site with AGAGA in species 1 and AAAA in species 2 is polymorphic. The neutral null hypothesis is equivalent to independence between the row and column in the contingency table.

To see why this is a valid test of neutrality, begin by assuming that all synonymous mutations are neutral, that all non-synonymous mutations are either strongly deleterious, neutral or strongly advantageous, and that advantageous mutations contribute little to polymorphism (but may contribute to substitutions). Under this model, the number

of synonymous (P_s) and non-synonymous (P_n) polymorphisms segregating in a sample of sequences from a population are (for an autosomal locus):

$$P_s = 4N_e u L_s k \; and \; P_n = 4N_e u f L_n k$$

where N_e: the effective population size, u: the nucleotide mutation rate, f: the proportion of amino-acid mutations which are neutral, L_s and L_n: the numbers of synonymous and non-synonymous sites, respectively, k: a constant reflecting the probability of observing a neutral variant [Eyre-Walker 2006].

The numbers of synonymous (D_s) and non-synonymous (D_n) substitutions are:

$$D_s = 2utL_s \; and \; D_n = 2utfL_n + a \tag{2.2}$$

where t: the time of divergence between the two species being considered, a: the number of adaptive substitutions.

It is evident that, if $a = 0$, then D_n/D_s is expected to equal P_n/P_s and this forms the basis of the MK test. It is also not difficult to show from these equations that the number of adaptive substitutions in a gene can be estimated by:

$$a = D_n - D_s.\frac{P_n}{P_s}$$

So, dividing this expression by D_n gives an estimate of the proportion of amino-acid substitutions driven by positive selection ($\alpha = a/D_n$) [Smith and Eyre-Walker 2002]:

$$\alpha = 1 - \frac{D_s P_n}{D_n P_s}$$

The above describes how to perform the MK test and estimate α for one gene. There are several methods for estimating the average value of α for data from multiple genes. The most basic method involves simply summing D_s, D_n, P_s, and P_n across genes and, despite its simplicity, this method usually agrees with more advanced approaches [Welch 2006].

There are several assumptions behind the MK method of estimating the proportion of amino-acid substitutions driven by positive selection, but the test is generally robust to violations of most assumptions. The exception is if fitness reducing mutations are only slightly deleterious. In this case, if population size has been stable, the estimate of α is an underestimate, because slightly deleterious mutations contribute relatively more to polymorphism than they do to divergence when compared with neutral mutations. On the other hand, if the population sizes have expanded, slightly deleterious mutations can

lead to an overestimate of α, because mutations that might have been fixed in the past, when the population was small, no longer segregate as polymorphisms [Eyre-Walker 2006].

2.3.5. Extensions of basic methods

Early studies using the d_n/d_s criterion took the approach of pairwise sequence comparison, averaging the gene sequence and over the whole time period separating the sequences. However, positive selection, if it occurs, may affect only a few sites which are not necessarily adjacent in the primary sequence (e.g. an active site) and probably take place over only a limited period of time [Golding and Dean 1998], while most sites are expected to be under negative selection [Siltberg and Liberles 2002]. Thus, pairwise comparisons, which average over time and sequence, rarely detect positive selection.

More formally, if we assume that synonymous mutations are neutral and non-synonymous mutations are either deleterious, neutral or advantageous then $d_s = 2ut$ and $d_n = 2utf/(1-\alpha)$ (derivable from equation 2.2 and the definition of α). Thus, d_n can only exceed d_s if $(1-\alpha) < f$. Values of f are typically less than 0.3, as judged by average d_n/d_s values [Roth and Liberles 2006], so the proportion of substitutions that are adaptive needs to be greater than 0.7 for adaptive evolution to be detectable if averaging across sites [Eyre-Walker 2006]. Such a proportion of adaptive substitution in the protein coding sequence of a gene is highly unlikely.

Fortunately, both counting and maximum likelihood methods can be extended to increase their power. The most obvious and commonly used method for improving the power of the counting methods is to use a sliding window approach which, instead of calculating the d_n/d_s of the full length of a protein-coding gene, computes the ratio on a window which slides along the primary sequence of the gene and is designed to detect a selective sweep [Endo et al. 1996, Fares et al. 2002]. A more advanced "windowing" approach is a 3D windowning method based on the tertiary structure of the protein [Berglund et al. 2005]. The rationale behind this method is that selection often affects specific binding pockets or interacting residues which may be distantly located in primary sequence, but are close in the tertiary structure. By applying these approaches the signal of positive selection is enhanced, resulting in the enhancement of the evidence for positive selection for certain sites and the discovery of new sites.

The likelihood approach under models of codon substitution can be extended to analyse multiple sequences on a phylogenetic tree and by allowing the ω parameter to vary across branches (branch models), it is possible to test for positive selection along particular branches of the tree. Yang has implemented several models that allow for different

levels of heterogeneity in the ω ratio among lineages [Yang 1998]. Moreover, it is possible to let the ω ratio vary among sites (sites models). Positive selection is then indicated by presence of sites with $\omega > 1$ rather than the ω ratio averaged over all sites being > 1. Finally, branch-site models have been developed [Yang and Nielsen 2002, Zhang et al. 2005] to enable the detection of local episodic natural selection.

2.3.6. Adaptive evolution

Population genetic data from *Drosophila* suggests that a very high proportion of amino acid substitutions, averaging approximately 40 percent across several studies, are driven by positively selected nucleotide substitutions [Eyre-Walker 2006]. A high percentage of nucleotide mutations in non-coding DNA have also been shown to have been fixed by selection. An extreme case is the untranslated region of mature mRNAs (UTRs) where 60% of fixed mutations are estimated to be adaptive [Andolfatto 2005]. Estimates in microorganisms such as *Escherichia coli* and some viruses are even higher [Eyre-Walker 2006]. However, within chordates, and more specifically *Homo sapiens* which is the main chordate species in which this kind of study has been carried out, this proportion has been estimated to be a lot lower [Bustamante et al. 2005].

Although not directly comparable to the population genetic data, the comparative genomic data also fails to detect high percentages of adaptive substitutions. For example, a systematic scan for adaptive evolution in chordates and embryophytes (higher plants), in which 15,462 chordate gene trees were generated (based on 348,142 genes), only returned 505 chordate branches with $d_n/d_s \gg 1$ using the full length of coding sequences [Roth et al. 2005]. This number would undoubtedly have been higher if a maximum likelihood branch (or branch-site) model or a 3D windowing method had been used, as this was shown to make a significant difference in the number of branches identified as being under positive selection in the case of plant sequences [Roth and Liberles 2006]. However, it is difficult to use such an approach on a large scale due to the limited amount of structural data available or due to the high computational requirements of maximum likelihood methods.

Despite the fact that adaptive evolution appears to have occurred on much more limited scale in chordate genomes than in for example *Drosophila*, there are many examples of adaptive evolution (see, [Yang 2006] on pages 287-289 for an extensive but not exhaustive list covering multiple species, or [Vallender and Lahn 2004] for a comprehensive review of genes affected by positive selection in humans). Independently of lineage, the genes that have been detected as affected by positive selection tend to fall into one of three broad categories: proteins involved in defence systems or immu-

nity (or avoiding defence systems), proteins involved in reproduction, and gene duplicates [Yang 2006]. The prominence of the first two categories is often explained by the Red Queen Principle which gets its name from the race in Lewis Carroll's "Through the looking glass" where the Red Queen says: "It takes all the running you can do, to keep in the same place". This is a metaphor for a species which continuously adapts in order to maintain its fitness relative to the species it is co-evolving with. It is relatively easy to see how this "arms race" evolution is applicable to proteins involved in defence and immunity systems and proteins involved in evading/penetrating these systems (host-pathogen interactions). It also applies to proteins involved in reproduction, in particular fertilization-related proteins, because sperm and ovum have similarly opposite functions: the spermatozon's functions are geared towards rapidly identifying and fertilising the ovum, while one of the functions that is key to the ovum is the ability to avoid polyspermic fertilisation. The presence of the third category "gene duplicates" is explained by the classical theory of gene duplicate retention [Ohno 1970] in which the duplication releases one of the duplicates from negative selection and opens the possibility for this gene to evolve a new function - neofunctionalisation (this topic will be more thoroughly explored in section 2.5).

The hunt for adaptive evolution has been very active in our own species and has returned a number of interesting examples in the functional categories in which one might expect them, perhaps because it was there that the search efforts were concentrated. Genes involved in dietary adaptation, sensory systems (trichormatic vision and taste) have all been shown to have undergone positive selection. Positive selection has also been detected in two genes associated with brain size (ASPM and Microcephalin) and, both genes, when mutated, are known to cause primary microcephaly (a disease characterised by a severe reduction in brain size) [Vallender and Lahn 2004].

2.4. Gene duplication

So far, we have mainly considered genetic variation caused by nucleotide mutations between orthologs (sequences separated by a speciation event). Following a speciation event a gene in the genome of the common ancestor will become two separately evolving genes in the descendent species (orthologs). For the absolute vast majority of genes, these orthologs will be under negative selective pressure to retain the function present in the common ancestor. It is this general principle that is being applied when researchers, interested in a particular human protein, identify the ortholog in a model organism and perform experiments in the model system with the aim of extrapolating

the results back to humans. Positive selection does occur between orthologs, particularly in proteins involved in arms races such as immune system proteins or proteins expressed in gametes, however, such positive selection is thought to only rarely involve functional changes.

Large scale mutations may occur involving whole sections of DNA. Of particular interest are events that result in the duplication of a section of DNA as such events result in additional genetic material potentially containing a gene. Duplication events are key drivers of evolution as they create redundancy and, thus, the opportunity for one of the duplicates to escape the eye of negative selection and to functionally diverge. Such functional divergence may take several forms, one of which involves the evolution of new function through the rise to fixation of beneficial mutations that are positively selected for [Ohno 1970].

There are two basic types of duplication of genetic material that can occur in DNA, these are small-scale duplication (SSD) and whole genome duplication (WGD). Small-scale duplication involves the duplication of a section of a chromosome and may result in the duplication of one or more open reading frames. The duplication may also encompass the associated transcription-regulating sequences (transcription start site, transcription factor binding sites and other gene regulatory sequences, see Figure 2.1) and, thus, results in a functional duplicate. Whole genome duplication, on the other hand, is the result of the duplication of all chromosomes, resulting in the duplication of all genes in the genome and all regulatory regions.

2.4.1. Smaller-scale duplication (SSD)

There are several mechanisms that may cause small scale duplication. The three most relevant with respect to gene duplication are described here, in decreasing order of their likelihood of producing a functional duplicate.

The first and most likely mechanism to produce a functional small-scale duplicate is unequal crossing-over which may occur during meiosis. During meiosis, prior to the first meiotic division, the diploid germ cell precursor undergoes DNA replication and the homologous duplicated chromosomes pair up and undergo recombination (exchange of homologous sections of homologous chromosomes), also called crossing-over (see Figure 2.5). If the homologous chromosomes pair up correctly, homologous sections of DNA are exchanged but, if they pair up incorrectly (for example, due to some other homologous genomic feature), what is known as an unequal crossing-over event can occur. This may result in the duplication of a gene as described in Figure 2.8. Such a duplication has a relatively high likelihood of also duplicating the transcrip-

tional start site and transcription factor binding sites (TFBS) of the affected gene, thus resulting in a functional duplicate copy. An unequal crossing-over event may cause the duplication of one (tandem duplication) or several genes (segmental duplication).

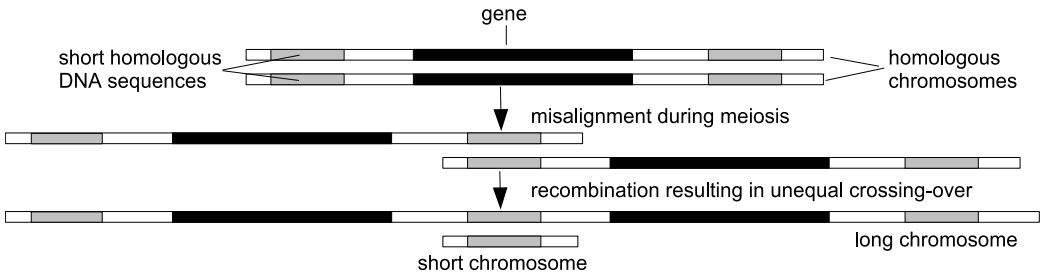

Figure 2.8.: Gene duplication by unequal crossing-over (adapted from [Page and Holmes 1996])

Another biological process that may result in gene duplication is retro-transposition [Walker et al. 1995]. Retrotransposons are sections of DNA that are able to make copies of themselves, usually via an RNA intermediary. DNA sequence transposes by first being transcribed into RNA by cellular RNA polymerases. A DNA copy of this RNA is then made using the reverse transcriptase enzyme. The DNA copy can then reintegrate into another site in the genome i.e. not necessarily in proximity to the source sequence. It is possible that a retrotransposon not only makes a copy of itself, but also copies adjacent sections of DNA which may contain genes. This is particularly likely if two transposons are located close to each other in the DNA sequence and a gene is located between them as the transposition mechanism may occasionally use the ends of two different elements (instead of the two ends of the same element) and thereby replicate the DNA between them [Alberts et al. 1997]. If a gene is affected in such a way by a retrotransposon, the protein coding section of the gene (possibly also accompanied by transcription factor binding sites) may get duplicated. There is then a remote possibility that the duplicate copy is expressed either because it was copied with its regulatory elements or because it was inserted next to functional TFBSs of another gene.

Finally, it is possible that mature mRNA transcripts from an expressed gene get reintegrated into the genome at another locus via the action of reverse transcriptase [Schacherer et al. 2004]. This results in a novel protein coding sequence without introns. Such a sequence is very unlikely to be transcribed as the original gene regulatory elements will not be present, but there is the remote possibility that a retro-transcribed mRNA comes under the control of the regulatory sequences of another gene or inde-

pendently evolves transcriptional capability and thus is expressed, see for example [McCarrey 1990].

Other mechanisms that may result in the duplication of sections of DNA containing genes, but which we have not described here, include DNA polymerase slippage and DNA-level transposition.

2.4.2. Whole genome duplication (WGD)

WGD is thought to occur through polyploidy (more than two sets of chromosomes). There are two types of polyploidy: allopolyploidy in which the polyploid originates by the fusion of the genomes of two different, but closely related, species; and autopolyploidy in which all the chromosomes are from the same species. The number of sets of chromosomes may be any number, but tetraploidy (four sets of chromosomes) is common as meiosis is not perturbed by this doubling in the number of chromosomes (this is therefore the situation we describe in the following section).

In a diploid organism, if two daughter cells which were produced at the end of mitotic telophase fuse into one, a tetraploid cell is produced. A tetraploid cell may also be produced by two DNA replications not intervened by mitosis. If a germ cell precursor is tetraploid, meiosis will produce diploid gametes and the union of two diploid gametes will produce a tetraploid zygote. Tetraploid zygotes in mammals occur with a non-negligeable frequency, but the condition is lethal [Carr 1967]. Even if the condition is not lethal, polyploids tend to be scarce in animal species due to the sex determining mechanism. When diploid organisms with the XY/XX sex determining mechanism become tetraploid, the male has to maintain the XXYY state and the female the XXXX state. During meiosis of the XXYY male, the four sex elements may pair off as the XX-bivalent and the YY bivalent, resulting in every gamete being XY. Thus, all zygotes produced by the mating of a tetraploid male and female will be of the XXXY type. If the XXXY type gives the male phenotype, then there will be no females. Alternatively, the XXXY may be sterile. Even if two XY bivalents are formed in male meiosis, in 50% of the cases X and Y will move to the same division pole at the first meiotic division, thus producing the XXXY type. Thus, polyploidy disturbs chromosomal sex determination [Ohno 1970]. The above explains why polyploidy evolution is rare in mammals, birds and reptiles. However, in amphibians and fish, the chromosomal determiners of the opposite sexes, the X and Y (for male heterogamety) and the Z and W (for female heterogamety) are still in an initial state of differentiation and may substitute for each other. This explains why polyploidy is observed in fish [Leggatt and Iwama 2003]

and amphibians [Ptacek et al. 1994]. Polyploidy is also common in plants for similar reasons [Bodt et al. 2005].

Most diploids that undergo a WGD and become tetraploids eventually revert to the diploid state (diploidization) as exemplified by *Arabidopsis thaliana* which, although no longer polyploid, is known to have undergone several relatively recent WGDs [Bowers et al. 2003]. A newly arisen autotetraploid has four homologous chromosomes. As long as four homologs get together to form a quadrivalent during meiosis, the four chromosomes would be randomly sorted into two sets of two at the end of the first meiosis. There is thus no possibility of functional diversification. The preferential formation of two separate bivalents is the prerequisite for diploidisation and this is thought to occur by the evolution of structural heterozygosity among the four homologous chromosomes [Ohno 1970]. Fish belonging to the suborder Salmonoidea (trout, salmon whitefish and graylings) appear to be autotetraploid species which have progressed towards the diploid state in various degrees via this mechanism [Ohno et al. 1968].

It has long been suggested that WGD events may be associated with important transitions, major leaps in evolution and adaptive radiations of species. In particular, Ohno's neofunctionalisation theory of gene duplicate retention was initially proposed to explain the adaptive radiation of vertebrates through two rounds of WGD in the ancestral chordate (referred to as 2R). Evidence for 2R is now strong [Dehal and Boore 2005] and as we shall see there is now mounting evidence that these events had an important influence on the gene content of vertebrate genomes (although the cause of retention is not limited to neofunctionalisation as originally thought).

2.4.3. Fixation of the duplication event

The duplication event occurs in an individual organism, but the duplication only becomes part of the species genome if it rises to fixation. If the initial duplication event is selectively neutral then it may rise to fixation by genetic drift which it will do with a probability that is proportional to the inverse of the effective population size. As we shall see, under many models of gene duplicate evolution, this is indeed the case, but under other models (e.g. dosage sensitivity or increased dosage) the duplication event might produce a fitness effect even immediately following duplication when the duplicate sequences have not diverged. Moreover, it has been suggested that if a gene is duplicated together with regulatory elements it might generally (regardless of any particular properties of the gene) have a negative fitness effect due to the metabolic cost of producing extra protein [Wagner 2005]. These fitness effects would affect the probability of fixation.

Irrespective of whether the initial duplication event has a fitness effect or not, it is generally assumed that if fixation occurs, it does so much faster than the resolution of the fates of the duplicate copies. Thus, most studies of the fates of gene duplicates consider fate determination as a separate step that occurs following the fixation process.

2.5. Gene duplicate retention

2.5.1. Pseudogenisation

Following fixation of a gene duplicate the genome will contain two copies of the duplicated gene with none, or very little, divergence between the two copies. Assuming that both copies are fully functional, then one of the two is redundant (although there is the possibility that only part of the gene was duplicated, e.g. the open reading frame without the TFBSs, in which case one copy is a pseudogene). This functional redundancy results in a reduction in the level of negative selective pressure that applied pre-duplication [Lynch and Conery 2000, Lynch and Conery 2003]. In most cases, this release from negative selective pressure will eventually lead to the fixation of a null mutation (mutation affecting either a coding or regulatory region that results in loss of function) by drift, as there is no loss of fitness if one of the copies pseudogenises.

Mutations that destroy (or simply debilitate) function occur by various mechanisms, examples include nucleotide substitutions, deletions, insertions, insertions of transposable elements, and unequal crossing-over between repeated transcription factor binding sites.

Pseudogenisation is clearly the fate of the majority of gene duplicates irrespective of whether they are the result of SSD [Lynch and Conery 2000, Lynch and Conery 2003] or WGD [Woods et al. 2005, Brunet et al. 2006, Kellis et al. 2004].

2.5.2. Neofunctionalisation

However, it is clearly the case that gene duplicates are retained since gene content varies across genomes and has clearly increased within specific lineages, see for example the Ensembl database of annotated vertebrate genomes [Birney et al. 2006]. Such discrepancies in gene number have long been suspected and it is this which lead Ohno to formulate his neofunctionalisation model [Ohno 1970]. The fundamental idea is that, although the majority of duplicates will pseudogenise due to the neutral fixation of a null mutation, some duplicates will be subject to beneficial mutations that confer a new function to the duplicate. These beneficial mutations might occur in coding DNA,

in regulatory sequence or in sequence controlling alternative splicing. Once one of the genes in the duplicate pair has neofunctionalised, there will be negative selective pressure to retain both copies. Because of the lack of genomic data at the time of its formulation, the theory that neofunctionalisation was driving gene retention remained for a long time untested against empirical data.

As we saw in sub-section 2.3.6, the large amounts of genomic data which have recently become available have revealed that adaptive substitutions occur on a large scale in *Drosopholia* between orthologous sequences, but are currently thought to occur on a more limited scale in chordate species. Key to the issue of whether or not this model may explain the retention of duplicates is the extent to which beneficial mutations occur at a sufficiently high rate in the period following duplication that the fixation of such a mutation outpaces that of a null mutation in a non-negligeable fraction of duplicate pairs. It is quite possible that this is the case, despite the low level of adaptive evolution observed in chordate orthologs, as the duplication event creates a release from negative selective pressure which does not happen for orthologous genes. Thus, the duplication creates a context in which a larger fraction of the total nucleotide mutation rate may be fitness enhancing.

Individual cases of neofunctionalisation following gene duplication have been identified [Bielawski and Yang 2001,Johnson et al. 2001,Maston and Ruvolo 2002,Rodríguez-Trelles et al. 2003, Wu 2005]. There is also more general evidence that gene duplicates are under positive selection [Moore and Purugganan 2003, Shiu et al. 2006], however it is unclear from these studies whether the positive selection is a result of the duplication itself (as in the increased dosage model, see section 2.5.6) or whether fitness enhancing functional changes may be involved.

Moreover, because of the uncertainty surrounding the rate of fitness enhancing mutations in gene duplicates, it has been suggested that neofunctionalisation may not be the mechanism driving the retention of duplicates, but that neofunctionalisation occurs later once the duplicate has been stabilised in the population through some other mechanism [Force et al. 1999]. In this case, neofunctionalisation may still be a prominent fate even though the adaptive nucleotide mutation rate may be very low, as retention is ensured by another process and the adaptive mutations have a long time period in which to accumulate free from competition from null mutations.

2.5.3. Subfunctionalisation

Once genomic data became available it became clear that the levels of retention of gene duplicates could be very high, in particular following WGD where retention rates can

easily reach 20% or more (e.g. X. laevis [Hughes and Hughes 1993], teleosts [van de Peer et al. 2003], maize [Ahn and Tanksley 1993]). The classical neofunctionalisation model was considered unable to explain these high levels of retention, due to the view that beneficial mutations are very rare, and the subfunctionalisation model was proposed as an alternative. This model, also refered to as the duplication-degeneration-complementation (DDC) model [Force et al. 1999], derives from the fact that many genes, particularly those involved in development and gene regulation, have multiple and independently mutable regulatory subfunctions which control timing and tissue specificity of gene expression. Taking into account this modularity of the regulatory regions of genes, the model demonstrates how degenerative mutations in complementary regions can lead to retention of both duplicate copies through the evolutionary requirement to retain all the regulatory regions of the original gene (see Figure 2.9). The fundamental observation is that if a gene has several independent regulatory modules controlling expression, for example in different tissues, then a null mutation to one regulatory region in one copy and a null mutation to another regulatory region in the other copy, will result in negative selection to retain both copies. As this model does not require beneficial mutations to explain the retention of both duplicates in a pair, it has been characterised as "near-neutral".

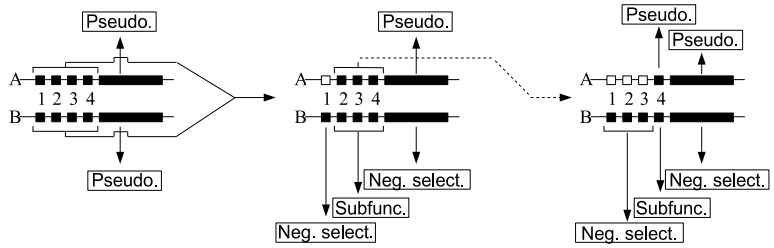

Figure 2.9.: Potential fates of a duplicate pair, adapted from [Force et al. 1999]

Small boxes: regulatory regions, *big boxes:* coding regions, *black box:* functional, *white box:* fixed null mutation

Arrows represent potential null mutations to regulatory or coding regions. The base of the arrow identifies the mutated region (if the base encompasses multiple regions, then this symbolizes a mutation to one of these regions) and the tips of the arrows point to a representation of the outcome if such a mutation is fixed in the population: *pseudo.:* pseudogenisation, *subfunc.*: subfunctionalisation, *neg. select.:* negatively selected against i.e. unlikely to reach fixation in the population.

The *dotted arrow* represents several intermediary states between the second and last state which are not drawn as the potential mutations from these states are identical to those in the second state. The model focuses on mutations fixed in the population, so the diagram shows the state of a single gamete.

Hughes [Hughes 1994] had previously proposed a similar model, sometimes referred to as "adaptive conflict". This model assumes an ancestral gene encoding two or more distinct, but pleiotropically constrained functions, in its protein coding sequence. Following duplication, the pleiotropic constraints are reduced and the duplicates specialise in distinct subsets of the functions of the ancestral gene, this division ensures retention

of both copies. This model is similar to the subfunctionalisation model in that there is a partitioning of functions between duplicates, but is also fundamentally different, in that the initial force driving retention of both copies is the accumulation of beneficial mutations which refine the subset of functions of each paralog. However, one can also envisage retention being driven by partitioning of function through deleterious mutations. Thus, although the subfunctionalisation model originally emphasised partitioning of expression patterns, it is also possible to think of the subfunctionalisation applying, for example, to active sites in the protein coding sequence. In this case, the probability of subfunctionalisation is an increasing function of the number of interactions or active sites (as opposed to the number of regulatory modules in the classical subfunctionalisation model).

Another variant of the subfunctionalisation model is quantitative subfunctionalisation. The idea is that mutations, instead of completely destroying function or expression, may merely have a debilitating effect. In this case, once the joint efficiency of a subfunction in both copies has been reduced to a level at which fitness begins to be reduced, any further degradation of the subfunction from either copy will be opposed by purifying selection [Force et al. 1999, Stoltzfus 1999].

The probability of subfunctionalisation is an increasing function of the number of regulatory modules, thus, there is reason to believe that the subfunctionalisation model might be at its most relevant as a mechanism of retention following WGD as all regulatory regions will be duplicated. This is to be contrasted with the situation following SSD where the extent to which regulatory regions are duplicated may be limited, particularly if such regions are widely dispersed in the genomic sequence [Kikuta et al. 2007].

2.5.4. Subfunctionalisation followed by neofunctionalisation

It was already noted in the original article [Force et al. 1999], as well as in subsequent publications [Lynch and Force 2000], that, subfunctionalisation may often occur in concert with neofunctionalisation: subfunctionalisation being the mechanism that ensures the initial retention, but the evolution of novel function being the ultimate fate of the retained duplicate. Since subfunctionalisation is driven by deleterious mutations which are known to be abundant, a model in which these two mechanisms are combined is not susceptible to the debate on the rate of beneficial mutations. The reasons one might expect a coupling of these two mechanismss are two-fold. First, subfunctionalisation stabilises the duplicate pair in the genome, thus increasing the probability that one of the genes is subject to rare beneficial mutations to a novel function. Second, the parti-

tioning of gene expression patterns may reduce the level of pleiotropic constraints that apply on single gene and, thus, allow a fine tuning of each member of the pair to its specific subfunction.

Using genome-wide protein-protein interaction data from yeast and gene (spatial) expression data from human, it has been possible to show that neither neofunctionalisation nor subfunctionalisation alone can adequately explain functional divergence of duplicate genes [He and Zhang 2005]. Instead, the analysis of this data reveals that rapid subfunctionalisation is accompanied by neofunctionalisation in a large proportion of duplicate genes. A simulation of duplicate protein evolution using lattice models has drawn similar conclusions [Rastogi and Liberles 2005].

2.5.5. Dosage balance

An alternative explanation for the high levels of retention following WGD is that dosage balance (stoichiometry) constraints apply to large numbers of genes.

The dosage balance model builds on the observation that the relative dosage of certain gene products in the cell have a critical effect on function, in particular for gene products forming complexes. For a complex formed by the binding of proteins A and B, there are numerous reasons why an excess of A might be deleterious: A could from homodimers with a different function from that of the AB heterodimer, it might be a regulatory subunit that competes with other regulatory subunits to bind the catalytic subunit B, it might be toxic by binding irreversibly to targets where AB should bind normally, or it could form toxic precipitates [Papp et al. 2003]. Recent evidence suggests that a major contributor to this balance effect are molecular complexes that function in various regulatory processes affecting gene expression [Birchler et al. 2005].

Thus, according to this model, a WGD should be followed by selective pressure to retain all genes encoding proteins that are dosage sensitive, as loss of one of the genes would create a dosage imbalance and is negatively selected against. Evidence that this might indeed be the case can be found for example in the *Arabidopsis thaliana* genome for which there is good evidence of several rounds of WGD [Bowers et al. 2003]. Genes retained in duplicate following WGD in this plant are not distributed evenly among Gene Ontology functional categories, instead there is an over-representation of genes involved in signal transduction and the regulation of transcription [Blanc and Wolfe 2004]. Genes with these functions are likely to be dosage sensitive and thus their overrepresentation among duplicates retained following WGD is supportive of the dosage balance model. Moreover, *Arabidopsis* genes retained in duplicate following one round of genome duplication are significantly more likely to be retained in dupli-

cate after a subsequent genome duplication [Seoighe and Gehring 2004], which is also consistent with this model.

It is important to note that, unlike the two first models of retention, dosage balance implies strong negative selection on both the coding and regulatory regions of the gene's DNA. Moreover, this model has the particularity that in the case of SSD of dosage sensitive genes, the duplication event is negatively selected against as the duplication of one gene of a set of mutually dosage sensitive genes has a negative fitness effect. Thus, although dosage balance may be an important cause of duplicate preservation following WGD, it is not a relevant model for the retention of duplicates following SSD.

As in the case of subfunctionalisation, it has been suggested that retention for stoichiometric reasons may be the initial cause of retention, but that the ultimate fate of the retained genes may be neofuntionalisation [Aury et al. 2006].

2.5.6. Robustness and increased dosage

The above models (neofunctionalisation, subfunctionalisation and dosage compensation) are widely considered to be the main models of gene duplicate retention as they are thought to have the potential to explain the retention levels observed either following SSD (neofunctionalisation), or WGD (dosage balance) or both (subfunctionalisation). But, there are also several other models of duplicate retention.

A very simple model is the "increased dosage" model. In this model, the ancestral gene is considered to have exhausted its mutational capacity to evolve higher levels of expression, thus a duplication of the ancestral gene is fitness enhancing as it doubles the maximum expression capacity. Because of this fitness effect, one would expect such a duplication event to rapidly rise to fixation through positive selection. However, given that retention levels are not insignificant following SSD and can be considerable following WGD, a large fraction of genes would need to be expressionally constrained in this way for this model to be an important explanation of gene duplicate retention. Although some genes may be retained through this mechanism, it seems unlikely that many genes suffer from "limited expressional capacity". Moreover, this model also implies that sequence divergence between duplicates would be limited or, at the very least, progress in a linear fashion. The fact that this is not the case [Lynch and Conery 2000, Lynch and Conery 2003], further weakens increased dosage as a major explanation of duplicate retention.

The observation that the deletion of a gene from a genome often has little phenotypic effect has lead to the suggestion that retained gene duplicates may be the source of this

robustness: the existence of gene duplicates may enable the deletion of one gene to be compensated for by a duplicate (the other prominent explanation for this compensation is that there exists alternative metabolic pathways and regulatory networks). Thus, it has been hypothesised that gene duplicates may be retained in genomes through negative selection for the buffering they provide against deleterious mutations.

The finding that, in *Sacchromyces cerevisiae,* at least a quarter of the gene deletions, that have no phenotype, are buffered by duplicate genes [Gu et al. 2003], was seen as supportive evidence. However, a later study in the multicellular *Caenorhabditis elegans* estimated a much lower contribution of gene duplicates to robustness [Conant and Wagner 2004], a result that has been confirmed by further work in *Sacchromyces cerevisiae* [He and Zhang 2006].

Moreover, even if gene duplicates were to explain a large fraction of the observed robustness to gene deletion, it does not necessarily follow that it is selection for robustness that drives duplicate copy retention. Indeed, gene redundancy may simply be an accidental by-product of gene duplication. Population genetic modeling showing that the selection pressure associated with robustness driven retention is very weak [Conant and Wagner 2004] would seem to confirm this view. And several empirical studies have concluded that negative selection for robustness is not an important driver of gene duplicate retention [Kuepfer et al. 2005, Ihmels et al. 2007].

2.5.7. Summary

There are two main mechanisms through which gene duplication can occur (small-scale and whole-genome duplication) and several models of gene duplicate retention, the most prominent of which are neofunctionalisation (either alone or in combination with subfunctionalisation), "pure" subfunctionalisation, and dosage balance. Other models of retention such as robustness and dosage balance may apply to individual cases, but are unlikely to be an important cause of retention. See Figure 2.10 for a graphical illustration of these models and Figure 2.11 for a summary of some of the details.

The mode of duplication determines a context for the duplicated gene and some models of gene duplication are not just dependent on features of the duplicated gene itself, but also on its context. Thus, some models of retention are associated with only one mode of duplication. For example, dosage balance is effectively only a candidate mechanism of retention following WGD.

The modes of duplication also vary in the levels of retention that they produce, with the retention rate following SSD being of the order of five percent in *Homo sapiens* [Hughes and Liberles 2007], whereas it can be several times higher following WGD.

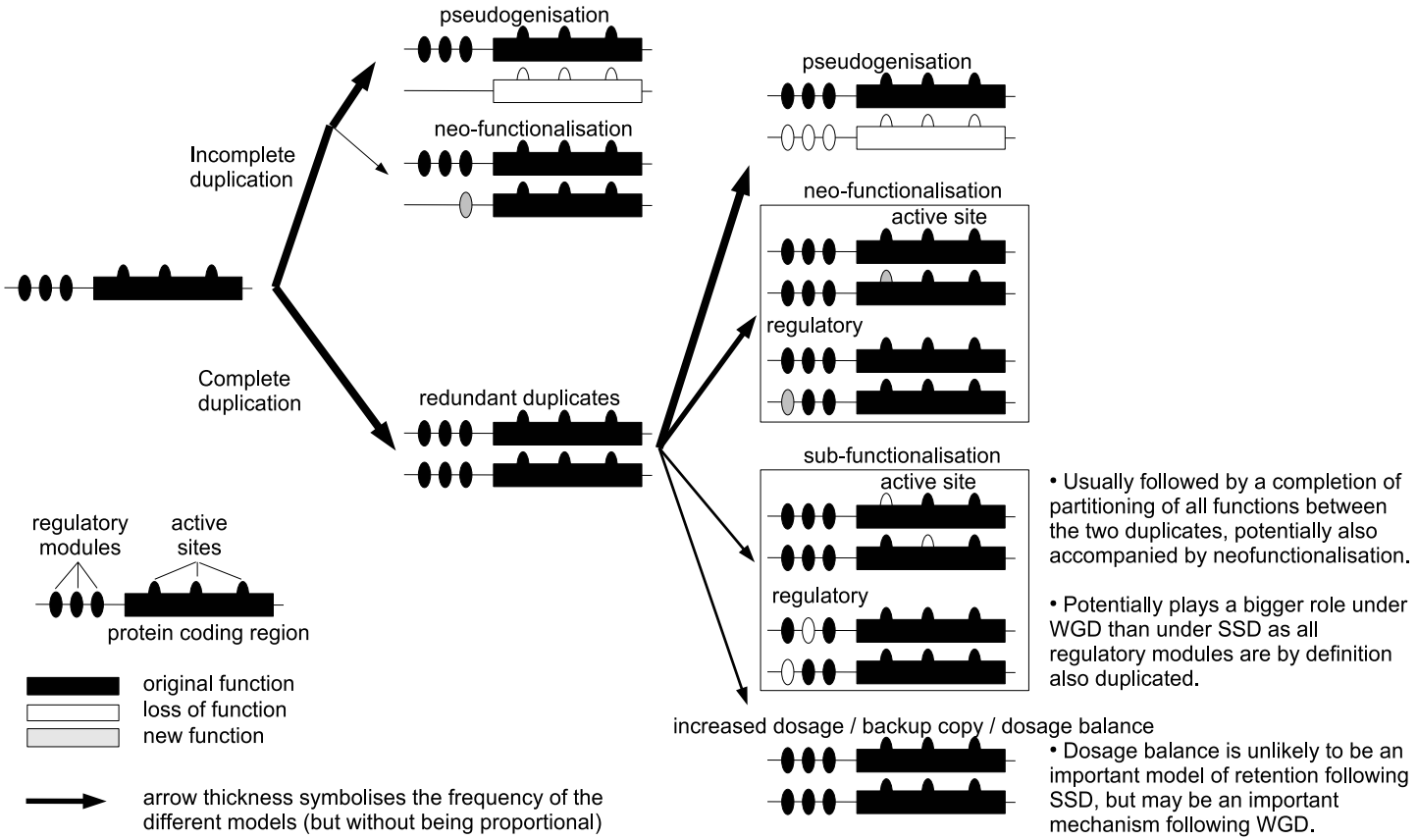

Figure 2.10.: Models of gene duplicate retention / evolution (adapted from [Moore and Purugganan 2005])

Model	Short description	Fitness effect of initial duplication event	Subsequent mutations: type of change (fitness effect)	Factors affecting the relevance of the model	Relevance of model	
					SSD	WGD
Pseudogenisation	One duplicate fixes a null mutation	Neutral	Loss of duplicate (neutral)	Rate of null mutation	By far the most common fate for a duplicate	
Neofunctionalisation	One duplicate fixes a fitness enhancing mutation (in regulatory or coding sequence, or gains a splice form)	Neutral	Beneficial mutation (positive)	Rate of beneficial mutation and magnitude of the fitness effect	Probably a common fate in SSD, perhaps preceded by subfunctionalisation. But unlikely to be able to explain the very high retention rates following WGD.	
Subfunctionalisation	Assuming genes have multiple regulatory modules or multiple active sites, duplicates undergo complementary loss of regulatory modules or active sites	Neutral	Complementary loss of regulations or active sites (near-neutral)	Number of regulatory modules or active sites per gene and the relative rates of null mutations to coding and regulatory regions of the sequence	Compatible with the retention rates following SSD. Lack of negative selective pressure on coding sequence and downward sloping hazard excludes "pure" subfunctionalisation as the main fate in SSD	Potentially plays an important role in WGD. First, because it is able to explain the high retention rate following WGD. Second, because all regulatory regions are guaranteed to be duplicated following WGD
Increasing dosage	Assuming that the transcriptional capacity of the ancestral gene is limited, the duplicate provides the ability to transcribe the gene at a higher rate	Positive	No changes	Extent to which gene expression levels may be limited by transcriptional capacity	Unlikely to be able to explain the retention rates following WGD and also SSD eventhough these rates are low. Is also imcompatible with a downward sloping hazard function. Also incompatible with release from negative selection on coding sequence.	
Backup copy / Robustness	Extra copies of genes results in increased robustness to null mutations to one of the copies	Positive	No changes	Magnitude of the fitness effect of carrying a backup copy	Same as for the "increased dosage model"	
Dosage balance	Certain genes are known to be dosage sensitive e.g. if they form a complex with other proteins or are involved in a metabolic pathway. If only one gene in a dosage sensitive "group" is duplicated it has a negative fitness effect, thus retention is unlikely. If all genes in a dosage sensitve group are duplicated, then the duplications are neutral and loss of one set of duplicates would be neutral, but loss of any individual members will be negatively selected against	SSD: negative WGD: neutral	Loss of one duplicate when all other genes have also been duplicated (negative)	Number of genes that are dosage sensitive and fitness effect of a dosage sensitive gene in dosage imbalance	Irrelevant as a model for retention in SSD.	Capable of explaining the high retention of duplicates following WGD

Figure 2.11.: Key features of models of gene duplicate retention

The only models that are considered to be able to produce such high levels of retention are subfunctionalisation and dosage balance.

Because of its neutrality, the "pure" subfunctionalisation model should be considered to be the null hypothesis when testing models of retention. Data on gene duplicates must first be shown to be inconsistent with this model, before the data is shown to be consistent with a selective model such as neofunctionalisation (positive selection) or dosage balance (negative selection).

It is important to recall that several of the models may apply to different types of genomic features e.g. both neo and sub-functionalisation may apply to regulatory regions, sequence controlling splice variants and to active sites. Given this, many types of data on gene duplicates have been examined to test the different models of retention, these include micro-array expression data, protein-protein interaction data, functional classes, and patterns of sequence divergence.

There appears to be a bias in the studies towards the study of retention following WGD. This bias is most probably driven by the fact that WGDs have the desirable characteristic of having duplicated all genes in the genome at a specific point in time, thus making for example the study of the retention of specific functional classes simple as all genes are known to have duplicated. There are currently no clear-cut results regarding the primary cause of retention following WGD, as some studies conclude in favour of subfunctionalisation while others favour dosage balance. The cause of this ambiguity is perhaps to be found in the fact that genes involved in regulation are the functional class which is most clearly over-retained following WGD (as compared to SSD). Genes involved in regulation have both complex regulation making them susceptible to subfunctionalisation and a tendency to be dosage sensitive, thus disentangling subfunctionalisation from dosage balance can be difficult.

The evidence that exists on the fate of duplicates following SSD suggests that they undergo some form of neofunctionalisation, with subfunctionalisation probably playing a part in the initial retention in a non-negligeable fraction of cases.

2.6. Gene families

2.6.1. Concept

As we have already seen, a gene may duplicate, resulting in a duplicate pair where both genes are initially identical. The most likely outcome of the subsequent evolution is that one of the genes looses its functionality and pseudogenises, but there is a small chance of retention. If both genes are retained, they will not remain identical for long

as selectively neutral non-synonymous mutations are bound to accumulate between the two copies whatever the mechanism of retention and, if retention is driven by neofunctionalisation, there is also the possibility of more radical non-synonymous mutations (see Figure 2.12 for a tree representation of gene family evolution).

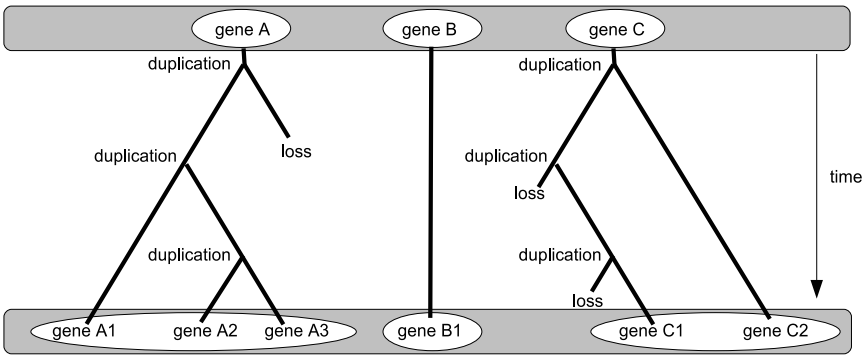

Figure 2.12.: Evolution of gene families (circles) within a genome

In most cases, even given neofunctionalisation and hundreds of millions of years of evolution, it will still be possible to detect similarity between the two sequences and from this to infer the common ancestry (homology). The reason this is possible is that, even in the case where a neofunctionalisation event separates two duplicates, the tertiary structure of the encoded proteins is usually maintained. This constraint on structure translates into constraints on the amino acids that are key to determining that structure. Thus, even in the absence of a structure for each of the duplicates, homology can be detected through sequence. There is debate about the extent to which neofunctionalisation drives duplicate retention but, when it does occur, it is usually considered to not involve fundamental changes in function, but rather fine tuning, e.g. the evolution of a slightly different affinity or specifity of the binding site of the duplicated protein. Thus, detectable similarity (and therefore homology) between two genes in a genome usually corresponds to shared function, and the definition of gene families based on detectable similarity satisfies both phylogenetic and functional considerations.

However, it can also be the case that a sequence diverges to such an extent that homology is no longer detected, for example if a gene undergoes a radical neofunctionalisation. In this case, although there is common ancestry, our limited insight into the true evolutionary process will lead to the genes being assigned to different families. From a phylogenetic perspective this is incorrect but, from a functional perspective, this is a desirable outcome as radical change in sequence is very likely to be driven by fundamental change in function. Thus, the equation of detectable homology with shared function may be a good rule of thumb for building gene families, as the resulting families will usually constitute a good evolutionary and functional classification of

the protein content of a genome. It is nevertheless important to remember that this is only an approximation: there are many examples of sequences without any detectable homology that have the same function, and of sequences with almost no sequence divergence that have different functions, see for example [Gaucher et al. 2003].

2.6.2. Similarity-based methods for building families

The sequence similarity approach to building gene families requires a measure of the similarity between all pairs of sequences, a similarity cutoff and a clustering algorithm to resolve the boundaries between gene families. Similarity measures are usually e-values from BLAST searches [Altschul et al. 1997] or PAM (point accepted mutation) distances [Dayhoff et al. 1978]. Clustering algorithms such as single or complete linkage are often applied. The advantage of complete linkage is that it ensures that there is a maximum distance between any two sequences in a cluster, whereas single linkage does not offer such a guarantee and thus produces a clustering in which the relatedness of members is less well defined. For both methods, the issue of where to set the similarity cut-off remains.

An alternative method that does not require a similarity cut-off is the Markov cluster algorithm [Enright et al. 2002]. This method does require parameters that affect the granularity of the clustering, but they are not directly related to the similarity of sequences within a cluster. This method also has the advantage of taking a more holistic consideration of all similarity measures simultaneously, rather than the sequential approach used when performing hierarchical clustering. Ensembl [Birney et al. 2006], a major repository for the annotation of vertebrate genomes, uses this method to build gene families.

2.6.3. Structure-based methods for building families

Another approach, which is more firmly anchored in biological reality, is to use structural information to hierarchically classify proteins, or more specifically, protein domains as this is the fundamental unit of structure. There are two main efforts to classify protein structures, SCOP - Structural Classification of Proteins [Andreeva et al. 2004] takes a more manual approach, while CATH [Greene et al. 2007] seeks to automate the classification process. Both databases have four main levels of classification which, although not identical, correspond roughly to each other. The higher levels are connected with similarity at the structural level as it is this type of similarity which remains detectable when sequence similarity no longer is, while the lower levels are connected

to homology that is detectable through sequence similarity. In the case of SCOP, the levels are, from top to bottom: class (all alpha helices, all beta sheets, α and β mixed, α and β segregated); folds (the same major secondary structures in the same arrangement and with the same topological connections); superamilies (low sequence similarity but very similar structure); families (high sequence similarity). Given a sequence it is possible to search these databases and determine which family the sequence is likely to belong to. This effectively also makes a prediction of the sequence's structure, even if the structure of the query sequence is not known.

2.6.4. Phylogenetic methods for building families

In the case where the sequences to be clustered come from a specific set of species it is possible to take a phylogenetic approach to clustering. For each gene in the outgroup species, one searches for the most similar sequence in all other species (best hit sequence). Once this sequence is identified, all sequences from other species, that are more similar to the best hit than to the outgroup sequence, are included in the family defined by the outgroup and best hit sequence [Dehal and Boore 2005, Blomme et al. 2006]. This method has the advantage of having a very clear phylogenetic definition. Its disadvantage is that the outgroup may have undergone some gene loss and, thus, for the orthologs of this sequence a gene family will not be defined. Moreover, this method like all others is not immune from the problems caused by highly divergent sequences.

2.6.5. Power-law distribution of gene family size

Despite the fact that there are different methods for building clusters of homologous genes and that the granularity of the clustering will always contain a level of arbitrariness, a clustering of all protein coding genes in a genome, irrespective of the method used, produces many small clusters and few large clusters [Yanai et al. 2000, Harrison and Gerstein 2002]. The functional form that best fits the data is the power-law [Huynen and van Nimwegen 1998, Luscombe et al. 2002]: $N = aF^b$ where F is the family size and N is the number of families of this size or, taking the natural logarithm, $\ln(N) = \ln(a) + b.\ln(F)$ i.e. a linear relationship on a log-log plot. The exponent b is usually in the range -4.0 to -2.75 and there is a weak positive correlation between the exponent and the logarithm of the number of genes in the genome [Huynen and van Nimwegen 1998].

A power-law distribution of gene family size is one of the most universal features of the gene content of genomes. An interesting topic of investigation, which we pursue in

one of the papers of this thesis, is determining the evolutionary mechanisms that drive this ubiquituous observation.

2.6.6. Phylogenetic trees

Once gene families have been defined, it is possible to compute a phylogenetic tree for the family. A phylogenetic tree is an attempt to infer the evolutionary history of the gene family (see Figure 2.12). There are many methods for carrying out such an inference from very simple methods based on hierarchical clustering of distances between sequences to computationally-demanding probabilistic methods such as maximum likelihood [Yang 1997] and bayesian inference [Ronquist and Huelsenbeck 2003]. Generally speaking, the more advanced methods produce a better inference [Yang 2006], but the improvement is often marginal. Once a gene tree has been computed from extant sequences, it is possible to infer speciations, duplications and losses on the tree through what is referred to as a gene tree / species tree reconciliation [Zmasek and Eddy 2001, Arvestad et al. 2003, Berglund et al. 2005].

3. Contributions

The core of this thesis consists of four papers that are all connected to the study of how the gene content of genomes is shaped by the processes of gene duplication, sequence divergence and gene loss. The purpose of this chapter is to position these papers in the context of the current knowledge as summarised by chapter 2. More specifically, additional background information is presented and the contribution of the paper to existing knowledge is explained. A summary of key results and ideas for further work are also presented. However, it may not be straightforward to understand the details of the discussion on further work before reading the papers (see appendix). It is therefore recommended that the reader return to this section once the papers have been read.

The papers included in this thesis are:

- Paper I: Timothy Hughes and David A. Liberles (2007). The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than sub-functionalisation. Journal of Molecular Evolution 65:574-588.

- Paper II: Timothy Hughes and David A. Liberles (2007). The power-law distribution of gene family size is driven by the pseudogenisation rate's heterogeneity between gene families. Submitted to Gene.

- Paper III: Timothy Hughes and David A. Liberles (2007). The whole genome duplications in the ancestral vertebrate are detectable in the distribution of gene family sizes of tetrapod species. Submitted to Journal of Molecular Evolution.

- Paper IV: Timothy Hughes, Diana Ekman, Himanshu Ardawatia, Arne Elofsson and David A Liberles (2007). Evaluating dosage compensation as a cause of duplicate gene retention in Paramecium tetraurelia. Genome Biology 8(5):213.

Two additional papers were produced and published in the course of the doctoral studies that lead to this thesis. Both papers fall within the field of evolutionary biology, but they were not included in the thesis as they have only a tangential relationship to the four other papers:

- Timothy Hughes, Young Hyun and David A. Liberles (2004). Visualising very large phylogenetic trees in three dimensional hyperbolic space. BMC Bioinformatics 29(5):48.

- Timothy Hughes and Tor Erik Rusten (2006). Origin and evolution of self-consumption: Autophagy. In, Origins and evolution of eukaryotic endomembranes and cytoskeletons, edited by Gaspar Jekely. Landes Bioscience.

3.1. Gene duplication and loss (paper I)

3.1.1. Context

Biology is primarily concerned with the study of all phenotypic aspects of living organisms and evolutionary biology is more specifically interested in elucidating how different phenotypes evolved. The introduction chapter has hopefully made clear that an important source of phenotypic change/novelty are retained gene duplicates. Three of the key processes in the retention of gene duplicates are gene duplication, gene loss, and gene sequence divergence between the genes in the pair. With the advent of whole genome sequencing, quantitative characterisations of these processes have been made possible [Lynch and Conery 2000, Lynch and Conery 2003]. Lynch and Conery's first paper is a seminal paper, in which all three processes are modeled. Essentially, the rates of gene duplication, gene loss and sequence divergence (as measured by the number of accumulated non-synonymous substitutions per non-synonymous site, d_n) are estimated by using the accumulation of synonymous substitutions per synonymous site between the genes in a pair (d_s) as a proxy for time since duplication. The idea of using d_s as a proxy for time is what enables the estimation of the rates for processes that we are not able to monitor through time by direct observation. However, following close inspection, it became clear that the models of sequence divergence and, more importantly, gene loss could be significantly improved.

3.1.2. Results

We therefore proceeded to build a dataset of recent gene duplicates for several fully-sequenced mammalian genomes (*Homo sapiens*, *Canis familiaris*, *Mus musculus*, *Rattus norvegicus*) to which we fitted what we consider to be improved models.

The results of the improved modeling broadly confirm the results of Lynch and Conery. We find that the duplication rate is of the same order of magnitude as the rate of mutation per nucleotide site. We also find that half lives of gene duplicates are very

short, of the order of 0.03 d_s and the rates of accumulation of non-synonymous substitutions between duplicated genes are similar to the earlier results [Lynch and Conery 2000, Lynch and Conery 2003]. The two significant differences are the model for the variance of the accumulation of non-synonymous substitutions and the Weibull function modeling of the loss rate.

The model of the variance is important for several reasons. First, because it enables an improved modeling of the duplicate gene data that can correct for the heteroscedasticity of d_n: the variance of d_n is not constant, but increases with d_s (the proxy for time since duplication). Second, because this characterisation of the variance can be used to differentiate between the modes of retention of gene duplicates (see section 3.2). Third, because it can be employed to differentiate functional gene duplicates from pseudogenes (see ideas for further work in sub-section 3.1.3).

The modeling of gene loss using a Weibull function shows that the rate of gene duplicate loss as a function of time since duplication is not a constant, but rather a decreasing function. This has implications for the likely mode of retention of gene duplicates but, it is also important in its own right as the correct functional form is needed both for modeling studies, which usually assume that the death rate is a constant, [Arvestad et al. 2003, Demuth et al. 2006] and also for an accurate estimate of the half-life of a gene duplicate.

3.1.3. Ideas for further work

The characterisation of the variance of the accumulation of non-synonymous substitutions can potentially be used to refine the way in which the d_n/d_s method is used for detecting pseudogenes in genome annotations.

Pseudogenes are complete or partial copies of genes that do not code for functional polypeptides. This lack of function is either a result of failure of transcription or translation, or a result of the production of a protein that does not have the same functional repertoire as the protein encoded by the normal paralog gene [Mighell et al. 2000]. Pseudogenes are generally divided into two main categories (processed and non-processed) which correspond to the mechanism through which they were generated.

Non-processed pseudogenes are generated through partial or complete segmental duplication of genomic DNA by unequal crossing-over. If the duplication is only partial, e.g. only part of the coding sequence is duplicated or gene regulatory regions were not duplicated, then the duplicate is likely to be a pseudogene “from birth”. On the other hand, if the duplication is complete, the duplicate will initially be functional. There is

a small probability that the duplicate remains functional, but the most likely outcome is that the duplicate pseudogenises [Lynch and Conery 2000], a result which has been confirmed in paper I.

Processed pseudogenes are retrotransposed mature mRNAs (also referred to as retro-pseudogenes) and far outnumber non-processed pseudogenes [Torrents et al. 2003]. They are typically characterised by an absence of 5'-promoter sequence and introns, and the presence of a 3'-polyadenylation tract. Due to these features, in the vast majority of cases, these retro-transposed sequences are pseudogenes from inception but, in a few instances, the retro-transposed sequence is maintained as a functional intronless gene [McCarrey 1990].

A common strategy for annotating genomes is to search the genome sequence for sections of DNA with protein-coding features. Such a search inevitably returns many pseudogenes as well as the functional genes which are of primary interest: for example, there is evidence that about one fifth of all annotated genes in C. elegans were in fact pseudogenes [Mounsey et al. 2002]. Processed pseudogenes have features, such as absence of introns and presence of 3'-polyadenylation tract, that enables them to be fairly accurately identified as non-functional. Non-processed pseudogenes lack these features and approaches usually evolve around searching for stop codons or frameshifts that disrupt the original reading frame, predicted by homology to known protein sequences. However, a sequence might be a pseudogene even if it lacks these features, for example, the lack of function may be due to the substitution of a functionally critical amino acid or due to the lack or disruption of promoters.

A feature that is shared by all pseudogenes is their lack of function and thus their freedom from the constraints of selection, thus over time they accumulate substitutions in a neutral fashion $d_n/d_s \simeq 1$ whereas functional genes, even if they are recent duplicates, tend to have $d_n/d_s < 1$ when this is calculated over the full length of the sequence, as shown in paper I. This feature of pseudogenes has been exploited to improve their detection [Torrents et al. 2003] but, in that study, the ratio was treated as a constant that does not change with time since duplication. Our characterisation of the accumulation of d_n as a function d_s shows that functional gene duplicates accumulate non-synonymous substitutions at a rate that decreases with time since duplication. Moreover, we have produced a quantitative model of the variance of d_n. It should thus be possible to exploit this additional information to refine the use of d_n/d_s as a criterion for identifying pseudogenes.

A method that is superior to the existing d_n/d_s ratio method [Torrents et al. 2003] has been developed [Coin and Durbin 2004], however this method requires detectable homology to a Pfam domain, a condition that is only satisfied for approximately 60

percent of sequences. Thus, a refinement of the d_n/d_s method would have the definite advantage of being more widely applicable and might even compete with the Receiver Operating Curve (ROC) of the Pfam based method.

3.2. Models of duplicate retention (paper I)

3.2.1. Context

From the background discussion in chapter 2, it should be clear that the most prominent models of duplicate gene retention are fundamentally different between SSD and WGD. Following SSD, the main models for gene duplicate retention are subfunctionalisation which may be considered as the neutral null hypothesis and neofunctionalisation. Dosage balance is not applicable as a model of retention following SSD as dosage sensitive duplicates are negatively selected against. Following WGD, subfunctionalisation can again be considered as the null hypothesis, but neofunctionalisation alone would be very unlikely to produce the high levels of retention observed following WGD. Dosage compensation, however, is consistent with high retention levels and is the most relevant alternative to subfunctionalisation in the WGD context. It is nevertheless important to recall that a gene may be stabilised in the genome by subfunctionalisation or dosage compensation, but subsequently undergo neofunctionalisation [Force et al. 1999, He and Zhang 2005, Rastogi and Liberles 2005].

In the literature, there is a strong focus on the study of retention following WGD and far fewer studies addressing SSD retention. This may be due to the fact that SSD has the disadvantage relative to WGD that not all genes are duplicated at one point in time, so it is difficult to study whether particular functional categories are preferentially retained or not. On the other hand, because SSD is an ongoing process and not a one-off event, there are techniques for producing a picture of how SSD duplicates evolve, something which is difficult to achieve with WGD duplicates. It is this feature of SSD that we have exploited in paper I to quantify two key characteristics of the duplicates: the rate of accumulation of non-synonymous substitutions per non-synonymous site and the rate of pseudogenisation. By deriving the predictions of the SSD retention models for these characteristics, we are able to test which model is the most consistent with the data.

Both neofunctionalisation and subfunctionalisation undoubtedly account for large numbers of the gene duplicate retentions, however, which is the dominant mode remains unclear. This is primarily due to the difficulty in directly identifying the changes in the sequence features that cause retention. In the case of the classical subfunctional-

isation model [Force et al. 1999], it is the complementary loss of regulatory modules that drives retention, but such sequences are difficult to identify computationally due to their short and degenerate nature. In the case of neofunctionalisation, it is the accumulation of beneficial mutations that change or add functionality that drive retention, by for example changing the specificity of an active site, but such mutations are notoriously difficult to identify due to their episodic and localised nature [Golding and Dean 1998].

Studies of small-scale duplication and retention have therefore used "indirect" evidence such as the species-wide levels of nucleotide polymorphisms in the progenitor and/or duplicate gene copies [Moore and Purugganan 2003] or the levels of duplicate gene retention in species with different population sizes [Shiu et al. 2006]. These studies have concluded that duplicates are under positive selection to be retained, but it remains unclear whether this is simply a result of duplication event itself (as in the increased dosage or robustness models) or whether one of the duplicates may have neofunctionalised. A study which is clearer on this point [He and Zhang 2005] uses functional genomic data (genome-wide protein-protein interaction data from yeast and gene spatial expression data from human) to clearly show that neofunctionalisation is very common, but that it tends to be preceeded by subfunctionalisation. In other words, neofunctionalisation is the ultimate fate (or mode of evolution), but subfunctionalisation is the mode of retention.

3.2.2. Results

Our derivation of the predictions of the different models of SSD retention show that the neofunctionalisation model is most consistent with the data on the pseudogenisation rate and the rate of sequence divergence of duplicate pairs. However, we are not able to clearly distinguish between neofunctionalisation from subfunctionalisation followed by neofunctionalisaton. These results are thus in agreement with existing results on small-scale duplicate retention [Moore and Purugganan 2003, He and Zhang 2005, Shiu et al. 2006].

That "pure" subfunctionalisation (subfunctionalisation without neofunctionalisation) does not play an important role, is perhaps not that suprising. Subfunctionalisation involves partitioning of function and a gene only has a limited number of functions, thus there is a limit to how many times these can be partitioned. At some point new functions must evolve for partitioning of function to be able to continue to play a role as a cause of retention. As there is currently little concrete evidence that retained whole genome duplicates are the source of novel function, this leaves only orthologs and small-scale

duplicates as potential sources. Our results indicate that retained small-scale duplicates do indeed probably play this role.

3.2.3. Ideas for further work

A general characteristic of studies of gene duplicate retention (both SSD and WGD) is that researchers choose a set of features of gene duplicates (e.g. sequence divergence rates, or expression patterns, or protein-protein interactions) and determine which model of retention is most consistent with the data. However, only a subset of all relevant features are analysed in any one study. An alternative approach would be to include all relevant features in one study. The downside of this approach is that it may be difficult to conclude in favour of a particular model of gene duplicate retention but, the upside of this more holistic approach would be that one may obtain a better quantification of how different features of genes contribute to the probability of retention.

A natural framework for such a study is survival analysis, which we use when modeling the pseudogenisation rate in paper I. Survival analysis is a commonly used technique in medical statistics to study the effect of a treatment. One of the key differences between medical survival analysis and the analysis that we performed in paper I is that additional covariates to time are included in the model, usually characteristics of the patient such as lifestyle, health and genetics. This makes it possible to test which covariates significantly contribute to the survival outcome. An existing theoretical framework therefore exists for extending the modeling of pseudogenisation to include characteristics of the gene duplicates.

However, in the case of gene duplicates, we do not directly observe duplicates through time in the same way as patients can be monitored for a medical survival analysis [Collett 2003]. However, we are able to determine whether a gene is a pseudogene or not (survival outcome), and it is possible to determine many of the features of genes and of pseudogenes. Assuming that it is possible to extract from the set of pseudogenes those that were functional on duplication (and exclude those that were not e.g. processed pseudogenes) then it should be possible to perform a logistic regression in which the dependent variable is the gene's state (functional or not, i.e. pseudogene) and the independent variables are features of the gene e.g. length, fold, sequence divergence, functional class, number of expression domains, number of interaction partners. The independent features would of course have to also include time since duplication, as this is the primary determinant of a duplicate's functional state. The details of how such a study may be carried out have not been fully explored, but a study embracing

this more holistic approach has the potential to make a contribution to the study of gene duplicate retention.

3.3. The distribution of gene family size (paper II)

3.3.1. Context

The power-law distribution of gene family size is one of the most (if not the most) ubiquituous feature of the gene content of genomes as it has been observed in species from all three domains of life (see subsection 2.6.5). It is observed for all fully-sequenced genomes and is also observed across multiple genomes if the species are evolutionarily distant [Enright et al. 2003]. The evolutionary mechanisms that are candidates as drivers of this characteristic of genomes are gene duplication, gene loss and sequence divergence (in bacteria, lateral gene transfer may be added to the list). Several papers have shown that it is possible to define theoretical evolutionary models that replicate the power-law distributions of genes or protein folds [Huynen and van Nimwegen 1998, Yanai et al. 2000, Qian et al. 2001, Karev et al. 2002, Kamal et al. 2005], however, none of these models have been validated using genomic data. In paper I, we have formulated models for all three key processes and fitted and tested the models using genomic data. In paper II, we use these models to explore through simulation their relationship with the power-law.

3.3.2. Results

Using a model of homologous gene evolution, we show that the power-law distribution of gene family size is driven by the pseudogenisation rate's heterogeneity across gene families and its correlation within families. Moreover, we show that gene duplication and pseudogenisation are necessary and sufficient for the emergence of the power-law, and are thus the key forces shaping the size of gene families. Our results are in agreement with one of the theoretical models [Huynen and van Nimwegen 1998] and, thus, complement those results by showing that the theoretical results are anchored in biological reality.

3.3.3. Ideas for further work

An interesting prediction of the model is that larger families should consist of genes with high duplicate retention rates. The models of gene duplicate retention make dif-

ferent predictions as to what duplicate features cause a high retention rate. The neofunctionalisation model predicts that proteins with plastic folds, which are able to accomodate many functions, have a higher probability of retention, while the subfunctionalisation model suggests that the number of regulatory regions or the number of interaction partners should be highly correlated with retention. It should be possible to investigate whether this prediction is satisfied by testing whether there is any significant correlation between gene family size and some of these features of their members.

One shortfall of our modeling is that we do not have an empirical estimate of the variance of the hazard function across gene families, instead we arbitrarily set this value and show how varying levels of this parameter affect the emergence of the power-law. Thus, this aspect of the model is not validated against genomic data. This does not detract from the fact that the biological anchoring of our model is greater than in previous studies of the power-law but, nevertheless, it would be interesting to quantify the level of pseudogenisation heterogeneity across gene families.

3.4. Hazard shift and WGD (paper III and IV)

3.4.1. Context

Whole genome duplication is a major genomic event which, at least initially, doubles the size of all gene families. If retention rates were as low as they are following SSD, i.e. of the order of a few percent [Lynch and Conery 2000, Lynch and Conery 2003], then the effect of WGD on the overall organisation of the genome would be minor. However, retention rates following WGD are much higher than following SSD, often of the order of 20 percent as in the fish specific WGD [Jaillon et al. 2004, Woods et al. 2005, Brunet et al. 2006] and even higher in other species [Hughes and Hughes 1993, Ahn and Tanksley 1993].

When clustering the protein coding sequences of five tetrapod species, we were intrigued to observe a strong deviation from the power-law distribution of gene family size. It was Ohno [Ohno 1970] that originally hypothesised (on rather weak evidence) that the ancestral vertebrate was subject to two rounds of WGD and a convincing case has now been built in favour of this claim [Dehal and Boore 2005]. Given the high retention levels following WGD, we hypothesised that the deviation from the power-law that we observed may have been the result of two whole genome duplications. In paper III, we extend the model of homologous gene family evolution developed in paper II to investigate whether our hypothesis is correct.

3.4.2. Results

We find that, in order to replicate the features of the empirical distribution, the simulation model must incorporate two WGD events. In addition, these WGDs must be such that a significant proportion of the gene duplicates generated in the WGDs have a higher retention rate than they do following small-scale duplication (SSD) and this shift affects primarily genes that have a very low probability of retention following SSD. Such a shift in retention is most likely to be driven by a shift in the pseudogenisation rate (or hazard shift). This requirement is consistent with what is known about duplicate retention following a WGD, namely that genes belonging to specific functional classes, such as genes regulating transcription, are much more likely to be retained following WGD than SSD [Blanc and Wolfe 2004, Maere et al. 2005, Blomme et al. 2006]. We conclude that the deviation from the power law, that we observe in the empirical data, is the result of the two WGDs that occurred in the ancestral vertebrate. This implies that the two ancient WGDs continue to have a structural effect on gene families approximately 500 million years after the initial events. The capacity of whole genome duplications to fundamentally change the architecture of gene families in a profound and lasting way is consistent with the observed correlation between WGDs and important evolutionary transitions.

The shift in the probability of retention, that we find is required, is consistent with both the subfunctionalisation model and the dosage balance model. It is, in fact, very difficult to distinguish between these two models as a source of duplicate retention following WGD. This issue is discussed at length in paper IV which reviews the hypothesis that dosage balance plays a major role in gene duplicate retention following WGD in *Paramecium tetraurelia*.

3.4.3. Ideas for further work

In our modeling of the evolution of gene families that span multiple species in paper III, we lack an empirical estimate of the rate at which orthologs are lost. It is generally assumed that the loss of orthologs is much less probable than the loss of paralogs. In paper I, we produced a quantification of the rate at which duplicates pseudogenise, but an equivalent estimate for orthologs is to our knowledge not available. A possible approach to this problem is to define gene families from the fully sequenced genomes of several species, build phylogenetic trees for these families and then infer orthology, paralogy and loss, through the reconciliation of the gene trees with the species tree [Zmasek and Eddy 2001, Arvestad et al. 2003, Berglund-Sonnhammer et al. 2006]. From such trees it should be possible to produce a quantification of ortholog loss.

Another obvious area for further research is to better characterise the evolutionary forces driving duplicate retention following WGD. As we discuss in paper III and IV, both subfunctionalisation and dosage balance are relevant models, but disentangling the two is difficult due to the overlaping predictions of the models.

Bibliography

[Ahn and Tanksley 1993] Ahn S, Tanksley SD (1993) Comparative linkage maps of the rice and maize genomes. Proc Natl Acad Sci U S A 90:7980–7984.

[Alberts et al. 1997] Alberts B, Bray D, Johnson A, Lewis J, Raff M, Roberts K, Walter P (1997) Essential cell biology. Garland Publishing.

[Altschul et al. 1997] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

[Andolfatto 2005] Andolfatto P (2005) Adaptive evolution of non-coding DNA in Drosophila. Nature 437:1149–1152.

[Andreeva et al. 2004] Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32:D226–D229.

[Anisimova and Liberles 2007] Anisimova M, Liberles DA (2007) The quest for natural selection in the age of comparative genomics. Heredity x:x.

[Arvestad et al. 2003] Arvestad L, Berglund AC, Lagergren J, Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics 19 Suppl 1:i7–15.

[Aury et al. 2006] Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, Arnaiz O, Billaut A, Beisson J, Blanc I, Bouhouche K, Câmara F, Duharcourt S, Guigo R, Gogendeau D, Katinka M, Keller AM, Kissmehl R, Klotz C, Koll F, Mouël AL, Lepère G, Malinsky S, Nowacki M, Nowak JK, Plattner H, Poulain J, Ruiz F, Serrano V, Zagulski M, Dessen P, Bétermier M, Weissenbach J, Scarpelli C, Schächter V, Sperling L, Meyer E, Cohen J, Wincker P (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature 444:171–178.

[Berglund et al. 2005] Berglund AC, Wallner B, Elofsson A, Liberles DA (2005) Tertiary windowing to detect positive diversifying selection. J Mol Evol 60:499–504.

[Berglund-Sonnhammer et al. 2006] Berglund-Sonnhammer AC, Steffansson P, Betts MJ, Liberles DA (2006) Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. J Mol Evol 63:240–250.

[Bielawski and Yang 2001] Bielawski JP, Yang Z (2001) Positive and negative selection in the DAZ gene family. Mol Biol Evol 18:523–529.

[Birchler et al. 2005] Birchler JA, Riddle NC, Auger DL, Veitia RA (2005) Dosage balance in gene regulation: biological implications. Trends Genet 21:219–226.

[Birney et al. 2006] Birney E, Andrews D, Caccamo M, et al. (51 co-authors). (2006) Ensembl 2006. Nucleic Acids Res 34:D556–D561.

[Blanc and Wolfe 2004] Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16:1679–1691.

[Blomme et al. 2006] Blomme T, Vandepoele K, Bodt SD, Simillion C, Maere S, van de Peer Y (2006) The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol 7:R43.

[Bodt et al. 2005] Bodt SD, Maere S, de Peer YV (2005) Genome duplication and the origin of angiosperms. Trends Ecol Evol 20:591–597.

[Bowers et al. 2003] Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438.

[Brunet et al. 2006] Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol Biol Evol 23:1808–1816.

[Bustamante et al. 2005] Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG (2005) Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157.

[Carr 1967] Carr DH (1967) Chromosome anomalies as a cause of spontaneous abortion. Am J Obstet Gynecol 97:283–293.

[Chamary et al. 2006] Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98–108.

[Coin and Durbin 2004] Coin L, Durbin R (2004) Improved techniques for the identification of pseudogenes. Bioinformatics 20 Suppl 1:i94–100.

[Collett 2003] Collett D (2003) Modelling Survival Data in Medical Research. CRC Press.

[Conant and Wagner 2004] Conant GC, Wagner A (2004) Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans. Proc Biol Sci 271:89–96.

[Darwin 1859] Darwin CR (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray.

[Dayhoff et al. 1978] Dayhoff M, Schwartz R, Orcutt B (1978) A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure 5:345–352.

[Dehal and Boore 2005] Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3:e314.

[Demuth et al. 2006] Demuth JP, Bie TD, Stajich JE, Cristianini N, Hahn MW (2006) The evolution of Mammalian gene families. PLoS ONE 1:e85.

[Edgar 2004] Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.

[Endo et al. 1996] Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. Mol Biol Evol 13:685–690.

[Enright et al. 2002] Enright AJ, Dongen SV, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584.

[Enright et al. 2003] Enright AJ, Kunin V, Ouzounis CA (2003) Protein families and TRIBES in genome sequence space. Nucleic Acids Res 31:4632–4638.

[Eyre-Walker 1996] Eyre-Walker A (1996) Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy? Mol Biol Evol 13:864–872.

[Eyre-Walker 2006] Eyre-Walker A (2006) The genomic rate of adaptive evolution. Trends Ecol Evol 21:569–575.

[Fares et al. 2002] Fares MA, Elena SF, Ortiz J, Moya A, Barrio E (2002) A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. J Mol Evol 55:509–521.

[Fay and Wu 2000] Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413.

[Force et al. 1999] Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.

[Gaucher et al. 2003] Gaucher EA, Miyamoto MM, Benner SA (2003) Evolutionary, structural and biochemical evidence for a new interaction site of the leptin obesity protein. Genetics 163:1549–1553.

[Golding and Dean 1998] Golding GB, Dean AM (1998) The structural basis of molecular adaptation. Mol Biol Evol 15:355–369.

[Goldman and Yang 1994] Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736.

[Greene et al. 2007] Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35:D291–D297.

[Gu et al. 2003] Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421:63–66.

[Harris 1966] Harris H (1966) Enzyme polymorphisms in man. Proc R Soc Lond B Biol Sci 164:298–310.

[Harrison and Gerstein 2002] Harrison PM, Gerstein M (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J Mol Biol 318:1155–1174.

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

[He and Zhang 2006] He X, Zhang J (2006) Higher duplicability of less important genes in yeast genomes. Mol Biol Evol 23:144–151.

[Hughes 1994] Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc Biol Sci 256:119–124.

[Hughes 1996] Hughes AL (1996) Adaptive evolution of genes and genomes. Oxford University Press.

[Hughes and Hughes 1993] Hughes MK, Hughes AL (1993) Evolution of duplicate genes in a tetraploid animal, Xenopus laevis. Mol Biol Evol 10:1360–1369.

[Hughes and Liberles 2007] Hughes T, Liberles D (2007) The Pattern of Evolution of Smaller-Scale Gene Duplicates in Mammalian Genomes is More Consistent with Neo- than Subfunctionalisation. J Mol Evol 65:574–588.

[Huynen and van Nimwegen 1998] Huynen MA, van Nimwegen E (1998) The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 15:583–589.

[Ihmels et al. 2007] Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS (2007) Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. Mol Syst Biol 3:86.

[Ina 1995] Ina Y (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J Mol Evol 40:190–226.

[Jaillon et al. 2004] Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biémont C, Skalli Z, Cattolico L, Poulain J, Berardinis VD, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigó R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quétier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Crollius HR (2004) Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431:946–957.

[Johnson et al. 2001] Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE (2001) Positive selection of a gene family during the emergence of humans and African apes. Nature 413:514–519.

[Kamal et al. 2005] Kamal M, Luscombe NM, Qian J, Gerstein M (2005) Analytical evolutionary model for protein fold occurrence in genomes, accounting for the effects of gene duplication, deletion, acquisition and selective pressure. Springer Science and Business Media.

[Karev et al. 2002] Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV (2002) Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evol Biol 2:18.

[Kellis et al. 2004] Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428:617–624.

[Kikuta et al. 2007] Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, Ghislain J, Pezeron G, Mourrain P, Ellingsen S, Oates AC, Thisse C, Thisse B, Foucher I, Adolf B, Geling A, Lenhard B, Becker TS (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res 17:545–555.

[Kimura 1983] Kimura (1983) The neutral theory of molecular evolution. Cambridge University Press.

[Kimura 1964] Kimura M (1964) Diffusion Models in Population Genetics. Journal of Applied Probability 1:177–232.

[Kimura 1968] Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626.

[Kuepfer et al. 2005] Kuepfer L, Sauer U, Blank LM (2005) Metabolic functions of duplicate genes in Saccharomyces cerevisiae. Genome Res 15:1421–1430.

[Leggatt and Iwama 2003] Leggatt R, Iwama G (2003) Occurrence of polyploidy in the fishes. Reviews in Fish Biology and Fisheries 13:237–246.

[Lewontin and Hubby 1966] Lewontin RC, Hubby JL (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics 54:595–609.

[Li 1993] Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99.

[Luscombe et al. 2002] Luscombe NM, Qian J, Zhang Z, Johnson T, Gerstein M (2002) The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome Biol 3.

[Lynch and Conery 2000] Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.

[Lynch and Conery 2003] Lynch M, Conery JS (2003) The evolutionary demography of duplicate genes. J Struct Funct Genomics 3:35–44.

[Lynch and Force 2000] Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473.

[Maere et al. 2005] Maere S, Bodt SD, Raes J, Casneuf T, Montagu MV, Kuiper M, de Peer YV (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A 102:5454–5459.

[Maston and Ruvolo 2002] Maston GA, Ruvolo M (2002) Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. Mol Biol Evol 19:320–335.

[McCarrey 1990] McCarrey JR (1990) Molecular evolution of the human Pgk-2 retroposon. Nucleic Acids Res 18:949–955.

[McDonald and Kreitman 1991] McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.

[Mighell et al. 2000] Mighell AJ, Smith NR, Robinson PA, Markham AF (2000) Vertebrate pseudogenes. FEBS Lett 468:109–114.

[Moore and Purugganan 2003] Moore RC, Purugganan MD (2003) The early stages of duplicate gene evolution. Proc Natl Acad Sci U S A 100:15682–15687.

[Moore and Purugganan 2005] Moore RC, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol 8:122–128.

[Mounsey et al. 2002] Mounsey A, Bauer P, Hope IA (2002) Evidence suggesting that a fifth of annotated Caenorhabditis elegans genes may be pseudogenes. Genome Res 12:770–775.

[Nei and Gojobori 1986] Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426.

[Nelson and Cox 2000] Nelson DL, Cox MM (2000) Lehninger Principles of Biochemistry. third edition,, Worth Publishers.

[Notredame et al. 2000] Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205–217.

[Ohno 1970] Ohno S (1970) Evolution by gene duplication. New York: Springer-Verlag.

[Ohno et al. 1968] Ohno S, Wolf U, Atkin NB (1968) Evolution from fish to mammals by gene duplication. Hereditas 59:169–187.

[Ohta 2002] Ohta T (2002) Near-neutrality in evolution of genes and gene regulation. Proc Natl Acad Sci U S A 99:16134–16137.

[Page and Holmes 1996] Page RDM, Holmes EC (1996) Molecular Evolution - A Phylogenetic approach. second edition,, Blackwell Science.

[Pamilo and Bianchi 1993] Pamilo P, Bianchi NO (1993) Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol 10:271–281.

[Papp et al. 2003] Papp B, Pál C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197.

[Ptacek et al. 1994] Ptacek M, Gerhardt H, Sage R (1994) Speciation by Polyploidy in Treefrogs: Multiple Origins of the Tetraploid, Hyla versicolor. Evolution 48:898–908.

[Qian et al. 2001] Qian J, Luscombe NM, Gerstein M (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J Mol Biol 313:673–681.

[Rastogi and Liberles 2005] Rastogi S, Liberles DA (2005) Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol 5:28.

[Rodríguez-Trelles et al. 2003] Rodríguez-Trelles F, Tarrío R, Ayala FJ (2003) Convergent neofunctionalization by positive Darwinian selection after ancient recurrent duplications of the xanthine dehydrogenase gene. Proc Natl Acad Sci U S A 100:13413–13417.

[Ronquist and Huelsenbeck 2003] Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

[Roth et al. 2005] Roth C, Betts MJ, Steffansson P, Saelensminde G, Liberles DA (2005) The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics. Nucleic Acids Res 33:D495–D497.

[Roth and Liberles 2006] Roth C, Liberles DA (2006) A systematic search for positive selection in higher plants (Embryophytes). BMC Plant Biol 6:12.

[Schacherer et al. 2004] Schacherer J, Tourrette Y, Souciet JL, Potier S, Montigny JD (2004) Recovery of a function involving gene duplication by retroposition in Saccharomyces cerevisiae. Genome Res 14:1291–1297.

[Schrödinger 1944] Schrödinger E (1944) What is life? Cambridge University Press.

[Seoighe and Gehring 2004] Seoighe C, Gehring C (2004) Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. Trends Genet 20:461–464.

[Shiu et al. 2006] Shiu SH, Byrnes JK, Pan R, Zhang P, Li WH (2006) Role of positive selection in the retention of duplicate genes in mammalian genomes. Proc Natl Acad Sci U S A 103:2232–2236.

[Siltberg and Liberles 2002] Siltberg J, Liberles D (2002) A simple covarion-based approach to analyse nucleotide substitution rates. Journal of Evolutionary Biology 15:588–594.

[Smith and Eyre-Walker 2002] Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415:1022–1024.

[Sober 1993] Sober E, ed. (1993) Conceptual Issues in Evolutionary Biology. second edition,, The MIT Press.

[Stoltzfus 1999] Stoltzfus A (1999) On the possibility of constructive neutral evolution. J Mol Evol 49:169–181.

[Tajima 1989] Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

[Thompson et al. 1994] Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.

[Torrents et al. 2003] Torrents D, Suyama M, Zdobnov E, Bork P (2003) A genome-wide survey of human pseudogenes. Genome Res 13:2559–2567.

[Vallender and Lahn 2004] Vallender EJ, Lahn BT (2004) Positive selection on the human genome. Hum Mol Genet 13 Spec No 2:R245–R254.

[van de Peer et al. 2003] van de Peer Y, Taylor JS, Meyer A (2003) Are all fishes ancient polyploids? J Struct Funct Genomics 3:65–73.

[Wagner 2005] Wagner A (2005) Energy constraints on the evolution of gene expression. Mol Biol Evol 22:1365–1374.

[Walker et al. 1995] Walker EL, Robbins TP, Bureau TE, Kermicle J, Dellaporta SL (1995) Transposon-mediated chromosomal rearrangements and gene duplications in the formation of the maize R-r complex. EMBO J 14:2350–2363.

[Watson and Crick 1953] Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171:737–738.

[Welch 2006] Welch JJ (2006) Estimating the genomewide rate of adaptive protein evolution in Drosophila. Genetics 173:821–837.

[Woods et al. 2005] Woods IG, Wilson C, Friedlander B, Chang P, Reyes DK, Nix R, Kelly PD, Chu F, Postlethwait JH, Talbot WS (2005) The zebrafish gene map defines ancestral vertebrate chromosomes. Genome Res 15:1307–1314.

[Wu 2005] Wu Q (2005) Comparative genomics and diversifying selection of the clustered vertebrate protocadherin genes. Genetics 169:2179–2188.

[Yanai et al. 2000] Yanai I, Camacho CJ, DeLisi C (2000) Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Phys Rev Lett 85:2641–2644.

[Yang 1997] Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556.

[Yang 1998] Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573.

[Yang 2006] Yang Z (2006) Computational Molecular Evolution. Oxford University Press.

[Yang and Nielsen 2002] Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908–917.

[Zhang et al. 2005] Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472–2479.

[Zmasek and Eddy 2001] Zmasek CM, Eddy SR (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 17:821–828.

A. Paper I

Journal of Molecular Evolutio 65, Hughes, Timothy and David A. Liberles, The Pattern of Evolution of Smaller-Scale Gene Duplicates in Mammalian Genomes is More Consistent with Neo- than Subfunctionalisation, pp. 574–588. Copyright 2007 Springer Science + Business Media. http://dx.doi.org/10.1007/s00239-007-9041-9

The Pattern of Evolution of Smaller-Scale Gene Duplicates in Mammalian Genomes is More Consistent with Neo- than Subfunctionalisation

Abstract

Gene duplication and the accompanying release of negative selective pressure on the duplicate pair is thought to be the key process that makes functional change in the coding and regulatory regions of genomes possible. However, the nature of these changes remains unresolved. There are a number of models for the fate of gene duplicates, the two most prominent of which are neofunctionalisation and subfunctionalisation, but it is still unclear which is the dominant fate. Using a dataset consisting of smaller-scale (tandem and segmental) duplications identified from the genomes of four fully sequenced mammalian genomes, we characterise two key features of smaller-scale duplicate evolution: the rate of pseudogenisation and the rate of accumulation of replacement substitutions in the coding sequence. We show that the best fitting model for gene duplicate survival is a Weibull function with a downward sloping convex hazard function which implies that the rate of pseudogenisation of a gene declines rapidly with time since duplication. Our analysis of the accumulation of replacement substitutions per replacement site shows that they accumulate on average at 64% of the neutral expectation immediately following duplication and as high as 73% in the human lineage. Although this rate declines with time since duplication, it takes several tens of millions of years before it has declined to half its initial value. We show that the properties of the gene death rate and of the accumulation of replacement substitutions are more consistent with neofunctionalisation (or subfunctionalisation followed by neofunctionalisation) than they are with subfunctionalisation alone or any of the other alternative modes of evolution of smaller-scale duplicates.

Keywords

Gene duplication - Chordata - Pseudogenisation - Neofunctionalisation - Subfunctionalisation - Positive selection - Nonsynonymous substitution

Supplementary materials to research article: The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than sub-functionalisation

Timothy Hughes and David A. Liberles

TH: Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway.
Telephone: (+47) 55 58 40 72. Email: tim@bccs.uib.no

DAL: Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA.
Telephone: (+1) 307 766 5206. Email: liberles@uwyo.edu.

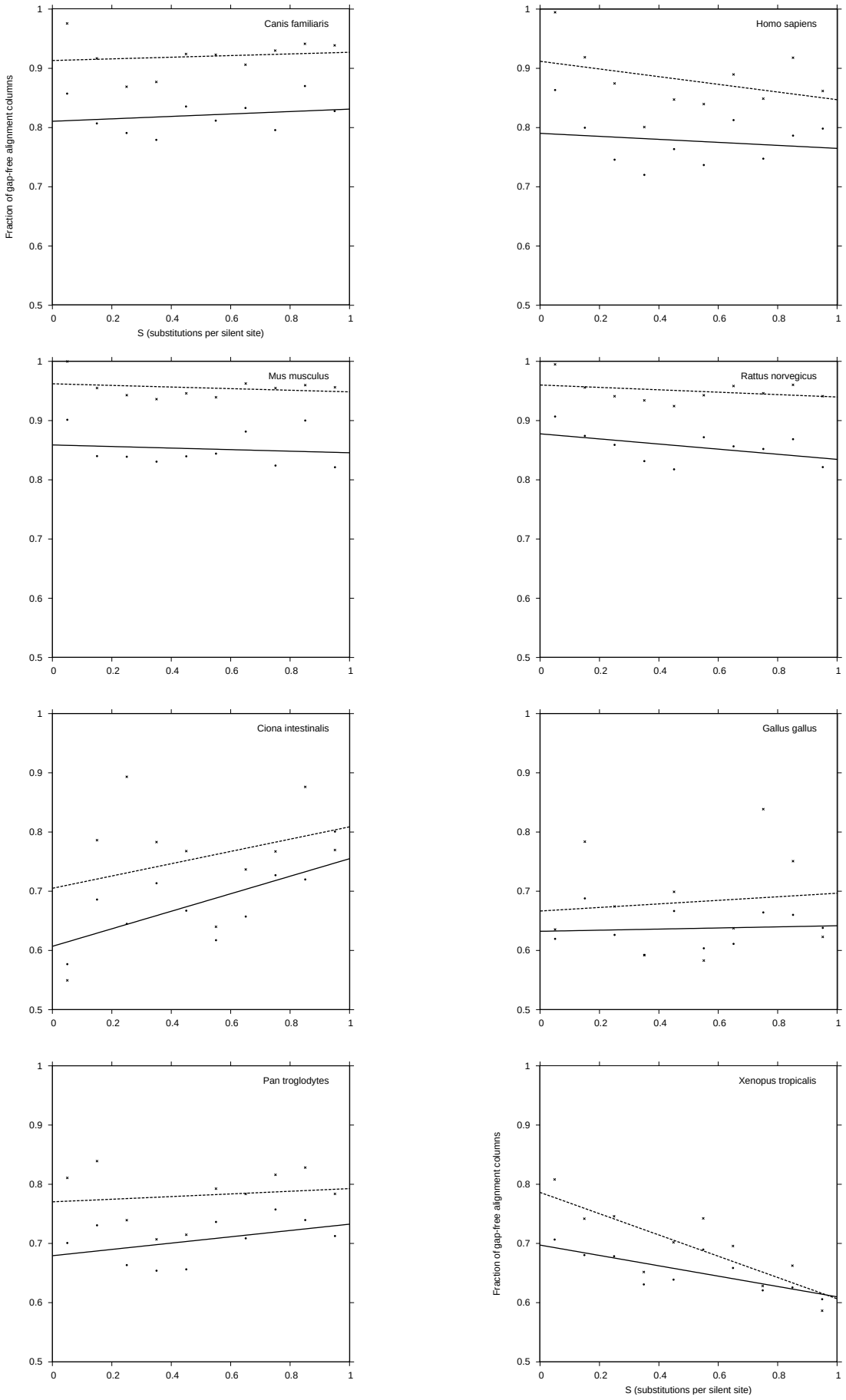

Figure 1: Alignment quality control (fraction of gap-free alignment columns)

Duplicate pairs are placed in groups of size 0.1 S. *Crosses:* group median. *Points:* group mean. *Dotted line:* linear equation fitted to median data. *Full line:* linear equation fitted to mean data.

Species	Duplications per 0.01S	Genes per genome	Duplications per gene per S
C. familiaris	147	18201	0.808
H. sapiens	460	22218	2.070
M. musculus	805	24460	3.291
R. norvegicus	246	21952	1.121

Table 1: Gene duplication rate estimates

Species	Subst. per silent site per BY
C. familiaris	2.94
H. sapiens	2.20
M. musculus	6.07
R. norvegicus	6.07

Table 2: Silent substitution rate estimates

No estimate was available for *C. familiaris*, so the artiodactyl rate was used as this is the nearest lineage for which an estimate was available (Yang and Nielsen 1998; Dimcheff et al. 2002; Springer et al. 2003; Axelsson et al. 2005).

The R code used for fitting the two models to the data and datasets for *C. familiaris* are available online at http://digitised.info. File paths (to model and dataset files) set in the main R script files will need to be adjusted to the correct values. Plots of the residuals are also available for the survival modeling. The file names and a description of their contents are as follows:

supMat/replacementSubstModeling:

bestAltSplices.tab (dataset)

model.repl.low.hughes (model)

outputFunctions.r (functions called from the main R script)

replSubsFitting.r (main R script)

supMat/survivalModeling:

model.survival (model)

outputFunctions.r (functions called from the main R script)

silentSubstCounts_bucketSize_0.01_median.tab (dataset)

survivalFitting.r (main R script)

plot_residuals_unrestricted_genusSpecies.pdf (plot of standardised residuals against fitted values of unrestricted model for model specification verification)

A description of the columns of the bestAltSplices.tab file (each row contains the data for one duplicate pair):

1 pair ID

2 number of codons in the alignment

3 number of gap free columns in the alignment

4 maximum likelihood with ω estimated

5 maximum likelihood with $\omega = 1$

6 replacement substitutions per replacement site (R) under model where ω estimated

7 silent substitutions per silent site (S) under model where ω estimated

8 replacement sites under model where ω estimated

9 silent sites under model where ω estimated

10 replacement substitutions per replacement site (R) under model where $\omega = 1$

11 silent substitutions per silent site (S) under model where $\omega = 1$

12 replacement sites under model where $\omega = 1$

13 silent sites under model where $\omega = 1$

14 Ensembl ID of first protein in pair

15 Ensembl ID of second protein in pair

A description of the columns of the silentSubstCounts_bucketSize_0.01_median.tab file (each row contains the summary data for one group of duplicate pairs where a group consists of all duplicate pairs with S within the interval of size 0.01):

1 median S value for the group

2 number of duplicate pairs in the group

B. Paper II

The power-law distribution of gene family size is driven by the pseudogenisation rate's heterogeneity between gene families

Timothy Hughes [a] David A. Liberles [b,*]

[a] *Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway*

[b] *Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA*

Abstract

Genome sequencing has shown that the number of homologous gene families of a given size declines rapidly with family size. A power-law has been shown to provide the best mathematical description of this relationship. However, it remains unclear what evolutionary forces drive this observation. We use models of gene duplication, pseudogenisation and accumulation of replacement substitutions, which have been validated and parameterised using genomic data, to build a model of homologous gene evolution. We use this model to simulate the evolution of the distribution of gene family size and show that the power-law distribution is driven by the pseudogenisation rate's heterogeneity across gene families and its correlation within families. Moreover, we show that gene duplication and pseudogenisation are necessary and sufficient for the emergence of the power-law.

Key words:
duplication, loss, homology, family, power-law

1 Introduction

Surveys show that, in the genome of every organism sequenced so far, the number of homologous gene families of a given size declines rapidly as family size increases, i.e. there are always many small homologous gene families and a few very large families. Different functions could potentially describe this

* Corresponding author.
Email address: liberles@uwyo.edu (David A. Liberles).

relationship between family size and number of families (exponential, yule, log-normal), but the power-law with a negative exponent has been found to provide the best fit to empirical data (Huynen and van Nimwegen, 1998; Luscombe et al., 2002). The power-law distribution of homologous gene family size has been reported for single genomes from archea, bacteria and eukaryota (Huynen and van Nimwegen, 1998; Yanai et al., 2000; Harrison and Gerstein, 2002) and across multiple genomes (Enright et al., 2003). In mathematical terms: $N = aF^b$ where F is the family size and N is the number of families of this size or, taking the natural logarithm, $\ln(N) = \ln(a) + b.\ln(F)$ i.e. a linear relationship on a log-log plot. The exponent b of the distribution is usually in the range -4.0 to -2.75 with a weak positive correlation (correlation coefficient of 0.63) between the exponent and the logarithm of number of genes in the genome (Huynen and van Nimwegen, 1998). This positive correlation implies a relative increase of the number of large clusters over the number of small clusters as the number of genes in the genome increases.

The power-law distribution is also observed for biological "parts" at different levels of organisation (Interpro families, protein superfamilies and folds, pseudogene families and pseudomotifs) and for many different attributes associated with these parts (their functions, interactions and expression levels) (Luscombe et al., 2002). Moreover, such distributions are not limited to biology and are often observed for the numbers of parts (or properties of the parts) of man-made systems that are not the product of design, for example, the relative sizes of cities (Zipf, 1949) and the connectivity of computers in the world wide web (Barabasi and Albert, 1999; Albert et al., 2000).

Several papers have shown that it is possible to define theoretical evolutionary models that replicate the power-law distributions of genes or protein folds (Huynen and van Nimwegen, 1998; Yanai et al., 2000; Qian et al., 2001; Karev et al., 2002; Kamal et al., 2005). These models are based on the process of gene duplication and add additional evolutionary processes such as gene death, point mutations or lateral gene transfer to achieve a power-law distribution. The shortcoming of these models is their theoretical nature i.e. none of the models build on empirically verified characterisations of the underlying processes. This results in the formulation of very different models (some of which are even biologically unrealistic), but all are found to be consistent with the power-law. There is thus no consensus on which forces drive the power-law distribution. Moreover, these models also frequently only include a subset of the relevant processes e.g. ignoring gene death or sequence divergence, but aim nevertheless to show what is minimally necessary for the power-law to emerge.

In this paper, we use models of gene duplication, pseudogenisation and accumulation of replacement substitutions, which have been validated and parameterised using genomic data, to build a model of homologous gene evolution.

These models were developed in previous work (Hughes and Liberles, 2007) which itself builds on the seminal work of Lynch and Conery (Lynch and Conery, 2000). We then use this model to simulate the evolution of the distribution of gene family size. We use the data output by the simulations to test whether our model of homologous gene evolution can qualitatively replicate a power-law distribution of gene family size. We find that, for our model to produce a power-law distribution, it must incorporate heterogeneity of pseudogenisation rates across families and correlation within families. Further, using the version of our model that can replicate the power-law distribution, we test whether all three of these forces (gene duplication, accumulation of replacement substitutions, and pseudogenisation) are essential to the emergence of the power-law. We find that the quantitative accumulation of replacement substitutions which drives sequence divergence can be considered superfluous, although qualitatively it is essential as substitutions causing loss of function or stop codons are major drivers of pseudogenisation which is necessary for the emergence of a power-law distribution.

2 Methods

2.1 Model overview

The power-law distribution is observed at many levels of biological organisation. We study its emergence at the gene family level because this is the lowest level of functional genomic organisation at which the power-law has been observed. Power-laws at higher levels such as superfamily or fold, are most probably the result of the power-law at this lower level as super-families and folds represent clusters of homologous genes (descended from a common ancestor gene), in the same way as gene families do, only more distant. The same processes (gene duplication and divergence), that more recently have generated gene families, generated clusters of homologous superfamilies and folds.

The most fundamental evolutionary forces driving the evolution of coding sequences in eukaryotes are: gene duplication, substitutions (silent and replacement), and indel events. Substitutions and indel events then drive both sequence divergence and pseudogenisation: stop codon substitutions, replacement substitutions, indel events, and substitutions in regulatory regions can all potentially produce a non-functional gene, which unless under selective pressure to be retained, will eventually pseudogenise. We could model how substitutions and indels drive qualitative sequence divergence and pseudogenisation, but this is difficult and would result in a very complex model. Instead, we model the duplication rate, the rate of accumulation of replace-

ment substitutions (quantitative sequence divergence) and the duplicate gene death rate. The gene death rate effectively incorporates the effect on pseudogenisation of replacement substitutions in coding regions (including stop codons), substitutions in regulatory regions and indels, but without explicitly modeling the process by which pseudogenisation occurs. This is a combination of the probability of such mutations as well as the probability of fixation mediated by population genetic forces plus selection.

These processes lead to the dynamics of gene family size. Duplication increases the number of genes in a family by one; replacement substitutions, if they accumulate in sufficient numbers, lead to a sequence breaking away from its family, thus reducing the number of genes in the family by one and creating a new family of size one; pseudogenisation reduces family size by one without creating a new family. Lateral gene transfer (which is an important process in bacteria) or ab-initio gene creation have the potential to play a similar role to duplication by producing new genes, but they are widely recognised to be extremely rare or non-existent in eukaryotes (Salzberg et al., 2001). Since the power-law has been observed for all sequenced higher eukaryotes, lateral gene transfer and ab-initio gene creation cannot be key to the existence of a power-law distribution and, therefore, we do not incorporate these processes in our model.

2.2 Key processes

The models of duplication, accumulation of replacement substitutions and pseudogenisation are taken directly from a previous study (Hughes and Liberles, 2007). The reader is referred to that article for the full details of the models, including the justification of the functional form of the equations as well as the parameter estimates presented in this section. These models describe the rate of gene duplication, the rate at which replacement substitutions accumulate between genes in a duplicate pair and the rate at which one of the genes in the pair pseudogenises. Time is measured through the accumulation of silent substitutions per silent site (S) between duplicate genes. S can be converted to real time by dividing by 2 and by the number of silent substitutions per billion years. Thus, given a rate of 2.20 silent substitutions per silent site per billion years for *H. sapiens* (Yang and Nielsen, 1998), $1\,S$ corresponds to 0.2 billion years - all parametrisations are from the *H. sapiens* estimates. In *H. sapiens*, we estimated that genes duplicate at a rate of 2.07 per gene per S (Hughes and Liberles, 2007). A duplicate pair i accumulates replacement substitutions per replacement site, R, according to the equation:

$$R_i = \theta_1 S_i + (\theta_2/\theta_3)(1 - \exp(-\theta_3 S_i)) \quad (1)$$

We use the following values of the parameters (Hughes and Liberles, 2007): $\theta_1 = 0.13$; $\theta_2 = 0.70$; $\theta_3 = 2.4$. Note that: $dR/dS = \theta_1 + \theta_2$ at $S = 0$ (which corresponds to the rate of accumulation of R immediately following duplication) and $dR/dS \rightarrow \theta_1$ as $S \rightarrow \infty$ for $\theta_3 > 0$ (which is the asymptotic rate of accumulation of R for a duplicate pair which gives an estimate of the rate of replacement substitution accumulation for sequences under negative selection).

The probability of pseudogenisation of one of the genes in a pair within Δt given that both genes are still functional at t is:

$$Pr(t < T < t + \Delta t / T > t) = -\frac{Q(t + \Delta t) - Q(t)}{Q(t)} \tag{2}$$

where $Q(t) = Pr(T > t)$ is the survival function: the probability that the time of death, T, is greater than t, i.e. the probality that both genes are still functional at time t. The hazard function $\lambda(t)$ is defined as the event (death/pseudogenisation) rate at time t conditional on survival to time t or later:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{Pr(t < T < t + \Delta t / T > t)}{\Delta t} = -Q'(t)/Q(t) \tag{3}$$

We have shown that the Weibull survival function $Q(t) = e^{\rho_1 t^{\rho_2}}$ provides an excellent fit to the data (Hughes and Liberles, 2007). Thus, we use this model of the survival function and S as a proxy for time:

$$\lambda(S) = -\rho_1 \rho_2 S^{\rho_2 - 1} \tag{4}$$

In *H. sapiens*, the fitted parameters are $\rho_1 = -4.1$ and $\rho_2 = 0.33$ which implies that the rate of pseudogenisation of a duplicate is a decreasing function of S (proxy for time since duplication) and not a time-independent constant (Hughes and Liberles, 2007).

2.3 Genes and gene duplicate pairs

A gene in our model has two key characteristics: it is either functional or pseudogenised, and it has a measure of the number of silent and replacement substitutions per site between itself and all homologous genes i.e. all genes that can be traced to a common ancestor through a series of duplication events. The model is initialised with a set of singleton genes i.e. genes that have no duplicates, and therefore each forms a family of size one. These are the "founding" genes of the homologous gene families. Because all key processes

are defined in terms of S, we define a "clock" which "ticks" in increments of 0.005 units of S. At each tick of the clock each gene's number of silent substitutions is incremented by half a tick, so that the distance between all genes increases by one tick. For each gene, we then detect the closest non-pseudogenised homolog which we define as the homologous gene with lowest R distance to the gene of interest. The S distance between the two genes is a measure of the time since the original duplication event and is used to compute the number of replacement substitutions per site the duplicate pair should be subject to in the timeframe of the current tick (equation 1) and the probability that one of the duplicates pseudogenises during the current tick (equation 2 with S as a proxy for time and the Weibull survival function). A gene that has no homologs (such as a founding singleton before it is duplicated) is assigned an S value of 1,000 which ensures that it accumulates R at a very low rate and is subject to a very low probability of pseudogenisation. This is a reasonable way to model singletons as singletons can be expected to have evolved some kind of specialised function that is under selective pressure to be retained in the genome. Finally, each gene is subject to a constant probability of duplication during each tick. The gene that results from a duplication is added to the set of homologous genes. It inherits the R and S distances to other genes from its parent and has a distance of $0\,R$ and $0\,S$ to its parent. A new gene is also assigned a replacement substitution error term (ε_i) which is added to equation 1, so as to incorporate into the model differences between genes in the rate of accumulation of replacement substitutions. Because the functional form of the distribution of this error term is not known, we draw an error term randomly from all residuals from the fitting of equation 1 to the *H. sapiens* gene duplicate data (Hughes and Liberles, 2007). This error term can be standardised through the fitted model of the variance as a function of S (Hughes and Liberles, 2007):

$$Var(\varepsilon_i) = \sigma^2(\tau_1 S_i + \exp(\tau_2(1 - \exp(-\tau_3 S_i)))) \tag{5}$$

with the following parameter estimates (Hughes and Liberles, 2007): $\sigma^2 = 3.55e-5$; $\tau_1 = 229.4$; $\tau_2 = 6.32$; $\tau_3 = 4.14$. Figure 1 illustrates graphically a hypothetical simulation scenario and the simulation code is available in the supplementary materials.

2.4 Clustering genes into gene families

At regular intervals during the simulation (every 0.02 S for $0 < S < 0.3$, every 0.2 for $0.3 < S < 3.0$, and every 1.0 for $S > 3.0$), we extract S and R for all duplicate pairs and use this data to calculate an age distribution of duplicate pairs and a plot of the accumulation of replacement substitution with time. This enables us to verify that our simulated genes are evolving in the same

way as the real *H. sapiens* duplicate gene data. We also compute a complete linkage clustering of all non-pseudogenised genes, which ensures a maximum R distance between genes in a group (Gan et al., 2007). We use complete linkage clustering (where all members in a family meet a distance threshold to all other members) and not single linkage clustering (where linkage of a gene to a single member of the cluster is sufficient to include the gene as part of the family) because it is intrinsic to single linkage clustering that the probability that a sequence is added to a cluster grows with the size of the cluster, thus causing this method to contain a bias towards larger clusters. Since we aim to test what processes are essential to the emergence of the power-law, i.e. a distribution that contains more large families than the alternative exponential function, we cannot use a method that contains a bias towards larger families. For each family size distribution, we fit the power-law and exponential functions to the data by ordinary least squares using the equations: $\ln(N) = \ln(a) + b.\ln(F)$ and $\ln(N) = \ln(a) + b.F$.

Computation time for the simulation increases rapidly with the number of initial singletons, therefore we start the simulations with 2,000 singletons. There is no reason to believe that the number of initialising singletons has any bearing on our results. The settings of each simulation and its output are described in table 1.

2.5 Empirical data

In order to determine a realistic maximum R distance between genes in the same family that can be used as the cutoff in the complete linkage clustering and also to provide an empirical distribution of gene family size for *H. sapiens* to which our simulated data can be compared, we build a clustering of protein coding genes.

Our basic dataset consists of all protein coding transcript sequences from the annotated genome sequence of *H. sapiens* from release 31 of Ensembl (Birney et al., 2006). First, we carry out low complexity masking of the sequences using CAST (Promponas et al., 2000) and then perform an all-against-all BLAST (Altschul et al., 1997) of the translated sequences (substitution matrix=BLOSUM62, gap opening cost=11, gap extension cost=1). In order to make the output of the all-against-all Blast manageable, the BLAST sequence pairs (query and target sequences) are filtered to remove any targets that do not satisfy all of the following criteria that should be satisfied by even very distant homologs: 20% similarity to the query, 60% coverage of the query, e-value $< 10^{-5}$. The e-values of the retained sequences are then used as input to the MCL clustering algorithm with the inflation parameter set to 4.0 (Enright et al., 2002). In order to ensure that the clusters do not contain very

distant homologs which are difficult to align, we further split each of these clusters using a complete linkage algorithm with a maximum length ratio criteria (longest/shortest < 2.5). We then align the sequences using MUSCLE with the default settings (Edgar, 2004) and perform complete linkage using a percentage gap difference criteria (percentage gap difference excluding the ends $< 40\%$). Each of the resulting families potentially contains alternative splices of the same gene, so we retain only the alternative splice form that has the best alignment to the other genes in the cluster. This method of homologous gene family construction was designed to maximise the probability of computing high quality alignments which can then be used to produce reliable measures of S and R.

From this gene family dataset, we compute the family size distribution. The power-law function fitted to this data has a slope of -2.53 (see figure 3 for a plot of this data). We also draw a random sample of 1,000 families from the dataset and use a modified Nei-Gojobori method (defined on pages 57 to 59 of (Nei and Kumar, 2000)) on the untranslated sequences to compute the R distance between all pairs of sequences in the family. For each family, we record the maximum R distance and then we compute the median of these maxima which produces a value of 0.56 R which we use as our clustering cutoff.

3 Results

3.1 Testing for the emergence of the power-law with the basic model

We first run the model exactly as described in the material and methods (model 1) until $S = 10.0$, which corresponds to approximately 2 billion years. Interestingly, the genome size increases during this long time frame, but only moderately. If we had an incorrect characterisation of the duplication rate or the pseudogenisation rate, the simulation may have generated no expansion at all or massive expansion. The fact that this did not occur is corroborative evidence that we are using reasonable characterisations of the duplication and pseudogenisation processes. The exponential function clearly provides the best fit to the size distribution of gene family size, although there are two points where there is a dramatic drop in the quality of the fit of the exponential function as measured by R^2 (an R^2 value of 1 indicates perfect fit of the equation to the data). These drops are due to the ephemeral emergence of larger gene families. The transitory existence of these larger families is clear from the plot of R^2 in the first row of figure 3 and from the simulation animations (see supplementary materials). The animation also contains figures of the age distribution of duplicate pairs and of the accumulation of replacement substitution between the genes in a pair. These provide a verification that the

simulation is evolving the genes in a way that replicates the real gene duplicate data (Hughes and Liberles, 2007).

It is quite possible that either the length of time the simulation is run for or the initial number of singletons are the reason a power-law distribution fails to emerge, so we ran two additional simulations: one where we extend the run time to $S = 30.0$ (model 2) and one where the initial number of singletons is $10,000$ (model 3). For models with a large number of genes such as model 3, complete linkage clustering is replaced with a faster approximate clustering: a gene belongs to a family of size l, if it has $l-1$ genes from which it is separated by $0.56\ R$, and the number of families of size l is the number of genes belonging to a family of size l divided by l. Both model 2 and 3 result in a failure of the power-law to emerge (see figure 1 in the supplementary materials).

3.2 Introducing variation in the rate of pseudogenisation across families

One assumption of our model is that all genes have the same hazard function. This is obviously an oversimplification, as different genes have different probabilities of retention under most prominent models of duplicate retention, e.g. the near-neutral subfunctionalisation model (Force et al., 1999) and the selective neofunctionalisation model (Ohno, 1970). We remove this assumption from the model by defining the probability of pseudogenisation of a duplicate pair i within a time interval Δt given survival until t as:

$$Pr'(t < T < t + \Delta t / T > t) = (1 + \upsilon_i) Pr(t < T < t + \Delta t / T > t) \qquad (6)$$

where υ_i is drawn from a normal distribution with mean and standard deviation as specified in table 1. This specification of the probability of pseudogenisation within Δt ensures that the expected hazard function for all genes is the same as that used in first run of the simulation and that the probability in equation 6 is proportional to the probability in equation 2. When a new gene is created by duplication, the error term υ_i can either be inherited from the gene that duplicated, in which case all genes descendent from a founding singleton will have the same hazard function; or a new error can be drawn from the distribution in which case there will be no correlation between the hazard functions of genes descendent from the same singleton. In the first simulation model within this group (model 4), we set the mean of the error distribution to 0 and the standard deviation to 0.2 and allow the error to be inherited by duplicates (table 1). The results of this simulation show that the fit of the exponential function to the distribution of gene family size begins to fall below the fit of the power-law function for values of $S > 2.0$ while the power-law function retains a good fit (see figure 3). The number of genes at $S = 2.0$ is 2,759 and by $S = 5.0$ this number has reached 3,853 which is at

the limit of where we are able to compute a complete linkage clustering of gene families, thus explaining why we only plot on the interval $0 < S < 5.0$. Interestingly, at $S = 2.0$, when the power-law has clearly emerged, the exponent of the power-law function is -3.3. This cannot be said to be close to the value of -2.53 estimated from the empirical *H. sapiens* data, but both values do fall within the upper part of the range for exponents of power-law distributions of gene family size, as expected for a genome of *H. sapiens*'s large size (Huynen and van Nimwegen, 1998). That they do not match more closely does not represent a failure of our model as it is not our aim to reproduce the *H. sapiens* distribution, we simply aim to test what features of an empirically validated model of homologous gene evolution are necessary to produce a power-law. The difference between the exponents can be explained by the fact that the empirical *H. sapiens* distribution has been generated over several hundreds of millions of years, during which time the rates of duplication, loss and divergence were not the same as the rates that prevail in the present.

This result suggests that heterogeneity of the hazard rate across families and correlation within families are important for the power-law distribution to emerge. To test whether this is the case, we run simulations with three variants of this basic model.

First, we remove inheritance of the error between duplicates to test whether correlation of the error within homologous genes is essential (model 5). Second, we modify the basic model by setting a positive mean for the error distribution, but we retain error inheritance and the same level of hazard heterogeneity (model 6). This raises the average hazard level applied in previous simulations and enables a test of the extent to which the power-law is dependent on the magnitude of the probability of pseudogenisation (we are careful to ensure that level of the hazard is not set so high that the genome does not expand). Third, we retain correlation within families (inheritance of the error) and heterogeneity between families, but we reduce the level of heterogeneity by reducing the standard deviation of the error distribution (model 7). In none of these models do we observe a clear emergence of the power-law as we did in model 4 (see rows 2, 3 & 4 of figure 2 in the supplementary materials). The fit of the exponential function is sometimes weaker than the fit of the power-law function particularly at higher values of S, but we do not observe a clear fall in the fit of the exponential function as we do in model 4. Model 7 is perhaps the model that gets closest to producing a power-law distribution: from $S = 2.0$ onwards, the exponential function provides a worse fit than the power-law and at specific points the difference in fit is substantial. Additional runs with a standard deviation set to a lower value than 0.1 resulted in a clearer failure of the power law to emerge.

In summary, the failure of these three simulations to produce a clear power-law distribution of gene family size demonstrates that heterogeneity of hazard

functions between gene families, correlation within families and a sufficiently low average absolute level of the hazard function are all essential to the emergence of the power-law.

3.3 Testing whether all three processes are essential

In all simulations up to this point, all three evolutionary process are modelled (duplication, sequence divergence, and pseudogenisation), however, it is not clear that all three are necessary for the power-law.

It is quite obvious that just one of these forces alone is not enough to generate the power-law distribution. Duplication alone would result in continuous expansion of gene family size resulting in no small families. Gene death alone would eventually result in no genes and therefore no families. Divergence alone would eventually result in only families of small sizes and in the extreme only families of size one.

However, it is less clear whether leaving one of the forces out would prevent a power-law. Removing duplication would obviously prevent a power-law as gene family size would not grow. Including only duplication and divergence (and removing pseudogenisation) could potentially produce a power-law, if divergence is sufficiently rapid to ensure that smaller family sizes do not become underrepresented. Duplication and death (and no sequence divergence) could potentially also generate a power-law, if sequence divergence is not playing an important part in generating new families.

We ran two simulations based on model 4, one where sequence divergence is removed (model 9) and one where pseudogenisation is removed (model 10). The removal of sequence divergence results in a family size distribution which is almost identical to the model including divergence (compare rows 2 and 3 of figure 3), indicating that pseudogenisation and duplication alone are sufficient to produce a power-law. Note, however, how the fit of the power-law function also begins to decline for $S > 2.5$. This decline is primarily due to the emergence of a few large gene families which may have been lessened had sequence divergence been present and split these families. The removal of pseudogenisation has a dramatic effect on the family size distribution, creating a large under-representation of small families (see row 4 of figure 3). This is evident already at $S = 1.0$, beyond this point we are not able to compute a gene family clustering due to the large number of genes in the simulated genome. Clearly, sequence divergence causing sequences to break away from their original family is not sufficient to prevent small families from becoming under-represented. In summary, we find that duplication and pseudogenisation are necessary and sufficient to produce the power-law, but that sequence divergence may play

a role in the maintenance of the power-law distribution by splitting families and thereby preventing perpetual growth of gene families with low hazard.

4 Discussion

The simulations show that the processes that are essential are gene duplication and pseudogenisation, and that the pseudogenisation rate must have a sufficient level of heterogeneity between families, a sufficient level of homogeneity with families, and a level that is sufficiently low in relation to the duplication rate to allow larger gene families to emerge and stabilise in the genome.

This result is in fact very close to that obtained with a theoretical model of evolution (Huynen and van Nimwegen, 1998), where it was suggested that, in order to explain the power-law distribution, the probabilities of duplications of genes within a gene family must not be independent of each other and the probabilities of deletions of genes within a gene family must not be independent of each other. This effectively means that the retention rate of gene duplicates must be correlated within a family which corresponds with our findings. The novelty of our results are three-fold. First, we model all three key processes and test whether sequence divergence is required for the emergence of the power-law, rather than assuming that it is not relevant. Second, we model pseudogenisation as a time-dependent process as we have shown this to be the case in genomic data (Hughes and Liberles, 2007). Third, and most importantly, we use empirically verified models of gene duplication, loss and sequence divergence. Thus, our results are an essential complement to the results of Huynen and van Nimwegen. They showed that it was *possible* to generate a power-law distribution of gene family size with a very simplified theoretical model of gene family evolution that ignored sequence divergence. We have shown that an empirically verified model that includes only duplication and pseudogenisation, but with heterogeneity of the pseudogenisation rate across families, actually does produce a power-law distribution.

4.1 Modeling of the duplication, pseudogenisation and divergence rates

We have assumed in our model that the duplication rate is time-invariant and equal for all genes. This is a simplification forced on us by the lack of detailed data on the duplication rate of individual genes. In this view, the duplication rate is viewed as the normalized per-gene rate of duplication in a single individual. The population genetic process of fixation and its interplay with other aspects of retention during and after fixation are not explicitly modeled. With this view of the duplication rate, the lack of per-gene variance is somewhat

justified. There is genomic heterogeneity in recombination rates and the existence of tandem duplicates will (through recombination and replication-based mechanisms) increase the per-gene rate of initial duplication. However, given the high rate of pseudogenization, other factors at the post-duplication retention level are expected to play a more major role in gene-specific variation in fate (and also to show greater evolutionary stability).

As previously mentioned, the direct causes of pseudogenisation at the sequence level are stop codon substitutions and frame shift insertions or deletions in coding sequence and deleterious substitutions in regulatory sequence, i.e. pseudogenisation is driven by sequence divergence. This process is complex and difficult to model, and, therefore, we use a pseudogenisation model with S as the independent variable. As a result, the complexity of how sequence substitution and/or indel events (or combinations of events) lead to pseudogenisation becomes incorporated into the hazard function. Further, the complex interplay between pseudogenisation and neofunctionalization (governed by functional sequence mutations, population dynamics and the process of fixation) is also incorporated into the hazard function and assumed to be a function of time (proxied by S) rather than modeled directly (see figure 2). However, these simplifications do not affect the validity of our conclusion that sequence divergence is not necessary for a power-law distribution to emerge.

The requirement that there be heterogeneity of pseudogenisation rates across gene families is consistent with the most prominent models of gene duplicate retention: the subfunctionalisation and neofunctionalisation models. In the subfunctionalisation model (Force et al., 1999), the probability of retention is an increasing function of the number of regulatory modules. Thus, genes with large numbers of regulatory modules have a lower pseudogenisation rate than genes with fewer regulatory modules (all other things being equal) and families consisting of such genes are likely to be larger. In the neofunctionalisation model (Ohno, 1970), the probability of retention is a function of the gene's propensity to accomodate beneficial mutations that fine tune or create new function. It has been established that gene function can affect the substitution rate and that this effect can be different for different functions after speciation events versus gene duplication events (Seoighe et al., 2003). This implies that the retention rate, as it is driven by the divergence rate is also different for different functional categories. Gene function is not the only important determinant of evolutionary rates, as simulations have shown different evolutionary dynamics and retention profiles based upon neo- and sub- functionalization for different protein folds (Rastogi et al., 2006). These function- and fold- specific variations in gene family evolutionary dynamics justify the use of different hazard functions for different gene families. Unfortunately, there are, to our knowledge, no empirical estimates of the variance of hazard functions across gene families, thus this aspect of the model is theoretical. However, this does not detract from the fact that the model employed here is solidly anchored in

biological reality as the key processes have been empirically validated and the existence of a variance in hazard is consistent with the prominent models of gene duplicate retention.

4.2 Use of H. sapiens estimates

The distribution of gene family size has been found to follow a power-law in a wide variety of evolutionarily-distant fully-sequenced genomes, thus, the most parsimonious inference is that this distribution first emerged at the very least many hundreds of millions of years ago, and has been maintained since. Ideally, in a study of the processes that contribute to the emergence of the power-law, we would have reliable estimates of them at the root of the eukaryotic tree. However, producing estimates of these processes in the distant past is effectively impossible due to the directly counteracting nature of duplication and loss, and the saturation of silent sites with time. The best we can do is to quantify the relevant rates in extant species (Hughes and Liberles, 2007). The rates of duplication, loss and divergence that prevailed in the distant past were obviously not the same as those estimated in *H. sapiens*. However, we believe that applying the *H. sapiens* rates in our simulations, as we have done, is a defensible approach: first, because the functional forms are unlikely to be different in an ancestral eukaryote even if it is very likely that the specific values of the parameters were different and, second, because we are then applying a set of rates that are known to be a good approximation to reality and consistent with each other rather than a set of potentially low quality and inconsistent estimates.

4.3 Preferential attachment and power laws

The power-law distribution has been observed for the connectivity of many different types of networks, e.g. web pages (Albert et al., 1999) and enzymes (Jeong et al., 2000; Wagner and Fell, 2001). It has been shown that two "generic" mechanisms of network evolution are sufficient for vertex connectivity to follow the power-law (Barabasi and Albert, 1999): i) the network expands continuously by the addition of new vertices, and ii) new vertices attach preferentially to sites that are already well connected. In our model the connections between the genes (vertices of the network) are not physical connections as in the connections between web pages by html links or the connections between metabolites by enzyme catalysed reactions, instead they are measures of sequence similarity which biologists use as a proxy measure for common origin (homology) and function. The expansion of the "network of homology" by the addition of new vertices is ensured by the duplication of

genes and an average pseudogenisation rate that is sufficiently low (as shown by model 6). But, these duplicated genes cannot preferentially attach and must attach to the gene from which it duplicated (and to that gene's homologs). However, if homologous families have homogeneity of hazard functions within the family and heterogeneity between families, then families with low hazard will grow while families with high hazard will not. Thus, although the way in which the connectivity of the "network of similarity" expands is driven by whether or not a new vertex (duplicate gene) is retained, homogeneity within families and heterogeneity between families ensures that it expands in a fashion that mimics preferential attachment. This observation of how the semblance of "preferential attachment" occurs in the case of gene families may be applicable to other domains as there are many types of network in which new vertices do not emerge ab-initio and then preferentially attach, but instead follow a model of evolution similar to homologous gene families, i.e. new vertices are already linked to existing vertices upon creation. Our finding shows how, in a network where the connectivity of a new vertex is predetermined, if there is heterogeneity of vertex death probability, "preferential attachment" is mimicked and a power-law emerges.

4.4 Speed of emergence and slope of power laws

In the simulation that actually produces a power-law distribution (model 4), the power-law had clearly emerged by $S = 2.0$. The emergence of the power-law is inevitably connected with genome growth (Barabasi and Albert, 1999) and by $S = 5.0$ the number of genes is 3,853. None of the other models (excluding the models where a process was removed) produced such growth in the number of genes. This growth is due to the fact that, in model 4, family founding genes with sufficiently low hazard functions are produced and used to initialise the simulation. These low hazard functions are the result of a high enough standard deviation of the error term and the fact that, when these genes duplicate, the duplicates inherit the same low hazard function through the inherited error term. Because these genes have a lower hazard function, they are more likely to be retained in the genome and, thus, grow in number and lower the average hazard function. Such a lowering of the average hazard function does not occur to the same extent in the other models; either because all genes have the same hazard function (models 1, 2, and 3); or because the error is not inherited, thus ensuring that a new duplicate has a hazard function drawn from a fixed distribution (model 5); or because the average hazard function is higher (model 6); or, finally, because the standard deviation of the error is too low, so that not enough families are founded by genes with a low enough hazard. By varying the value of the mean or the standard deviation of the error term distribution in model 4, we affect the emergence of the power-law: increasing the mean (model 6) or decreasing the standard

deviation (model 7) slow down or prevent the emergence of the power-law, whereas decreasing the mean or increasing the standard deviation result in a more rapid emergence of the power-law (unreported data). If the error mean and standard deviation are set to values that ensure a rapid emergence of the power-law, the simulation produces a power-law distribution with a slope that gradually becomes less and less negative. At the same time, the number of genes grows more rapidly due to families founded by genes with low hazard becoming larger and more numerous. This is consistent with the empirical observation that there is a weak positive correlation between the exponent of the power-law and the logarithm of the number of genes in the genome (Huynen and van Nimwegen, 1998). This positive correlation implies a relative increase of the number of large clusters over the number of small clusters as the number of genes in the genome increases. These large clusters are most likely those containing genes with two key properties: a relatively low hazard function and a tendency for this to be inherited by duplicates.

4.5 Conclusion

We have shown that duplication and pseudogenisation are necessary and sufficient for the emergence of a power-law distribution of gene family size. The duplication rate needn't be different for different families, but the death rate must have a certain degree of heterogeneity between families and homogeneity within families. This requirement is consistent with both the neofunctionalisation and subfunctionalisation models of gene duplicate retention. In addition, the average death rate must be low enough that the number of genes expands. We find that the role of accumulation of replacement substitutions leading to sequences forming new families after passing a divergence threshold is not essential for the emergence of a power-law. However, that is not to say that replacement substitutions are not important, as they are indeed important to the retention/pseudogenization process where their role is simplified into the hazard function which is critical to the emergence of the power-law. Moreover, sequence divergence might play a role in a "quantitative" sense in the maintenance of the power-law by "splitting" families and, thus, preventing the emergence of very large families which do not fit the power-law.

5 Acknowledgements

The work has been funded by FUGE, the functional genomics platform of the Norwegian research council.

References

Albert, R., Jeong, H., Barabasi, A. L., 1999. Internet: diameter of the world-wide web. Nature 401, 130–131.

Albert, R., Jeong, H., Barabasi, A. L., Jul 2000. Error and attack tolerance of complex networks. Nature 406 (6794), 378–382.

Altschul, S. F., Madden, T. L., Schffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., Sep 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389–3402.

Barabasi, A. L., Albert, R., Oct 1999. Emergence of scaling in random networks. Science 286 (5439), 509–512.

Birney, E., Andrews, D., Caccamo, M., et al. (51 co-authors)., Jan 2006. Ensembl 2006. Nucleic Acids Res. 34 (Database issue), D556–D561.

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

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

Enright, A. J., Kunin, V., Ouzounis, C. A., Aug 2003. Protein families and TRIBES in genome sequence space. Nucleic Acids Res. 31 (15), 4632–4638.

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

Gan, G., Ma, C., Wu, J., 2007. Data Clustering: Theory, Algorithms, and Applications. Society for Industrial and Applied Mathematics.

Harrison, P. M., Gerstein, M., May 2002. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J. Mol. Biol. 318 (5), 1155–1174.

Hughes, T., Liberles, D., Oct 2007. The Pattern of Evolution of Smaller-Scale Gene Duplicates in Mammalian Genomes is More Consistent with Neo- than Subfunctionalisation. J. Mol. Evol. published online.

Huynen, M. A., van Nimwegen, E., May 1998. The frequency distribution of gene family sizes in complete genomes. Mol. Biol. Evol. 15 (5), 583–589.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., Barabsi, A. L., Oct 2000. The large-scale organization of metabolic networks. Nature 407 (6804), 651–654.

Kamal, M., Luscombe, N. M., Qian, J., Gerstein, M., 2005. Analytical evolutionary model for protein fold occurrence in genomes, accounting for the effects of gene duplication, deletion, acquisition and selective pressure. Springer Science and Business Media, Ch. 10.

Karev, G. P., Wolf, Y. I., Rzhetsky, A. Y., Berezovskaya, F. S., Koonin, E. V., 2002. Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC. Evol. Biol. 2 (1), 18.

Luscombe, N. M., Qian, J., Zhang, Z., Johnson, T., Gerstein, M., Jul 2002. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome Biol. 3 (8).

Lynch, M., Conery, J. S., Nov 2000. The evolutionary fate and consequences of duplicate genes. Science 290 (5494), 1151–1155.

Nei, M., Kumar, S., 2000. Molecular evolution and phylogenetics. Oxford University Press.

Ohno, S., 1970. Evolution by gene duplication. New York: Springer-Verlag.

Promponas, V. J., Enright, A. J., Tsoka, S., Kreil, D. P., Leroy, C., Hamodrakas, S., Sander, C., Ouzounis, C. A., Oct 2000. CAST: an iterative algorithm for the complexity analysis of sequence tracts. Bioinformatics 16 (10), 915–922.

Qian, J., Luscombe, N. M., Gerstein, M., Nov 2001. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J. Mol. Biol. 313 (4), 673–681.

Rastogi, S., Reuter, N., Liberles, D. A., Nov 2006. Evaluation of models for the evolution of protein sequences and functions under structural constraint. Biophys. Chem. 124 (2), 134–144.

Salzberg, S. L., White, O., Peterson, J., Eisen, J. A., Jun 2001. Microbial genes in the human genome: lateral transfer or gene loss? Science 292 (5523), 1903–1906.

Seoighe, C., Johnston, C. R., Shields, D. C., Apr 2003. Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation. Mol. Biol. Evol. 20 (4), 484–490.

Wagner, A., Fell, D. A., Sep 2001. The small world inside large metabolic networks. Proc. Biol. Sci. 268 (1478), 1803–1810.

Yanai, I., Camacho, C. J., DeLisi, C., Sep 2000. Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Phys. Rev. Lett. 85 (12), 2641–2644.

Yang, Z., Nielsen, R., Apr 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46 (4), 409–418.

Zipf, G., 1949. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley Press.

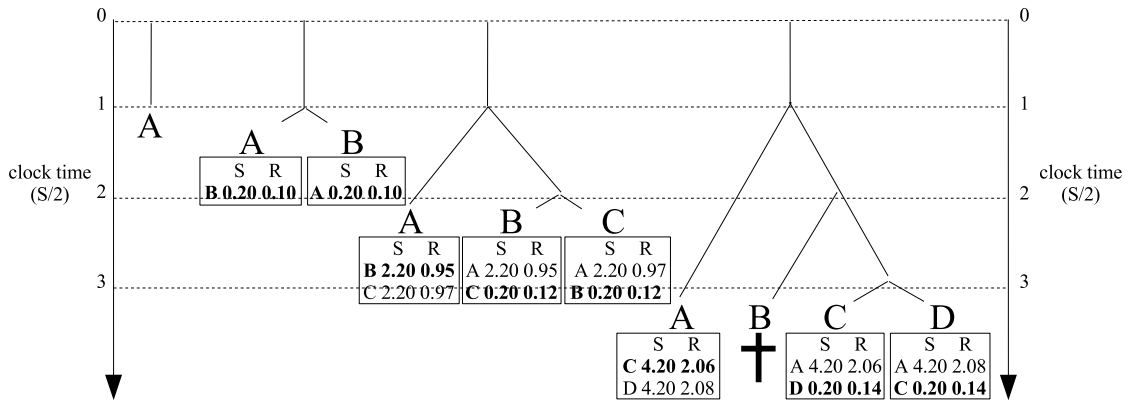

Fig. 1. Simulation of homologous gene family evolution (Example scenario)
Boxes contain distances to homologous genes (first column: distances in units of silent substitutions per silent site, S; second column: distances in units of replacement substitutions per replacement site, R). R accumulates more slowly than S (see equation 1) and does so in a non-deterministic fashion. This explains why the R distances A-B & A-C (in the third tree), and A-C & A-D (in the fourth tree) are not equal. The non-pseudogenised gene with the shortest R distance to a gene of interest is that gene's closest homolog (indicated in bold in the figure). We assume that silent substitutions per silent site S accumulate at a constant rate between genes in a duplicate pair. Thus, real time (represented by the axis) can be measured in units of $S/2$.

The figure depicts the following scenario:
1. A gene family is founded by a singleton gene A which initially is the only member of the family.
2. At some point gene A is subject to a duplication and produces gene B. We illustrate the situation where the duplication happened 0.1 clock ticks ago i.e. the distance between the two genes is 0.2 S. At this point A and B are each others closest non-pseudogenised homologs.
3. Then, at a later point gene B duplicates and produces gene C. We again depict the situation shortly after this duplication and assume that B has accumulated slightly fewer replacement substitutions than C since the duplication. Thus, B is A's closest homolog, and B and C are each others closest homologs.
4. Finally, C duplicates and produces gene D and some time before this duplication occurs gene B pseudogenises. Because C has accumulated slightly less replacement substitutions since the duplication, C becomes A's closest homolog, and C and D are each others closest homologs.

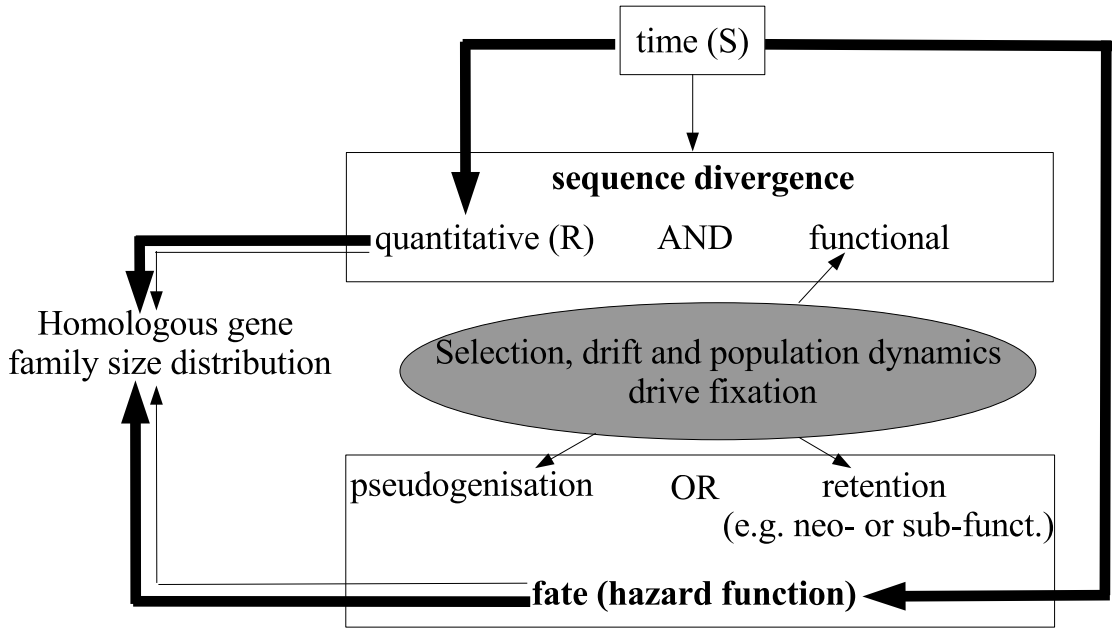

Fig. 2. Summary of simplifying aspects of the model

Thin arrows: actual processes.

Bold arrows: processes modeled and incorporated into the simulation.

The figure illustrates how our model simplifies reality by only modeling the quantitative aspect of sequence divergence and by modeling retention fate through the hazard function (with S as the independent variable). Thus, the complex process, through which most duplicates pseudogenise while others are retained due to specific functional mutations that reach fixation, is not explicitly modeled.

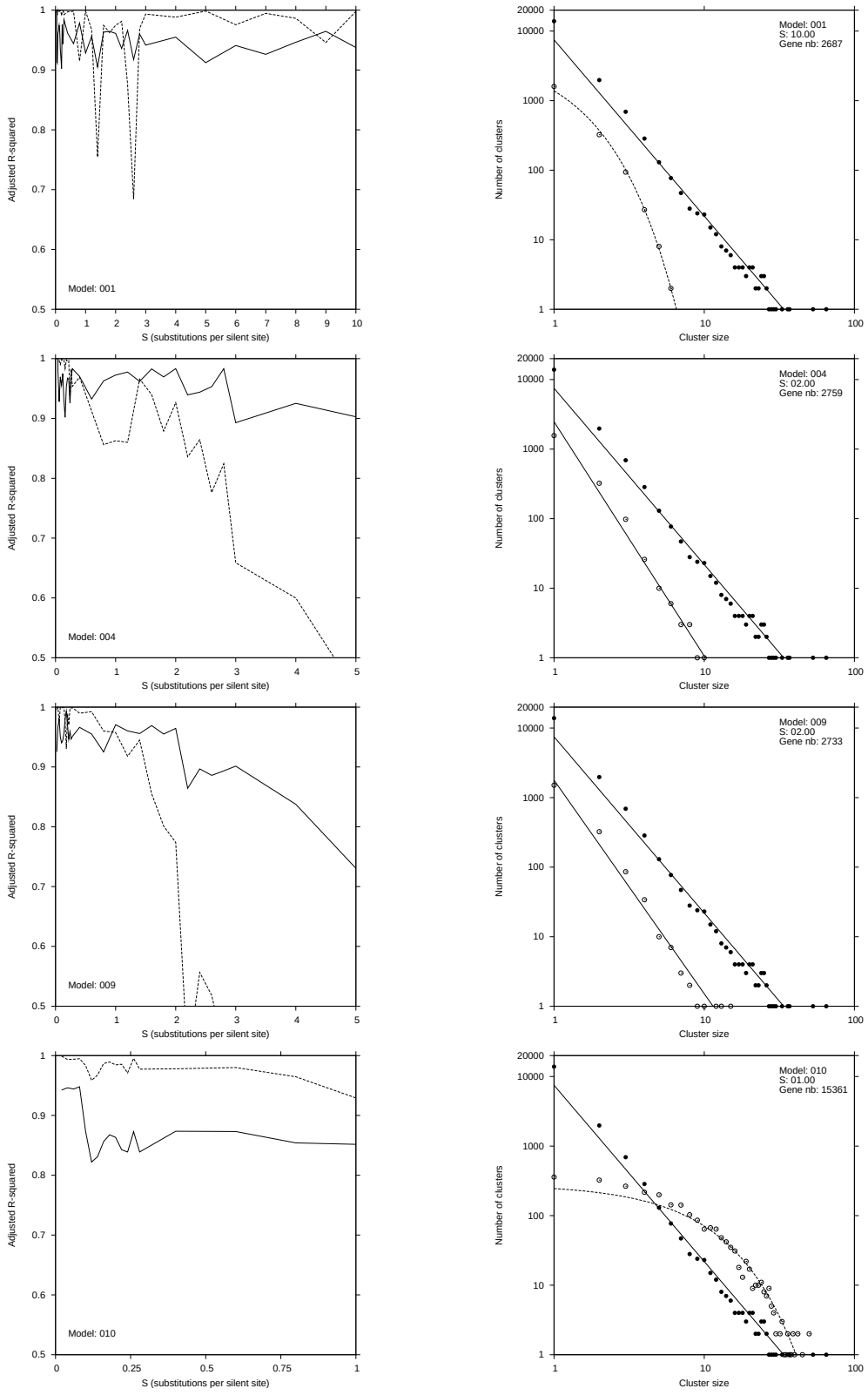

Fig. 3. Function fit and family size distribution plots (see table 1 for model details)
First column: R^2 of fit of exponential and power-law function to fam. size distrib. at successive S.
Second column: fam. size distrib. and best fitting function (exponential or power-law) at specific S.

Solid line: power-law function. *Dotted line:* exponential function.
Black points: real gene family size data computed from the Ensembl annotation of the *H. sapiens* genome.
White points: gene family size data computed from the simulation.

Model	Initialisation singletons	Max S	Probability function error: Mean	Stand. dev.	Inherited error	Processes: Gene death	Repl. subst.	Results: Best fit at max S	Genome size at max S
Same hazard fct									
1. Basic	2,000	10.0	0.0	0.0	na	yes	yes	exponential	2,687
2. Longer S	-	***30.0***	-	-	-	-	-	exponential	2,375
3. Higher initial nb of genes	***10,000***	10.0	-	-	-	-	-	exponential	13,429
Different hazard fcts									
4. Basic	2,000	5.0	0.0	0.2	yes	yes	yes	power	3,853
5. Error not inherited	-	10.0	-	-	***no***	-	-	unclear	3,000
6. Higher average hazard	-	10.0	**0.2**	-	-	-	-	unclear	2,667
7. Lower error variance	-	10.0	-	***0.1***	-	-	-	unclear	3,076
Process removed									
8. Basic (identical to run 4)	2,000	5.0	0.0	0.2	yes	yes	yes	power	3,853
9. No repl. subst.	-	5.0	-	-	-	***no***	-	power	3,937
10. No gene death	-	1.0	-	-	-	-	***no***	exponential	15,361

Table 1
Details of simulation models

The models are divided into three groups. Within each group the first model is the basic model for the group and subsequent models within the group vary one of the features of the basic model (bold text) with all other features remaining the same as the basic model for the group (dash).

Initialisation singletons: the number of genes used to initialise the simulation.

Max S: the time for which the model is run measured in silent substitutions per silent site.

Mean and standard deviation: the mean and standard error of the normal distribution from which the hazard error is drawn.

Inherited error: indicates whether the error is inherited by a duplicate or whether a new error is sampled from the normal distribution for the new gene.

Processes: duplication is always applied but gene death (pseudogenisation) and replacement substitutions may be removed from the model.

Best fit at max S: whether the exponential or power-law function provides the best fit at max S.

Genome size at max S: the number of non-pseudogenised genes at max S.

Supplementary materials

Timothy Hughes [a] David A. Liberles [b,*]

[a] *Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway*

[b] *Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA*

Abstract

This file contains information on the supplementary materials for the article entitled "The power-law distribution of gene family size is driven by the pseudogenisation rate's heterogeneity between gene families". This includes: animation files, R code and figures of gene family size distribution for all models.

1 Simulation animations

Unzip the file animations.zip. Open the animations.html file in a browser. This file provides a summary of all available animations and from this file it is possible to link the details of each animation.

2 Simulation code

The code is written in Java and is available in the simulationCode.zip file. This file can be unzipped and its contents can be inspected to gain an understanding of the details of the simulations' code. A good starting point for such an inspection is the "simulations" directory which is structured in the same way as table 1. The code is extensively commented.

To actually run one of the simulations, a user will need to:
1. modify the file paths in the relevant simulation .java file.
2. ensure that all classes of the .jar file are on the classpath.

* Corresponding author.
Email address: liberles@uwyo.edu (David A. Liberles).

3. ensure the maths package colt.jar file is also on the classpath (available from http://dsd.lbl.gov/~hoschek/colt).
Note, however, that this code was not designed as a software application and, therefore, running modified code might not be straight forward.

3 Figures for all models

Figures of the distribution of gene family size for all models (see table 1 in the article for an overview over all models).

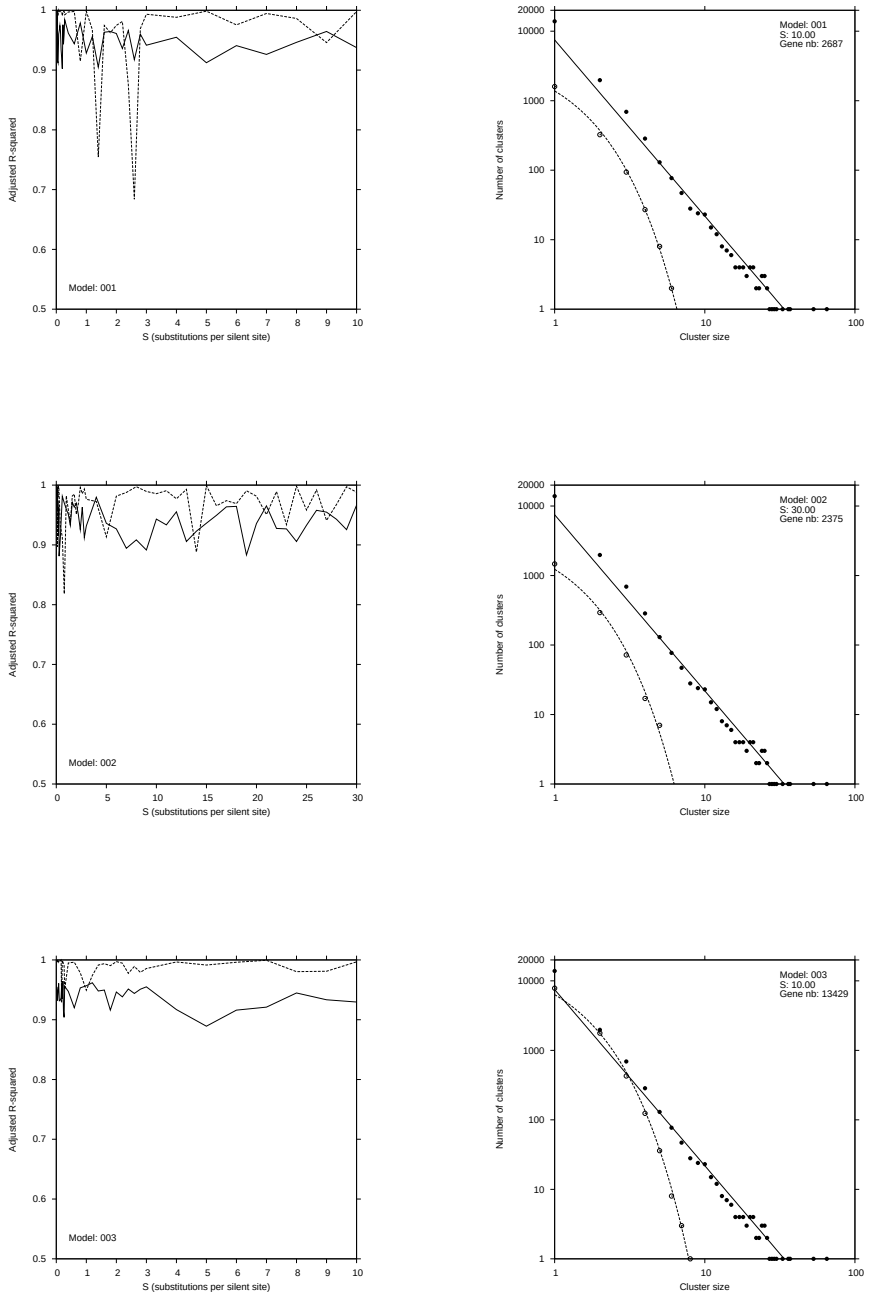

Fig. 1. Group 1 models (same hazard function for all genes)
Model 1: basic; *model 2:* longer S; *model 3:* higher number of initial genes.
Solid line: power-law function. *Dotted line:* exponential function.
Black points: real gene family size data computed from the Ensembl annotation of the *H. sapiens* genome.
White points: gene family size data computed from the simulation.

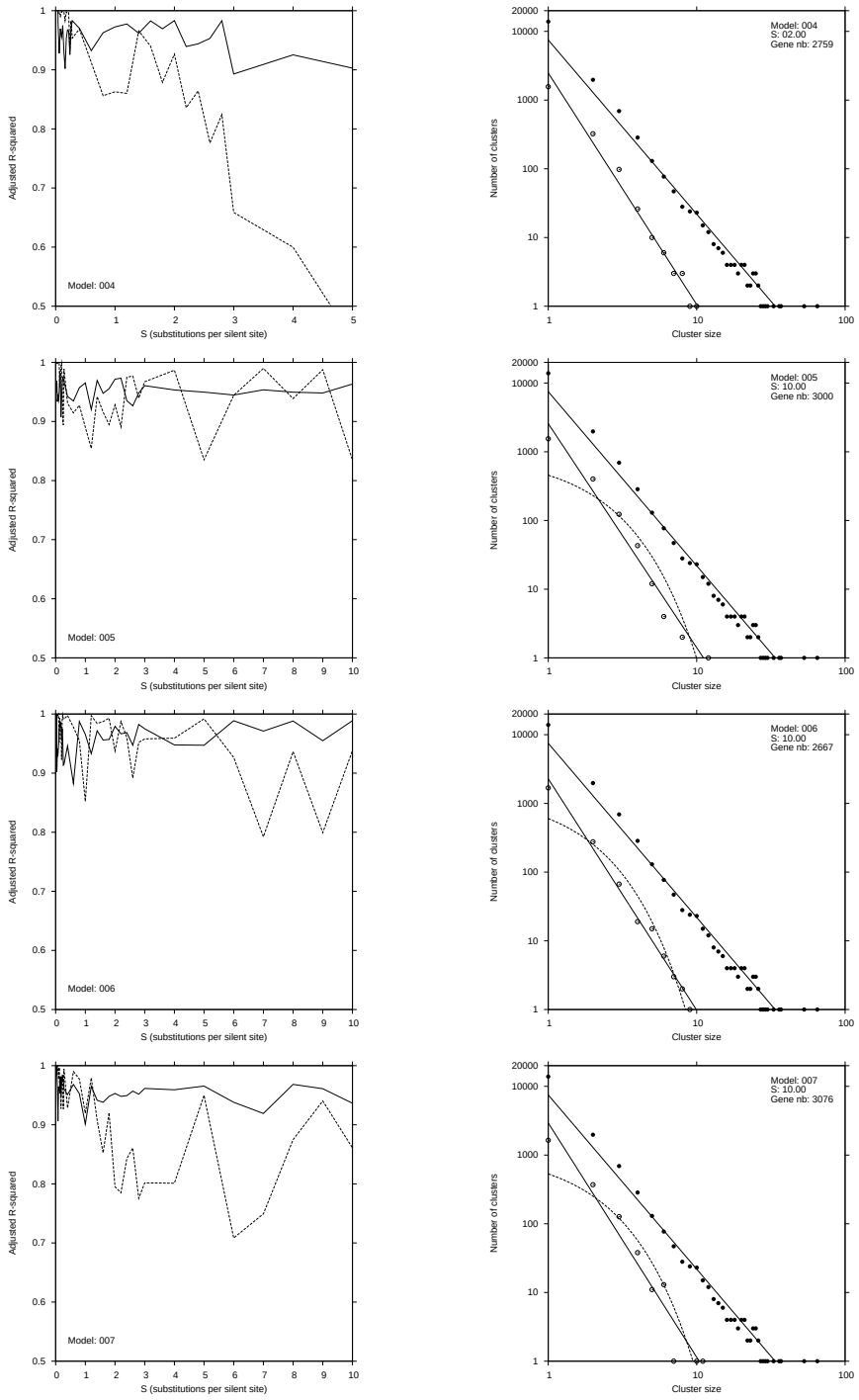

Fig. 2. Group 2 models (different hazard functions for different genes)
Model 4: basic; *model 5:* error not inherited; *model 6:* error mean greater than 0; *model 7:* low error standard deviation.
Solid line: power-law function. *Dotted line:* exponential function.
Black points: real gene family size data computed from the Ensembl annotation of the *H. sapiens* genome.
White points: gene family size data computed from the simulation.

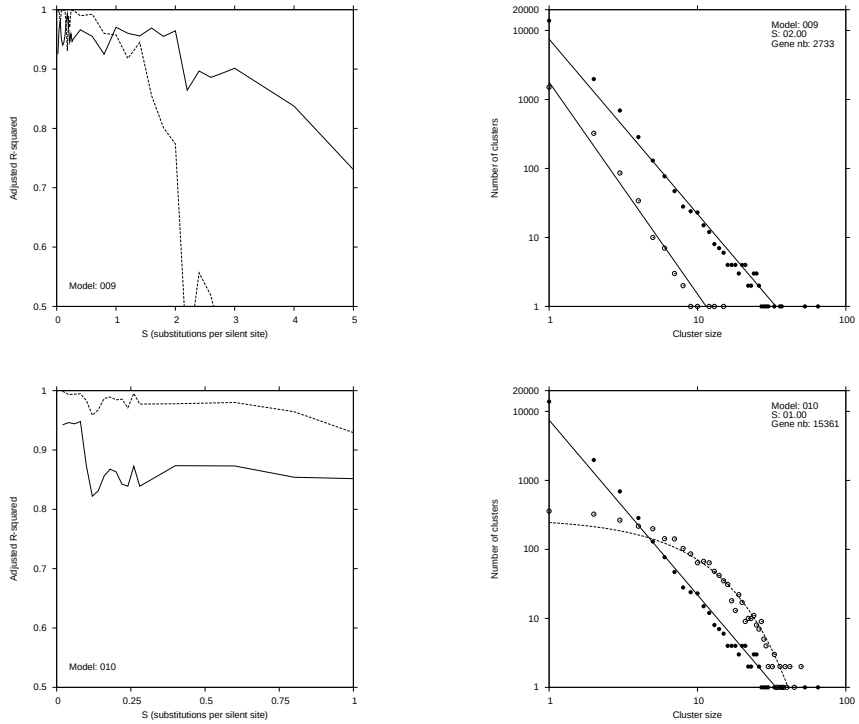

Fig. 3. Group 3 models (one process removed)
Model 9: no replacement substitution; *model 10:* no pseudogenisation
Solid line: power-law function. *Dotted line:* exponential function.
Black points: real gene family size data computed from the Ensembl annotation of the *H. sapiens* genome.
White points: gene family size data computed from the simulation.

C. Paper III

The whole genome duplications in the ancestral vertebrate are detectable in the distribution of gene family sizes of tetrapod species

Submitted to the Journal of Molecular Evolution

Timothy Hughes and David A. Liberles

TH: Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway.
Telephone: (+47) 55 58 40 72. Email: tim@bccs.uib.no

DAL: Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA.
Telephone: (+1) 307 766 5206. Email: liberles@uwyo.edu.

Correspondence: TH and DAL

Keywords: gene duplication, whole genome duplication, pseudogenisation, non-synonymous substitution, gene family size, power-law distribution, speciation.

Abstract

A clustering of all protein coding genes from the complete genomes of five tetrapod species into gene families, shows a clear deviation from the expected power-law distribution of gene family size. We hypothesise that at least part of the deviation is the result of the two whole genome duplications (WGD) that are now known, with reasonable certainty, to have occurred prior to the fish-tetrapod split. We build a model of homologous gene family evolution and perform simulations to show that speciations alone cannot produce a distribution that resembles the empirical data. In order to replicate the features of the empirical distribution, the simulation must incorporate two WGD events. In addition, these WGDs must be such that a significant proportion of the gene duplicates generated in the WGDs have a higher retention rate than they do following small-scale duplication (SSD). This requirement is consistent with what is known about duplicate retention following a WGD, namely that genes belonging to specific functional classes, such as genes regulating transcription, are much more likely to be retained following WGD than SSD. We conclude that the deviation from the power law that we observe in the empirical data is the result of the two WGDs that occurred in the ancestral chordate. This implies that the two ancient WGDs continue to have a structural effect on gene families approximately 500 million years after the initial events. On the one hand, this is a surprising result given the limited retention of duplicates generated by a WGD and the continual small-scale duplication which further weakens the signal created by the fraction of duplicate pairs that are retained. On the other hand, WGD's capacity to fundamentally change the architecture of gene families in a profound and lasting way is consistent with the observed correlation between WGDs and important evolutionary transitions.

Introduction

There are different methods for building clusters of homologous genes, but a clustering of all protein coding genes in a genome, irrespective of the method used, produces many small clusters and few large clusters (Huynen and van Nimwegen 1998; Yanai et al. 2000; Harrison and Gerstein 2002). The functional form that best fits the data is the power-law (Huynen and van Nimwegen 1998; Luscombe et al. 2002): $N = aF^b$ where F is the family size and N is the number of families of this size or, taking the natural logarithm, $\ln(N) = \ln(a) + b.\ln(F)$ i.e. a linear relationship on a log-log plot (see figure 1). The exponent b is usually in the range -4.0 to -2.75 and there is a weak positive correlation between the exponent and the logarithm of the number of genes in the genome (Huynen and van Nimwegen 1998).

We have previously shown through simulation what are likely to be the key processes causing the emergence of a power-law (Hughes and Liberles 2007*b*). If we assume a constant small-scale gene duplication rate (tandem and segmental duplications), and we use specifications of the pseudogenisation rate and sequence divergence rate that have been validated using genomic data, the minimal requirements for such a distribution to emerge from an initial set of singleton genes are that the genes are subject to duplication and loss, and that there is heterogeneity of the rate of loss across gene families. Once a power-law distribution emerges, and assuming that there are no large-scale duplication events, the intuition as to why such a distribution is maintained is simple: genes in all families duplicate, but the vast majority rapidly pseudogenise (Lynch and Conery 2000; Lynch and Conery 2003; Hughes and Liberles 2007*a*). Some families may have a lower pseudogenisation rate causing such families to increase in size relative to families with a higher rate, but this is a slow process due to the strength of pseudogenisation. In addition, sequence divergence will be a moderating factor allowing older duplicates to accumulate sufficient replacement substitutions and split away from their original family, thus limiting the size of large families (Hughes and Liberles 2007*b*).

Of course, lineage-specific expansions and contractions do occur in certain gene families, for example, on the human lineage, the GAGE gene family seems to be expanding while the olfactory receptor gene family seems to be contracting (Gilad et al. 2003). This has been termed the revolving door mechanism (Demuth et al. 2006) and is expected to be non-random and related to both gene function (Maere et al. 2005) and protein fold (Rastogi et al. 2006). However, this dynamic process does not appear to cause major deviations from the power-law distribution of gene family sizes, probably because it only affects a limited number of families in any one lineage.

It has also been shown that the power-law distribution applies to the clustering of genes from multiple complete genomes, if the species concerned are evolutionary distant (Enright et al. 2003). We were therefore initially intrigued by the observation of a strong deviation from the power-law when clustering the genes from the complete genomes of five tetrapod species, with clear "waves" with a period of size 5, visible for sizes 1 to 15 (see figure 2). These "waves" are very pronounced: not only do the frequencies not follow a linear relationship in the log-log plot, but it is also the case for several sizes that the frequency of size x is less than the frequency of size $x + 1$. However, in this case, although the species' divergence times spanned several hundred million years, they were not as distantly related as in the Enright et al. study. Moreover, a convincing case, based on gene family phylogenetic trees and genomic map position data, has been made in favour of the hypothesis that the genome of the ancestral vertebrate was subject to two whole genome duplications (Dehal and Boore 2005). We hypothesise that the deviation

from the power-law observed in the empirical data is at least partly the result of these two whole genome duplications.

It is obvious that if a whole genome duplication reaches fixation, it will initially dramatically alter the gene family size distribution: the frequency of families of size x will become the frequency of families of size $2x$ and odd sized families will become effectively non-existent. Speciation initially has a similar effect if we consider a clustering of all genes from both descendent species. However, it is less straightforward to establish the effect of small-scale duplication and loss following the initial event: whether a power-law distribution returns (and if so on what time-scale) and what the effect of a combination of WGD and multiple speciations has on the gene family size distribution.

There has been considerable debate surrounding the hypothesis of none, one or two WGDs in the ancestral chordate (0R/1R/2R). Initial efforts to prove the WGD hypothesis centered on the size of gene families. The opponents of WGD argued that there was no signal of a WGD in this kind of data (Friedman and Hughes 2001). A recent study has produced strong evidence of 2R through the genomic mapping of paralogous regions known to have arisen before the fish-tetrapod split (Dehal and Boore 2005). However, they too claim that there is no signal of WGD in gene family size data. These findings directly contradict our hypothesis that the pattern we observe in the empirical data is caused by the ancestral WGDs. The reason that these studies found no signal of the WGDs in gene family size data, is that they did not compare the empirical distribution to the distribution that would be expected in the absence of WGDs. In previous work (Hughes and Liberles 2007*b*), we have shown that small-scale duplication and loss results in a power-law distribution and, thus, is the expected distribution in the absence of a WGD. In this paper, we extend our model of homologous gene family evolution to incorporate WGD and speciation, and use it to simulate the evolution of the distribution of gene family size under different scenarios (WGD, speciation, and WGD followed by multiple speciations). The output of the simulations show that, in order to produce a deviation from the power-law which is consistent with the empirical data, the simulation must incorporate not only speciations but also two WGDs. We conclude that the deviation from the power law that we observe in the empirical data is the result of the two WGDs that occurred in the ancestral chordate.

Results

The simulations use a model of homologous gene evolution which incorporates gene duplication, sequence divergence and pseudogenisation. It is impossible to obtain accurate estimates of these processes for the 500 million years that separate us from the two WGDs,

as a second best the model is parameterised using *H. sapiens* data (see the discussion for the detailed reasons for this approach). Time is measured in units of silent substitutions per silent site between duplicate pairs (S) and, in *H. sapiens*, 1 S corresponds approximately to 230 million years (Yang and Nielsen 1998). The model allows not only for small-scale duplication (SSD) caused by tandem and segmental duplication, but also for speciation and WGD events (see the "Materials and Methods" section for full details of the model). WGD events are implemented in two different ways: the first implementation assumes that a gene duplicate's pseudogenisation rate (also referred to as "hazard rate") following a WGD is the same as following SSD (the "no hazard shift" model); in the second implementation, gene duplicates that have a high hazard rate following SSD are given a very low hazard rate following WGD, while gene duplicates with a low hazard rate following SSD keep the same hazard rate following WGD (the "hazard shift" model). This second implementation is consistent with models such as dosage balance (Aury et al. 2006) and subfunctionalisation (Force et al. 1999) which have been developed to explain the higher retention following WGD as compared to SSD, and there is mounting evidence that these models fit the data on the retention of duplicates generated through WGD (see the discussion for the details of this point).

Whole genome duplication

Immediately following the fixation of a WGD event, the frequency of families of size x will become the frequency of families of size $2x$ and odd sized families will become effectively non-existent. Assuming that the distribution of gene family sizes followed a power-law prior to the WGD, then the post-WGD distribution of even sized families should also follow a power-law with the same exponent but a higher intercept (see figure 3). As time passes since the WGD, if small-scale gene duplication returns, and gene loss and sequence divergence happens in an SSD manner (for all genes irrespective of whether they were generated by the WGD or not), then we would expect to see the families of odd sizes increase in number as even sized families increase in size by 1 through SSD and decrease in size by 1 through loss (see figure 4). Loss should be rampant due to all genes being recent duplicates which are subject to a high pseudogenisation rate (Hughes and Liberles 2007*a*). This should result in the return of the power-law with a similar exponent to the original power-law and an intercept below the intercept that prevailed immediately following the WGD, but higher than the pre-WGD intercept due to the retention of some of the WGD duplicates. However, if we consider that a significant proportion of gene families are subject to a downward shift in the hazard rate following WGD and that families experiencing this shift are drawn from families that have a high hazard rate following SSD, then the return to the power-law will be slower. The reasons for this are two-fold

and are both connected to the tendency for the hazard shifted families to retain their post-WGD size: they are less likely to diminish in size through loss because of the reduced pseudogenisation rate that applies to the duplicates generated by WGD and less likely to increase in size due to the high hazard rate that applies to duplicates that may be generated by SSD after the WGD.

In order to verify these predictions, we carry out two simulations: one where WGD duplicates behave in the same manner as small-scale duplicates (the "no hazard shift" model) and one where a certain fraction of families are subject to a hazard shift following WGD (the "hazard shift" model). The reader is referred to the "Speciation and whole genome duplication extensions" sub-section in the "Materials and Methods" section for the details of the "no hazard shift" and "hazard shift" implementations of the WGD event. In both cases, the WGD is carried out at $S = 2.0$ as it takes approximately this time for the power-law distribution to emerge from the initial state of the model through small-scale duplication and loss. We consider the power-law distribution to have returned for a certain size range when the frequency of these sizes has an approximately linear downward sloping relationship in a log-log plot and when, for all sizes within this range, the frequency of size x is stably greater than the frequency of size $x + 1$.

As predicted, there is a relatively rapid return towards the pre-WGD distribution for the "no hazard shift" model (see figure 5). Note that families of size one remain under-represented 0.2 S after the WGD due to the fact that increase in the number of families of this size are driven by loss from families of size 2 as opposed to loss and duplication for all other family sizes. Nevertheless, by $S = 2.3$ (i.e. 0.3 S after the WGD), the power-law has returned for all sizes for the "no hazard shift" model (see figure 6). In contrast, in the "hazard shift" model, singletons are still under-represented due to the lower loss rates "hazard shifted" families are subject to (see figure 6). The comparison in figure 6 also shows how the "hazard shift" model results in a higher retention of WGD duplicates and, thus, less of a shift back towards the origin than in the "no hazard shift" model.

Speciation

If we consider the clustering of the genes of two genomes that recently speciated, the initial effect of the speciation is the same as that of a WGD. Subsequently, as after WGD, duplication and loss should begin to smooth the distribution, but loss is more restricted following speciation because it is not associated with a sudden increase in the number of young duplicates which are subject to a higher pseudogenisation rate (Hughes and Liberles 2007*a*) . This should slow the return to the power-law as compared to both WGD models. In addition, there is expected to be little loss of singletons in each species which we model by ensuring that singleton families have a near-zero probability of losing

their last member. These singletons form families of size two when clustering the genes from both genomes, and the restricted loss of singletons from each species means that the number of families of size one can effectively only increase through the divergence of sequences from larger-sized families causing the sequence to "break-away" to form a new singleton family. Thus, the increase in the number of families of size one is expected to be particularly slow.

We run a simulation in which we perform two speciation events. The qualitative predictions are fullfilled as expected. Note in particular the differences between the two speciation events: after the first speciation, there is a big disruption to the distribution and a slow return to the power-law; whereas after the second, the disruption is less because only one of the two species radiates and, thus, the distribution for sizes greater than two recovers a power-law shape more rapidly (see supplementary materials, figure 1 for the details of the first speciation event at $S = 2.0$ and figure 2 for the second at $S = 2.5$). As explained earlier, the very high under-representation of families of size one after the first speciation (and of families of size one and two after the second speciation) is to be expected, but is perhaps exagerated in the simulation as our model effectively does not allow singletons to be lost from a genome (see figure 7).

The underrepresentation of smaller gene family sizes observed here has also been observed in the distribution of gene family sizes in gene family databases like TAED (Roth et al. 2005) that were built using entirely different methodology and as such is not an artifact of the gene family construction process. This may in fact reflect a set of core functions, where deletion from a genome is highly deleterious. While the core set of functions necessary for parasitic prokaryotic life has been estimated to be 500-600 genes (Koonin 2003), that set for vertebrate life might be expected to be larger. It has been suggested that informational genes (involved in the retention of biological information) represent a phylogenetic core in bacterial species (Rivera et al. 1998) and a similar expanded core might be expected in vertebrate species.

Whole genome duplication followed by speciation

Our first two sets of simulations have established that: 1) following a WGD without hazard shift a power-law distribution returns within a relatively short period of time for all sizes, 2) for a WGD with hazard shift, a power-law also returns, but families of size 1 are underrepresented and the intercept is higher than in the "no hazard shift" model, 3) following multiple speciation events, a power-law returns, but it takes longer than following WGD, the intercept is significantly higher than pre-speciation, and sizes less than the number of species are under-represented relative to the power-law defined by larger sizes. This strongly suggests that the under-representation of family sizes less than the number

of species, which we observe in the empirical data, can be explained by the speciations, but that speciation events alone or WGD events without hazard shift cannot explain the "waves" observed in the empirical data for family sizes greater than the number of species (see figure 2). In fact, given the output of the simulation of WGD with hazard shift, it is also difficult to see how it can explain a deviation from the power-law for these larger family sizes: we are able to detect a signal when comparing the output of simulations with and without hazard shift (see figure 6) but, if we were to observe the output of the "hazard shift" model alone, it would be difficult to argue that a signal was still observable.

We now run simulations that combine both WGD and speciation events to investigate whether the signal that remains following a WGD with hazard shift continues to be detectable when it is followed by multiple speciation events and whether the qualitative features of the signal are consistent with our empirical observation of "waves" with a period equal to the number of species.

We start by running the same two simulations as in the WGD section, i.e. with and without hazard shift, but we let the WGD be followed by two speciations. The results of the simulations confirm that the absence of a signal for the WGD without hazard shift persists following two speciations as can be seen in figure 8 where the power-law returns within 0.1 S of the second speciation. On the other hand, the underrepresentation of singletons produced by the WGD with hazard shift (as observed in figure 6) is not removed, but does shift to another size due to the speciation event: immediately prior to the second speciation, i.e. 0.5 S since the first speciation, the frequency of families of size three is almost equal that of families of size four (see the white dot distribution in the first graph of figure 9). This is the result of the reduced loss from "hazard shifted" families. Singletons are under-represented before the first speciation due to the reduced loss, thus, following the first speciation which results in two species, families of size two consisting mainly of one gene from each species are underrepresented and size four which consists mainly of a duplicate pair generated by the WGD for each species are overrepresented. Interestingly, the second speciation event would appear to make the signal of the WGD event with hazard shift stronger (see the clear and persistent underrepresentation of families of size five relative to size six in graphs 2, 3 and 4 of figure 9).

To understand how the frequency of families of size five can be less than the frequency of families of size six following the second speciation in the model with hazard shift, we investigate the details of the gene family size distribution immediately prior to the second speciation and compare it to the equivalent situation for the model without hazard shift (see table 1). There are two clear differences between the simulated data for the two models prior to the second speciation. First, the frequency of size four in the model with hazard shift is more than double that in the model without hazard shift. Second,

the species composition of size 4 is radically different between the two models with a clearly higher frequency of two sequences from each species in the model with hazard shift. Since families of size 4 with two sequences from each species become families of size 6 following the second speciation, this explains how families of size six become over-represented relative to families of size five. The speciation event is effectively combining these two signals to produce a stronger unified signal which is detectable as a clear deviation from the power-law distribution.

Finally, we run a simulation with two hazard shifted WGDs followed by two speciations to test whether a simulation can produce a distribution that resembles the empirical data, which we know with reasonable certainty was subject to two WGDs prior to the fish-tetrapod split dated to approximately 476 million years ago (Blair and Hedges 2005). We separate the WGDs by 0.2 S which might correspond approximately to the time between 2R, but estimates of the time separating the two events vary between 10 and 100 million years (Lundin et al. 2003). The separation of the speciation events is set arbitrarily to 0.5 S which corresponds to approximately 115 million years i.e. we do not ensure that the speciation times correspond precisely to their estimates in the literature. However, the goal of this work is not to develop a model that can reproduce the exact quantitative features of the empirical distribution of gene family size, but rather to apply the *H. sapiens* parameterisations to get an indication of whether the deviation from the power-law observed in the empirical data is the result of the two rounds of WGD. Given the above and also the raging debate about the large errors in timing major divergence times in the chordate species tree (Graur and Martin 2004; Hedges and Kumar 2004), we do not consider this discrepancy to be of importance.

The simulation produces data with clear peaks at 3, 6, 9, and 12 (see figure 10). Two rounds of WGD with hazard shift should result in an over-representation of families of size 4 and the subsequent two speciations produce three species, thus explaining the peak at 12. The peaks at 3, 6 and 9 are due to the fact that not all genes with a hazard shift will retain all duplicate copies following the WGD: some will lose one duplicate (size 9), some will lose two (size 6), and most will lose three (size 3). This is qualitatively very similar to the empirical data of figure 2 with the only difference being that the data plotted in figure 2 is for five species instead of the three (generated by the two speciations). The fact that the waves are not detectable at larger family sizes in the empirical data is probably due to the fact that larger families, due to their large number of genes have a higher probability of containing a gene that duplicates. This results in a less stable size and thus a lack of conservation of the signal of the WGD.

Discussion

Small- and large-scale gene duplication

The redundancy generated by gene duplication has long been hypothesized to provide the raw material from which new function can evolve (Ohno 1970) and, as such, is of great interest. Small-scale gene duplication is known to occur in many species at a high and relatively constant level through tandem and segmental duplication (Lynch and Conery 2000; Lynch and Conery 2003). Whole genome duplication is also known to occur and has the potential to rapidly and dramatically change the gene content of a genome, but can be difficult to detect if it occurred in the distant past. Probably the two most studied WGDs are those originally hypothesised by Ohno to have occured prior to the fish-tetrapod split (Ohno 1970). There is now strong evidence for two rounds of WGD (2R), possibly in quick succession, prior to the divergence of ray-finned and lobe-finned fish (Wang and Gu 2000; Dehal and Boore 2005) and a ray-finned fish specific WGD (3R) prior to the radiation of teleosts (Christoffels et al. 2004; Vandepoele et al. 2004). However, there has been considerable debate about the number of WGDs in the ancestral chordate with some arguing for none (Friedman and Hughes 2001; Friedman and Hughes 2003), others one (McLysaght et al. 2002) or two (Abi-Rached et al. 2002).

There are two main reasons why ancient large-scale duplication events such as whole genome duplications (WGD) can be difficult to detect. First, the retention rate of duplicates generated through a WGD is often not very high, although it is generally thought to be higher than the retention rate of SSD. This results in a weak signal in the genome, e.g. it is estimated that only approximately 20 percent of duplicates were retained in pufferfish and zebrafish following the fish-specific whole genome duplication (Jaillon et al. 2004; Woods et al. 2005; Brunet et al. 2006). Second, there is a high level of small-scale duplication and loss: genes are constantly subject to a high probability of being duplicated through a small-scale duplication event and the resulting duplicates are themselves subject to a high probability of pseudogenisation (Lynch and Conery 2000; Lynch and Conery 2003; Hughes and Liberles 2007*a*).

The 2R debate

To prove the occurrence of a WGD it is necessary to statistically test whether the null hypothesis, that the data were produced by the background process of SSD, can be rejected in favour of the hypothesis of a large scale duplication. To carry out such a test it is necessary to obtain estimates of the background small-scale duplication and pseudogenisation rates. If the event is hypothesised to have occurred in the relatively recent past, then it can

be possible to tackle this problem as a significant percentage of the duplicates generated by the whole genome duplication are still functional and it is possible to get a relatively accurate estimate of the small-scale duplication and loss process since the hypothesised event (Maere et al. 2005). If, however, the hypothesised event is more ancient, obtaining an accurate estimation of the small-scale duplication and loss process that applied during the period since the hypothesised WGD and formulating a statistical test is impossible.

In the absence of a formal test, advocates and opponents of 2R have studied many features of homologous gene families in order to gather corroborative evidence for their hypotheses, e.g. the number of genes in multigene families, timing of duplications, and genomic location of paralogons. The most recent study, which uses the full genome sequences of *Ciona intestinalis*, *Homo sapiens*, *Mus musculus* and *Takifugu rubripes*, provides all these types of data (Dehal and Boore 2005). First, using the gene annotations, they build gene families such that each family includes all (and only) the descendents of a single gene in the ancestral chordate. They then build phylogenetic trees for these families and infer which nodes in these trees are duplications. Finally, they plot the genomic map positions of the genes that duplicated prior to the fish-tetrapod split. They find that only the genomic location data provides clear evidence of 2R. The data on the number of genes per family for a given species (as well as data on number and timing of duplications) is dismissed as unsupportive, as opponents of the 2R hypothesis had previously done (Hughes et al. 2001).

Although we agree with the conclusion that the ancestral vertebrate is very likely to have undergone two WGDs, we do not agree that there is no sign of these events in gene family size data.

Deviation from the power-law

The motivation for this study was the observation that a clustering of all genes from five tetrapod species produced a distribution of gene family sizes with “waves” with a period of five (see figure 2). This represents a clear deviation from the expected power-law distribution. A study of gene family phylogenetic trees from fully-sequenced vertebrate genomes (Blomme et al. 2006) has shown that a large proportion of duplicated genes in extant vertebrate genomes are ancient and were created at times that coincide with the proposed whole genome duplication events. The same study also established that regulatory genes have a higher probability of being duplicated and retained through WGD than SSD, and it was noted that this is consisent with the dosage balance hypothesis. These findings suggest that WGD has the potential to fundamentally and persistently modify the distribution of gene family size, however, the study did not directly address this issue.

Through several sets of simulations, we have shown that the deviation from the power-law that we observe in the empirical data for sizes less than five can be attributed to the speciation events, but for larger sizes the "waves" are best explained by the two ancient WGDs. Moreover, we have shown that the WGD event must have a specific characteristic, namely that a significant fraction of genes must undergo a hazard shift in WGD as compared to the hazard they are subject to following SSD. Critics might be tempted to suggest that the deviations from the power law are an artifact of the clustering algorithm. This, however, is very unlikely given that the empirical data was clustering using the MCL algorithm (Enright et al. 2002) which is the algorithm used by Ensembl (Birney et al. 2006) to produce gene families, whereas the simulated data was clustered using complete linkage (see "Materials and Methods" section).

These results are of interest for several reasons. First, because it shows that "simple" data on the size of gene families can provide an indication of ancient large scale duplication (the data is simple in the sense that we do not need to build gene families defined with an outgroup, build phylogenetic trees for these families, infer duplications and locate the duplicated genes in the genome). Note, however, that we do not claim that gene family size provides as strong evidence for WGD as the spatio-temporal data of the Dehal and Boore study (Dehal and Boore 2005). Second, it is remarkable that the signal is still detectable given that approximately 500 million years of small-scale duplication and loss separate us from the WGD events. This suggests that there are gene families with retention rates that differ radically between large-scale and small-scale duplication, or more specifically, that there exists a significant fraction of genes with low retention following SSD which are subject to high retention following WGD. Third, it shows that a WGD modifies the structure of the genome's gene content in a profound and persistent way, a finding which is consistent with the observed correlation between WGD and major evolutionary transitions.

Molecular basis of the model of whole genome duplication and retention

A key component of the model needed to produce simulated data that qualitatively matches the empirical data is the WGD with hazard shift. This is consistent with both the dosage balance (Aury et al. 2006) and subfunctionalisation model (Force et al. 1999).

The theory behind dosage balance is that certain categories of protein coding genes, such as proteins forming complexes or enzymes in a metabolic pathway, are very sensitive to the stoichiometry of their interaction partners. If such a gene duplicates through a small-scale duplication, it will have a negative fitness effect by disturbing the stoichiom-

etry and will be selected against. If, on the other hand, it is duplicated in a WGD, it is duplicated with all its interaction partners and the stoichiometry of the interacting partners is unchanged. Thus, the fitness effect of the duplicates is neutral. However, the loss of any of the duplicates will have a negative fitness effect by upsetting the stoichiometry. Under such a model there are two broad types of genes, those that are dosage sensitive and those that are not. For those that are dosage sensitive, the probability of retention depends on whether the duplication occurred through small or large-scale duplication; for those that are not dosage sensitive, the scale of the duplication is immaterial. Moreover, the genes that are dosage sensitive will tend to be those with a low probability of retention following small-scale duplication, but will have a high probability of retention following WGD.

In the subfunctionalisation model, where retention is driven by temporal or spatial partitioning of expression through complementary loss of regulatory regions between a duplicate pair (Force et al. 1999), the probability of retention is an increasing function of the number of regulatory modules. Thus, as long as all regulatory modules are duplicated, the duplicates hazard rate should not depend on whether the duplication was small-scale or whole-genome. It is, however, becoming increasingly evident that a gene's regulatory modules are not necessarily located in the immediate vicinity of the gene's promoter and may even extend into and beyond adjacent transcriptional units (Kikuta et al. 2007). If many regulatory blocks are distant from the genes they regulate, then, under the subfunctionalisation model, such genes would have a lower hazard rate if duplicated in a whole genome duplication event than if duplicated in a small-scale event. In addition, if a gene subfunctionalises following WGD, each duplicate is left with only a subset of the ancestral regulatory regions, thus the probability of retention following a future small-scale duplication is reduced. Therefore, the implementation described above is also relevant for the classical subfunctionalisation model. Another variant of the subfunctionalisation model involves the partitioning of function in the coding sequence causing, for example, the partitioning of interaction partners. Here, as a protein has many interacting partners, this provides opportunities for both subfunctionalization of interactions (proportional to the number of interactions) as well as neofunctionalization of each interacting partner as an independent probability. Once a copy of a duplicated interacting partner neofunctionalizes, the subfunctionalization of the interactions with each partner can then lead to fixation of the duplicates. Through this mechanism, a hazard shift is generated after WGD that is different depending upon the number and nature of physical interactions with other proteins.

From a theoretical point of view, it is difficult to determine whether it is dosage balance or subfunctionalisation that is most likely to cause a hazard shift. Because of

the negative selection against SSD for dosage sensitive genes and the negative selection against loss of duplicates following WGD, dosage balance appears to be a strong candidate. However, this shift may only be transient because, if the negative selection against loss is stochastically overcome in small effective population size vertebrates, cooperative positive selection for rapid gene loss will follow. The hazard shift associated with subfunctionalisation may be of a more permanent nature.

Empirical data also support the hazard shift model: genes involved in transcription regulation are a large functional class which multiple studies have shown is preferentially retained following WGD compared to SSD (Blanc and Wolfe 2004; Maere et al. 2005; Blomme et al. 2006). But, again this can be interpreted either as caused by dosage balance, as regulatory genes are often functional in complexes, or as caused by subfunctionalisation, as they are also often subject to complex regulation involving many enhancer regions.

In our model, the SSD rate is assumed to be a constant and WGD is assumed to duplicate all genes. In both cases, we assume that the duplications reach fixation. This is clearly a simplification as all genes do not duplicate with a given frequency. However, for models of duplication and retention where the initial duplication event is neutral such as subfunctionalisation, this should be an acceptable simplification. In the case of dosage balance, the initial duplication has a negative fitness effect if the duplication was small-scale. This results in a reduced chance of fixation i.e. a lower probability of retention. Our model does not capture this as we model differential retention rates across families exclusively through different pseudogenisation rates. Thus, this feature of dosage balance is modeled through a higher than average hazard rate following SSD for a fraction of the gene families rather than a reduced duplication rate. This approach is justified by the need to build a model that is consistent with multiple modes of gene duplication and retention while restricting the modeling to the genomic level (rather than descending to the population genetic level where we do not have estimates of the key processes).

It is important to emphasise that although the original retention of duplicates following WGD is likely to be driven by subfunctionalisation or dosage balance, there is a strong possibility that the ultimate fate of at least one of the duplicates is neofunctionalisation (Ohno 1970). This is because the retention mechanism, particularly in the case of subfunctionalisation, may reduce the level of pleiotropic constraint that was exerted on the ancestral gene prior to duplication, thus allowing at the very least fine tuning of function (Lynch and Force 2000; Rastogi and Liberles 2005).

Use of Homo sapiens estimates

The key processes in our model are gene duplication, gene loss and sequence divergence. These processes are obviously variable across lineages and time. Ideally, we would have reliable estimates of these processes for the past 500 million years. However, producing estimates of these processes in the distant past is effectively impossible due to the directly counteracting nature of the processes of duplication and loss, and the saturation of silent sites with time. As a second best, we use the estimates for these processes obtained from data on recent SSD duplicates in *Homo sapiens* (Hughes and Liberles 2007*a*). Estimates for other species are available, but these are also for recent duplicates, so we decided to use the high quality human data rather than create a consensus between multiple species. Although, the numerical values of parameters of the equations are different across species the functional forms are the same, thus there is no reason to believe that the functional forms were any different in the past although the parameter values were undoubtedly different.

Again, due to the distant nature of the WGD events, we built a model that only qualitatively matches existing theories for the retention of duplicate genes. We have no basis for the numerical parametrisations of the "hazard shift" model (proportion affected by hazard shift, magnitude of hazard shift, and functional form of the hazard rate following WGD). As a result of this, we are only able to produce simulated data that qualitatively matches the empirical data, as can be seen through the comparison of figures 2 and 10. We could have fine tuned parameters in the model and, thus, obtained a better fit between the simulated and empirical data. For example, by increasing the proportion of families affected by the hazard shift or increasing the size of the hazard shift, the deviation from the power law in the simulated data would have been stronger and quantitatively more similar to the empirical data, but this would have been misleading. Moreover, it is unnecessary as we are not claiming to have produced a precise reconstruction of the evolution of the distribution of gene family sizes in the tetrapod lineage. We aimed only to show that the deviation from the power-law in the empirical data was the result of two rounds of WGD and we consider that our qualitative results show this to be highly probable.

Conclusion

In this study, we have used a model of gene family evolution to produce an approximate characterisation of the effects of whole genome duplication and speciation on the distribution of gene family size. We find that for our simulations to produce the kind of deviation from the power law observed in the empirical distribution of gene family size for several tetrapod species, it is necessary that a significant proportion of genes are subject to a high

probability of retention following WGD and that these genes also have a low probability of retention following small-scale duplication. Whether this difference in probability of retention is the result of the fixation of duplication events being selected against following SSD and loss selected against following WGD (as in the dosage balance model) or whether it is due to a shift in the pseudogenisation rate between SSD and WGD (as in the subfunctionalisation model) is not known. Given that it is difficult to imagine what other type of genomic event would disrupt the distribution in this way and that strong evidence already exists for two WGDs in the ancestral vertebrate approximately 500 Mya, we find it logical to conclude that the pattern, that we observe in the empirical distribution of gene family size for tetrapods, is the result of the ancient WGDs. This implies that WGD may profoundly and persistently modify the distribution of gene family size.

Materials and Methods

Empirical data

Our empirical data consists of the longest protein coding transcript sequence for every gene of the annotated genome of the following species from release 31 of Ensembl (Birney et al. 2006): *Gallus gallus*, *Canis familiaris*, *Mus musculus*, *Rattus norvegicus* and *Homo sapiens*. First, we carry out low complexity masking of the translated sequences using CAST (Promponas et al. 2000) and then perform an all-against-all BLAST (Altschul et al. 1997) (substitution matrix=BLOSUM62, gap opening cost=11, gap extension cost=1). In order, to make the output of the all-against-all BLAST manageable, the BLAST sequence pairs (query and target sequences) are filtered to remove any targets that do not satisfy all of the following criteria that should be satisfied by even very distant homologs: 20% similarity to the query, 60% coverage of the query, e-value $< 10^{-5}$. The e-values of the retained sequences are then used as input to the MCL clustering algorithm with the inflation parameter set to 4.0 (Enright et al. 2002). This procedure produces a clustering of all genes into gene families from which the distribution of gene family size can be computed (see figure 2).

Simulated data

Overview

In order to theoretically investigate the effect of either a speciation event or a whole-genome duplication on the power-law distribution, we need a model of homologous gene family evolution. We use as our starting point the model developed in a previous paper

(Hughes and Liberles 2007*b*). We repeat here a description of this original model and we extend it to incorporate speciation and whole genome duplication events (the reader is referred to the original paper for the rationale behind the model).

Basic model

We model the rate of gene duplication, the rate at which replacement substitutions per replacement site accumulate between genes in a duplicate pair and the rate at which one of the genes in a pair pseudogenises. These models are taken directly from our previous study (Hughes and Liberles 2007*a*) which built on earlier work on the same topic (Lynch and Conery 2000; Lynch and Conery 2003). Time is measured through the accumulation of silent substitutions per silent site (S) between duplicate genes. In *H. sapiens*, under the assumption of a constant rate of small-scale duplication, we have estimated that genes duplicate at a rate of 2.07 per gene per S (all parametrisations used here are the result of fitting the models to duplicate gene pair data from the *H. sapiens* full genome sequence annotation). A duplicate pair i accumulates replacement substitutions per replacement site (R), according to the equation:

$$R_i = \theta_1 S_i + (\theta_2/\theta_3)(1 - \exp(-\theta_3 S_i)) + \varepsilon_i \quad (1)$$

$$Var(\varepsilon_i) = \sigma^2(\tau_1 S_i + \exp(\tau_2(1 - \exp(-\tau_3 S_i)))), E(\varepsilon_i) = 0 \quad (2)$$

where the ε_i are assumed to be independent random variables for i varying from 1 to n. We use the following fitted values of the parameters (Hughes and Liberles 2007*a*): $\theta_1 = 0.13$; $\theta_2 = 0.70$; $\theta_3 = 2.4$; $\sigma^2 = 3.55e-5$; $\tau_1 = 229.4$; $\tau_2 = 6.32$; $\tau_3 = 4.14$.

The probability of pseudogenisation of one of the genes in a pair within Δt given that both genes are still functional at t is:

$$Pr(t < T < t + \Delta t / T > t) = -\frac{Q(t + \Delta t) - Q(t)}{Q(t)} \quad (3)$$

where $Q(t) = Pr(T > t)$ is the survival function: the probability that the time of death, T, is greater than t, i.e. the probability that both genes are still functional at time t. The hazard function $\lambda(t)$ is defined as the event (death/pseudogenisation) rate at time t conditional on survival to time t or later:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T < t + \Delta t / T > t)}{\Delta t} = -Q'(t)/Q(t) \quad (4)$$

We have shown that the Weibull survival function $Q(t) = e^{\rho_1 t^{\rho_2}}$ provides an excellent fit to the data (Hughes and Liberles 2007*a*). Thus, we use this model of the survival function and S as a proxy for time:

$$\lambda(S) = -\rho_1 \rho_2 S^{\rho_2 - 1} \tag{5}$$

In *H. sapiens*, the fitted parameters are $\rho_1 = -4.1$ and $\rho_2 = 0.33$ which implies that the rate of pseudogenisation of a duplicate is a decreasing function of S.

A gene in our model has two key characteristics: it is either functional or pseudogenised, and it has a measure of the number of silent and replacement substitutions per site between itself and all homologous genes i.e. all genes that can be traced to a common ancestor through a series of duplication events. The model is initialised with a set of singleton genes i.e. genes that have no duplicates, and therefore each forms a family of size one. These are the "founding" genes of the homologous gene families. Because all key processes are defined in terms of S, we define a "clock" which "ticks" in increments of 0.001 units of S. At each tick of the clock each gene's number of silent substitutions is incremented by half a tick, so that the distance between all genes increases by one tick. For each gene, we then detect the closest non-pseudogenised homolog which we define as the homologous gene with lowest R distance to the gene of interest. The S distance between the two genes is a measure of the time since the original duplication event and is used to compute the number of replacement substitutions per site the duplicate pair should be subject to in the timeframe of the current tick (equation 1) and the probability that one of the duplicates pseudogenises during the current tick (equation 3) with S as a proxy for time and the Weibull survival function). A gene that has no homologs (such as a founding singleton before it is duplicated) is assigned an S value of 1,000 which ensures that it accumulates R at a very low rate and is subject to a very low probability of pseudogenisation. This is a reasonable way to model singletons as singletons can be expected to have evolved some kind of specialised function that is under selective pressure to be retained in the genome. Finally, each gene is subject to a constant probability of duplication during each tick. The gene that results from a duplication is added to the set of homologous genes. It inherits the R and S distances to other genes from its parent and has a distance of 0 R and 0 S to its parent.

As the model stands, all genes are subject to the same rates of the three main processes. We introduce differences between genes by introducing error terms to the rate of sequence divergence and the rate of pseudogenisation.

Errors for the rate of sequence divergence are available as the fitting of equation 1 produced residuals. Unfortunately, the functional form of the distribution of the error term in equation 1 is not known so, instead, we draw an error term randomly from all residuals from the fitting of equation 1 to the *H. sapiens* gene duplicate data (Hughes and Liberles 2007*a*). This error term can be standardised through the model of the variance as a function of S (see equation 2). A new gene is assigned this error term and this is used in

equation 1 to calculate the number of replacement substitutions per replacement site the duplicate pair should be subject to in a tick (each gene in a pair is subject to half the value predicted by equation 1 when using the gene's specific error term).

In order to be able to accomodate different genes having different rates of pseudogenisation, we modified the definition of the probability of pseudogenisation of a duplicate pair i within a time interval Δt given survival until t, by defining:

$$Pr'(t < T < t + \Delta t / T > t) = (1 + \upsilon_i) Pr(t < T < t + \Delta t / T > t) \quad (6)$$

where $\upsilon_i \sim N(\mu, \sigma^2)$. We want to be able to control the extent to which hazard rates are correlated within families. Thus, when a new gene is created by duplication, the error term υ_i is either inherited from the gene that duplicated, in which case all genes descendent from a founding singleton will have the same hazard function and the heterogeneity between families is determined by σ^2; or a new error can be drawn from the distribution in which case there will be no correlation between the hazard functions of genes descendent from the same singleton.

Model validation and results

In the original paper where we first presented this model, we successfully tested that the simulated evolution of gene duplicates using this model matches the real *H. sapiens* duplicate gene data i.e. that the rates of pseudogenisation and the rate of accumulation of replacement substitutions are the same in the simulated and empirical data.

At regular intervals during the simulation, we extract R for all duplicate pairs and use this data to compute a complete linkage clustering of all non-pseudogenised genes. We use an empirically derived maximum distance of 0.56 R between genes in the same family as the cutoff value in the clustering process (Hughes and Liberles 2007*b*). From this clustering, we can compute the distribution of gene family size.

In our previous paper, we found that the power-law distribution of gene family size failed to emerge if all genes in all families had the same hazard function or if all genes had different hazard functions. The key conclusion was that it is necessary for υ_i to be correlated within a family (inheritance of υ_i) and for there to be sufficient heterogeneity between families ($\upsilon_i \sim N(0, 0.04)$). We found that, under such circumstances, a power-law distribution of gene family size had clearly emerged by $S > 2.0$. This corresponds to approximately 460 million years, given a rate of 2.20 silent substitutions per silent site per billion years for H. sapiens (Yang and Nielsen 1998). Thus, in all our modeling in this paper, we use this configuration of the model which means that, in the absence of WGD, all genes in the same family have the same hazard rate. We run the simulation

until $S = 2.0$, so that a power-law emerges in the distribution of gene family size. We initiate the model with 1,000 singleton families as the number of genes grows rapidly when the genome is subject to WGD and speciation, and the complete linkage clustering is compute-time intensive. Only once the power-law distribution has emerged, do we disrupt it by subjecting the genome to a speciation or WGD event.

Speciation and whole genome duplication extensions

A speciation event is modeled by copying every non-pseudogenised gene (including S and R distances to all homologs) and labeling the genes with the species of the genome in which they exist. This information is then used to limit the search for the closest non-pseudogenised homolog to genes that belong to the same species. This ensures that both genomes evolve independently after speciation as only genes from the same species (and not orthologs) play a role when computing the rate of sequence divergence or the probability of pseudogenisation, but orthologs do get clustered together when homologous gene families are built.

A whole genome duplication is modeled by duplicating all genes of a specific species giving them a distance of 0 R and 0 S to their parent as in small-scale duplication. Due to the one-off nature of WGD, we do not have a quantitative characterisation of the rate of sequence divergence and pseudogenisation as a function of time since WGD. We are thus forced to define these rates as best we can. We leave the sequence divergence rate the same as for small-scale duplicates as this process does not appear to play a crucial role in dynamics of the power-law distribution. For the hazard rate, we implement two options.

The first option is to simply consider that the duplication event to which each gene is subject in the WGD is the same as a SSD, i.e. that each gene inherits the pseudogenisation error term from its parent. However, there is evidence that the retention rate is higher following many WGD, e.g. following the fish-specific WGD (Woods et al. 2005). The most prominent models that have been put forward to explain this higher retention rate of gene duplicates that arise through WGD are dosage balance (Aury et al. 2006) and subfunctionalisation (Force et al. 1999). Although the models are fundamentally very different, they both share that some genes are more highly retained following WGD and that these genes maybe those that stand a below average chance of being retained following SSD.

To incorporate such features in our model, we divide the families into two categories; those that have a high hazard rate error (defined as $\upsilon_i > 0.1$) following small-scale duplication and those with a low hazard rate (defined as $\upsilon_i < 0.1$). We choose this definition as it makes approximately one third of the genes dosage sensitive and thus subject to a hazard shift. We have no data on the proportion of genes that may be subject to such a hazard shift but, given that, following the fish specific WGD, it has been estimated that 20

percent of duplicates were retained (Jaillon et al. 2004; Woods et al. 2005; Brunet et al. 2006) whereas retention rates following SSD are only a few percent (Hughes and Liberles 2007*a*), one third is not an unreasonably large fraction. When a WGD occurs, genes with a high hazard rate error are duplicated and both the original gene and the duplicate are given a hazard rate error (υ_i) drawn from $N(-0.85, 0.0025)$ which ensures a very high probability of retention. We refer to this as hazard shift. Genes with a low SSD hazard rate are duplicated and inherit the error of the duplicated gene as in SSD. When a hazard shifted gene is subsequently duplicated in a small-scale duplication event, the duplicate does not inherit the hazard shift, instead the υ_i shared by all genes in the family prior to the WGD is restored.

Acknowledgements

The work has been funded by FUGE, the functional genomics platform of the Norwegian research council.

References

Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H (2002) Evidence of en bloc duplication in vertebrate genomes. Nat Genet 31:100–105.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, Arnaiz O, Billaut A, Beisson J, Blanc I, Bouhouche K, Câmara F, Duharcourt S, Guigo R, Gogendeau D, Katinka M, Keller AM, Kissmehl R, Klotz C, Koll F, Mouël AL, Lepère G, Malinsky S, Nowacki M, Nowak JK, Plattner H, Poulain J, Ruiz F, Serrano V, Zagulski M, Dessen P, Bétermier M, Weissenbach J, Scarpelli C, Schächter V, Sperling L, Meyer E, Cohen J, Wincker P (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature 444:171–178.

Birney E, Andrews D, Caccamo M, et al. (51 co-authors). (2006) Ensembl 2006. Nucleic Acids Res 34:D556–D561.

Blair JE, Hedges SB (2005) Molecular phylogeny and divergence times of deuterostome animals. Mol Biol Evol 22:2275–2284.

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16:1679–1691.

Blomme T, Vandepoele K, Bodt SD, Simillion C, Maere S, van de Peer Y (2006) The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol 7:R43.

Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol Biol Evol 23:1808–1816.

Christoffels A, Koh EGL, Chia JM, Brenner S, Aparicio S, Venkatesh B (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol 21:1146–1151.

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3:e314.

Demuth JP, Bie TD, Stajich JE, Cristianini N, Hahn MW (2006) The evolution of Mammalian gene families. PLoS ONE 1:e85.

Enright AJ, Dongen SV, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584.

Enright AJ, Kunin V, Ouzounis CA (2003) Protein families and TRIBES in genome sequence space. Nucleic Acids Res 31:4632–4638.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.

Friedman R, Hughes AL (2001) Pattern and timing of gene duplication in animal genomes. Genome Res 11:1842–1847.

Friedman R, Hughes AL (2003) The temporal distribution of gene duplication events in a set of highly conserved human gene families. Mol Biol Evol 20:154–161.

Gilad Y, Man O, Pääbo S, Lancet D (2003) Human specific loss of olfactory receptor genes. Proc Natl Acad Sci U S A 100:3324–3327.

Graur D, Martin W (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends Genet 20:80–86.

Harrison PM, Gerstein M (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J Mol Biol 318:1155–1174.

Hedges SB, Kumar S (2004) Precision of molecular time estimates. Trends Genet 20:242–247.

Hughes AL, da Silva J, Friedman R (2001) Ancient genome duplications did not structure the human hox-bearing chromosomes. Genome Res 11:771–780.

Hughes T, Liberles D (2007*a*) The Pattern of Evolution of Smaller-Scale Gene Duplicates in Mammalian Genomes is More Consistent with Neo- than Subfunctionalisation. J Mol Evol 65:574–588.

Hughes T, Liberles DA. The power-law distribution of gene family size is driven by the pseudogenisation rate's heterogeneity between gene families. submitted.

Huynen MA, van Nimwegen E (1998) The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 15:583–589.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biémont C, Skalli Z, Cattolico L, Poulain J, Berardinis VD, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigó R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quétier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Crollius HR (2004) Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431:946–957.

Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, Ghislain J, Pezeron G, Mourrain P, Ellingsen S, Oates AC, Thisse C, Thisse B, Foucher I, Adolf B, Geling A, Lenhard B, Becker TS (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res 17:545–555.

Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol 1:127–136.

Lundin LG, Larhammar D, Hallböök F (2003) Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. J Struct Funct Genomics 3:53–63.

Luscombe NM, Qian J, Zhang Z, Johnson T, Gerstein M (2002) The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome Biol 3.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.

Lynch M, Conery JS (2003) The evolutionary demography of duplicate genes. J Struct Funct Genomics 3:35–44.

Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473.

Maere S, Bodt SD, Raes J, Casneuf T, Montagu MV, Kuiper M, de Peer YV (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A 102:5454–5459.

McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. Nat Genet 31:200–204.

Ohno S (1970) Evolution by gene duplication. New York: Springer-Verlag.

Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C, Ouzounis CA (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. Bioinformatics 16:915–922.

Rastogi S, Liberles DA (2005) Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol 5:28.

Rastogi S, Reuter N, Liberles DA (2006) Evaluation of models for the evolution of protein sequences and functions under structural constraint. Biophys Chem 124:134–144.

Rivera MC, Jain R, Moore JE, Lake JA (1998) Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci U S A 95:6239–6244.

Roth C, Betts MJ, Steffansson P, Saelensminde G, Liberles DA (2005) The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics. Nucleic Acids Res 33:D495–D497.

Vandepoele K, Vos WD, Taylor JS, Meyer A, de Peer YV (2004) Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci U S A 101:1638–1643.

Wang Y, Gu X (2000) Evolutionary patterns of gene families generated in the early stage of vertebrates. J Mol Evol 51:88–96.

Woods IG, Wilson C, Friedlander B, Chang P, Reyes DK, Nix R, Kelly PD, Chu F, Postlethwait JH, Talbot WS (2005) The zebrafish gene map defines ancestral vertebrate chromosomes. Genome Res 15:1307–1314.

Yanai I, Camacho CJ, DeLisi C (2000) Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Phys Rev Lett 85:2641–2644.

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409–418.

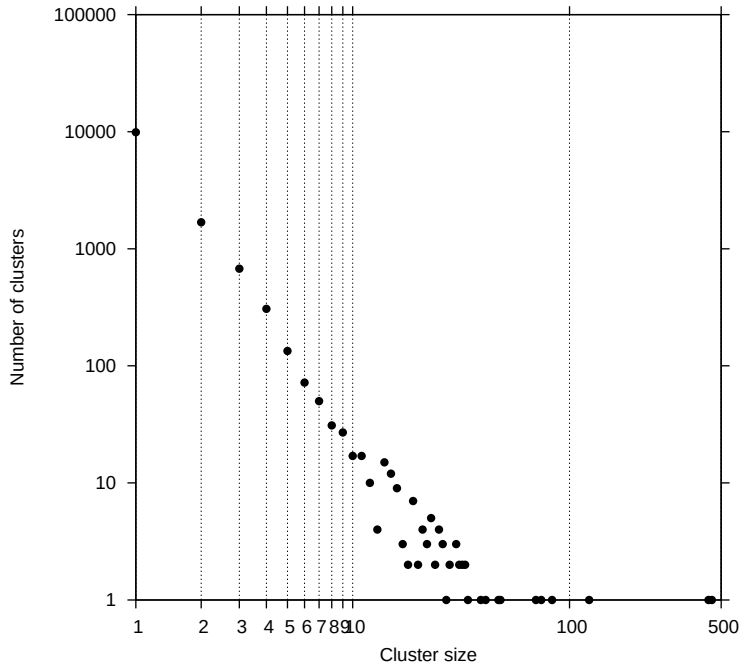

Figure 1: Distribution of gene family size for *H. sapiens (*from a clustering of all putative genes from the fully sequenced genome*)*

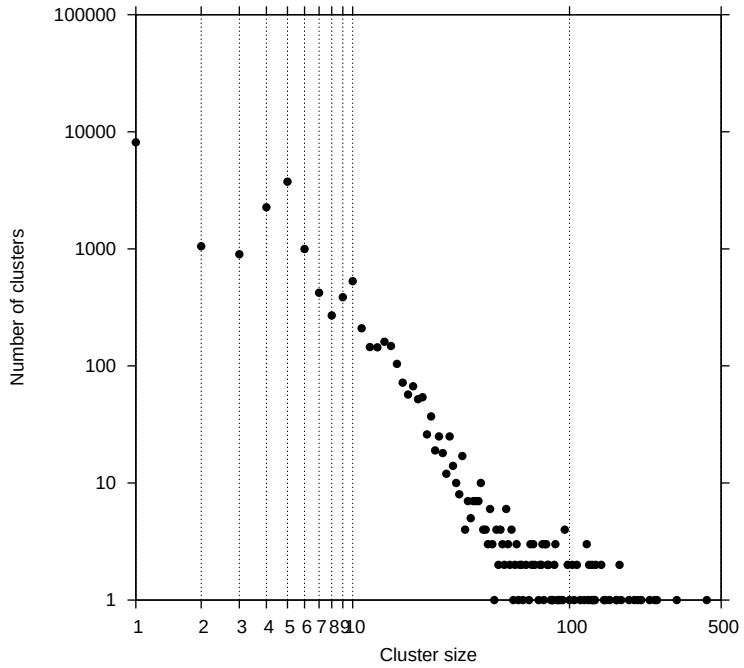

Figure 2: Distribution of gene family size from five tetrapod species (from a clustering of all putative genes from the fully sequenced genome*s of G. gallus, C. familiaris, M. musculus, R. norvegicus and H. sapiens)*

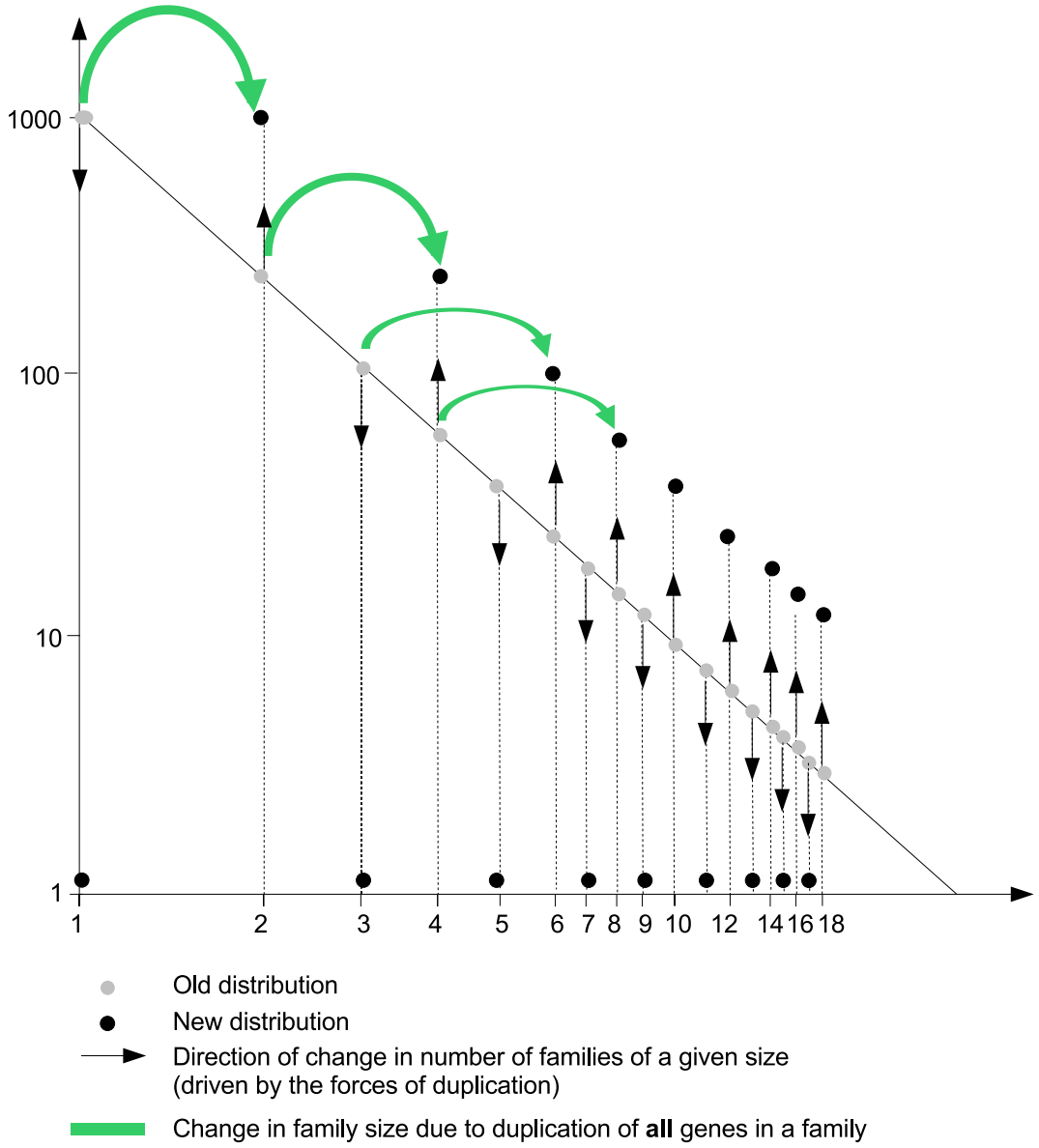

Figure 3: Qualitative description of the immediate effect of a WGD on a power-law distribution of gene family size

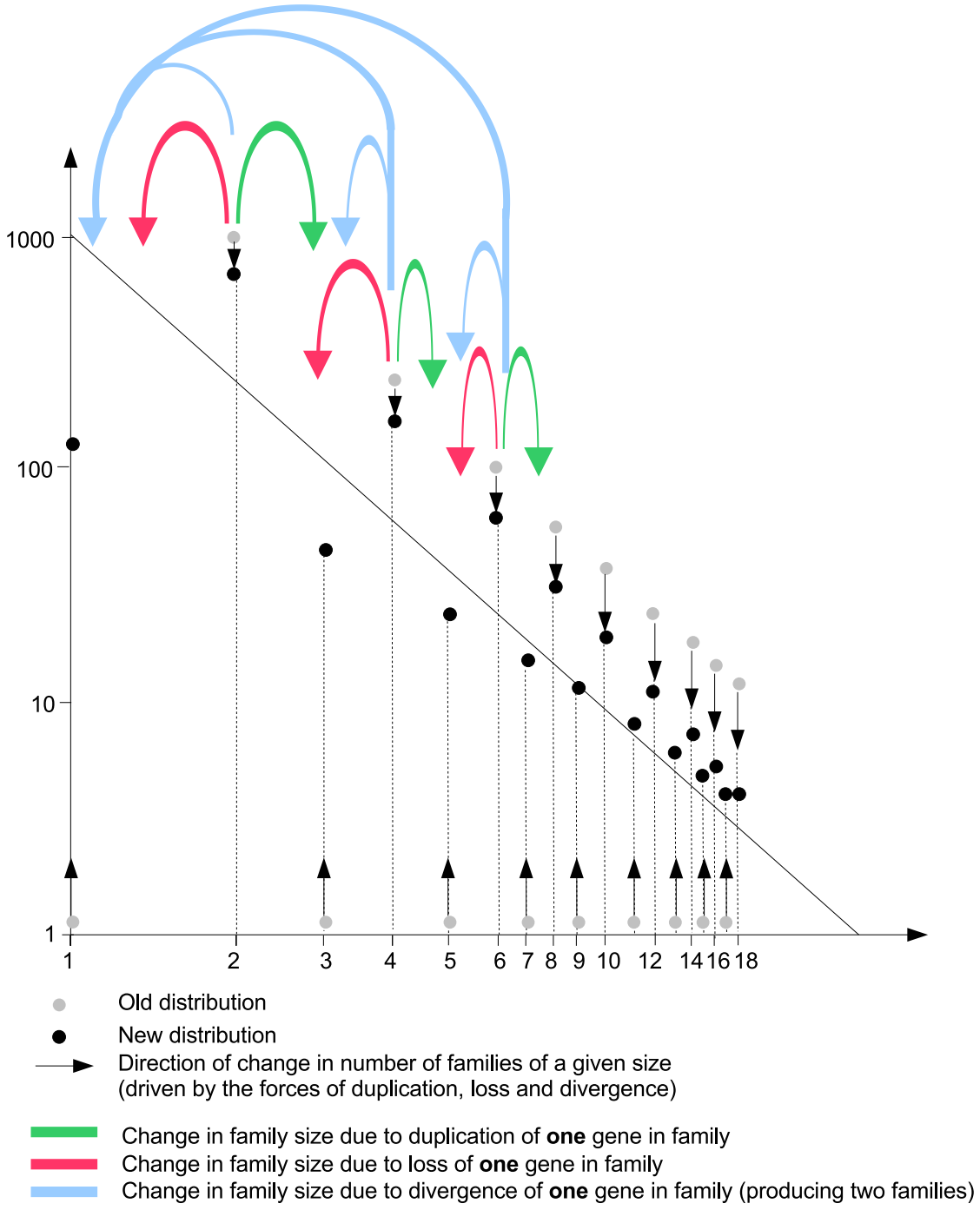

Figure 4: Qualitative description of the gene family size distribution sometime after the WGD

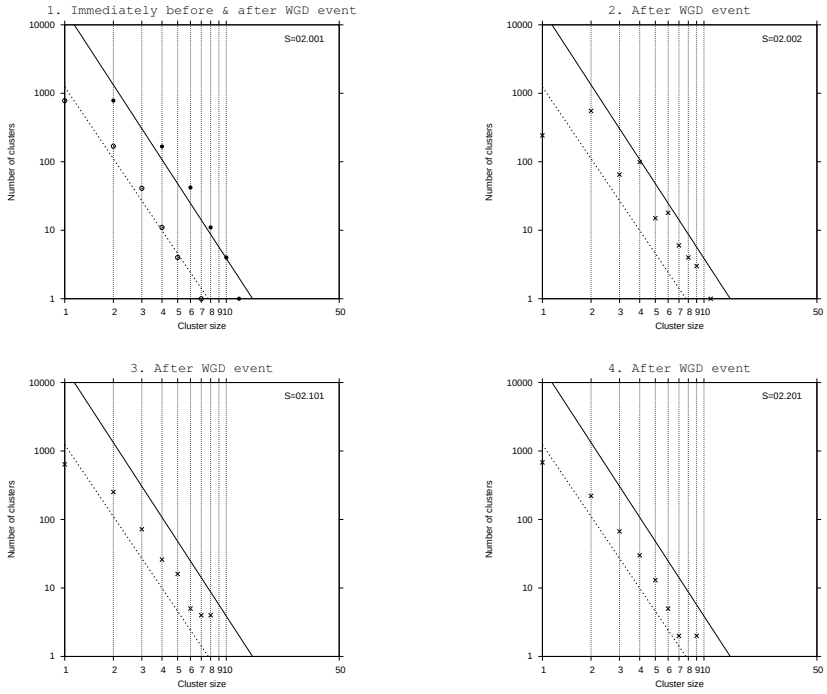

Figure 5: WGD *without hazard shift* at $S = 2.0$ and subsequent change in the distribution of gene family size

White dots and dotted line: data immediately *prior* to the WGD and fitted power-law equation
Black dots and solid line: data immediately *after* the WGD and fitted power-law equation
Crosses: data at later time points (higher S)

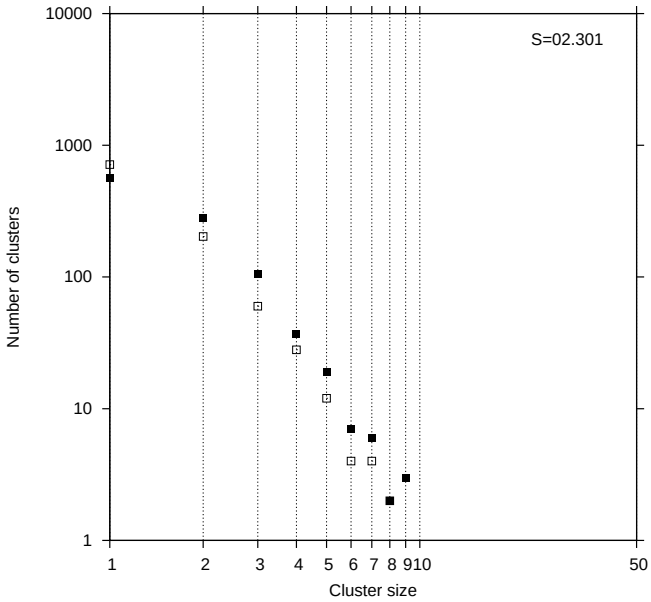

Figure 6: Comparison of distributions following WGD at $S = 2.0$ *with hazard shift* (black squares) and *without hazard shift* (white squares), represented at $S = 2.3$

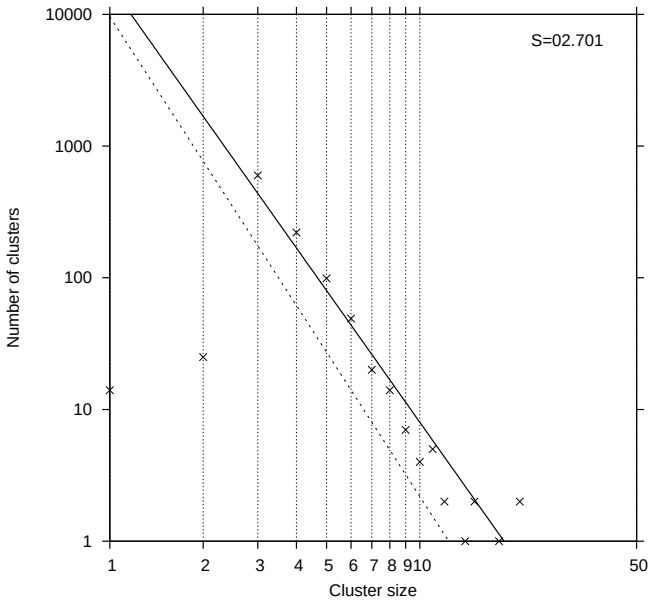

Figure 7: Distribution following a first speciation at $S = 2.0$ and a second speciation at $S = 2.5$, represented at $S = 2.7$

Dotted line: equation fitted to the distribution data immediately *prior* to the second speciation
Solid line: equation fitted to the distribution data immediately *after* the second speciation
Crosses: data at $S = 2.7$

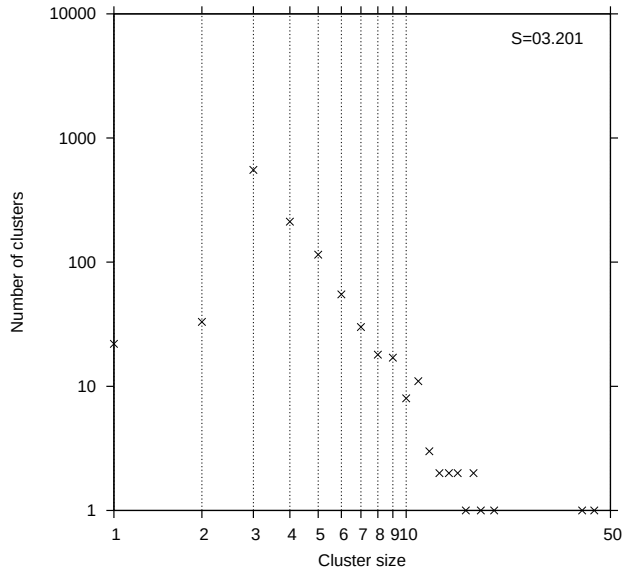

Figure 8: WGD *without hazard shift* followed by two speciations, represented at $S = 3.2$ (WGD at $S = 2.0$, first speciation at $S = 2.5$, and second speciation at $S = 3.0$)

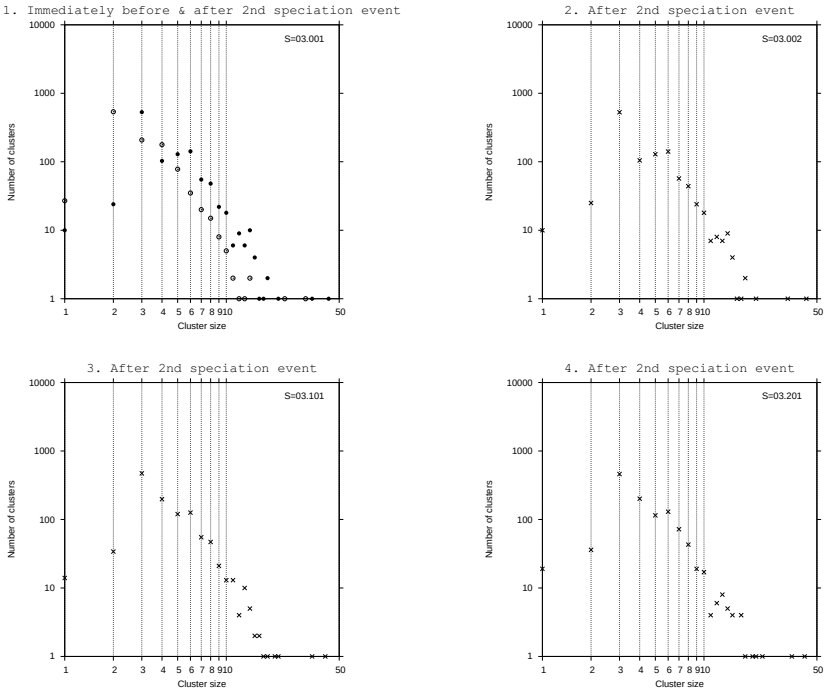

Figure 9: WGD *with hazard shift* followed by two speciations, represented at $S = 3.0$ (WGD at $S = 2.0$, first speciation at $S = 2.5$, and second speciation at $S = 3.0$)

White dots: data immediately *prior* to the second speciation
Black dots: data immediately *after* the second speciation
Crosses: data at later time points (higher S)

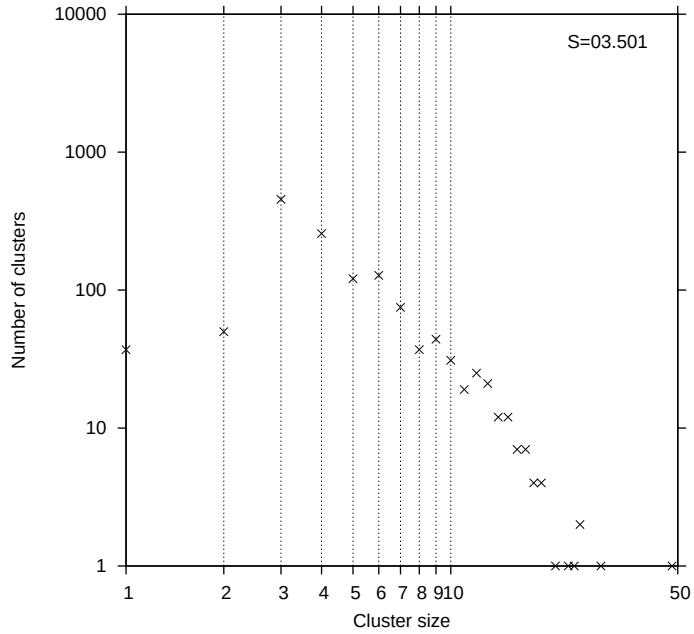

Figure 10: Two WGDs *with hazard shift* followed by two speciations, represented at $S = 3.5$
(WGDs at $S = 2.0$ and 2.2, speciations at $S = 2.7$ and 3.2)

Prior to 2nd speciation						After 2nd speciation
Family size	Frequency (%)		Species composition	Frequency (%)		Family size
	No hazard shift	Hazard shift		No hazard shift	Hazard shift	
1	3.2	2.4	A	35.3	37.0	1
			B	64.7	63.0	2
2	59.3	48.0	AA	0.6	0.9	2
			AB	98.6	98.1	3
			BB	0.8	0.9	**4**
3	20.3	18.6	AAA	0.5	1.0	3
			AAB	47.7	46.2	**4**
			ABB	51.4	51.4	**5**
			BBB	0.5	1.4	**6**
4	6.9	15.9	AAAA	0.0	0.0	**4**
			AAAB	36.5	13.5	**5**
			AABB	36.5	72.5	**6**
			ABBB	27.0	14.0	7
			BBBB	0.0	0.0	8

Table 1: Gene family size distribution for the genes of species A and B which arose at $S = 2.5$ immediately prior to a second speciation of B at $S = 3.0$. Prior to the first speciation, the common ancestor of A and B was subjected to a WGD.

Supplementary materials

Timothy Hughes and David A. Liberles

TH: Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway.
Telephone: (+47) 55 58 40 72. Email: tim@bccs.uib.no

DAL: Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA.
Telephone: (+1) 307 766 5206. Email: liberles@uwyo.edu.

Correspondence: TH and DAL

Keywords: gene duplication, whole genome duplication, pseudogenisation, non-synonymous substitution, gene family size, power-law distribution, speciation.

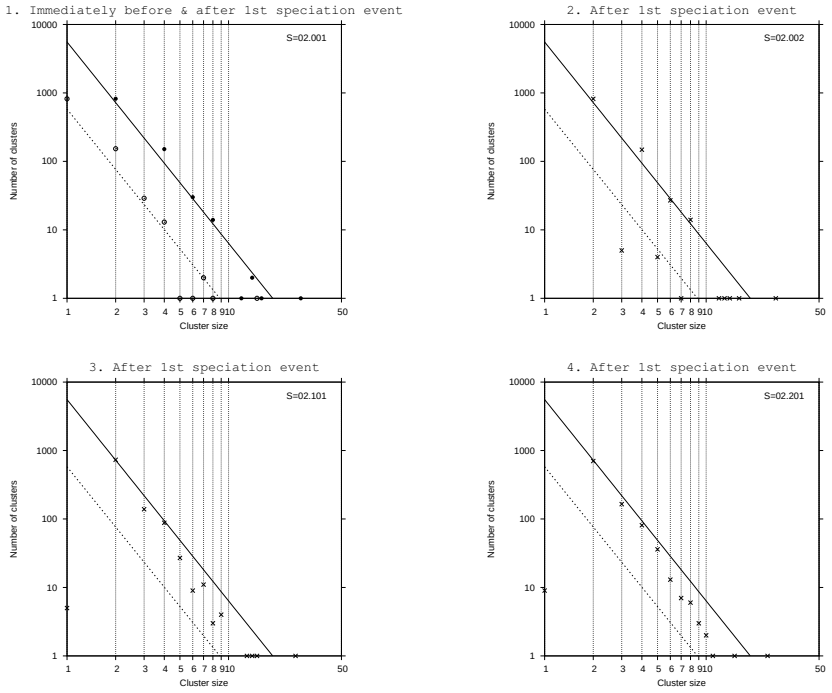

Figure 1: First speciation at $S = 2.0$

White dots and dotted line: data immediately *prior* to the speciation and fitted power-law equation
Black dots and solid line: data immediately *after* the speciation and fitted power-law equation
Crosses: data at later time points (higher S)

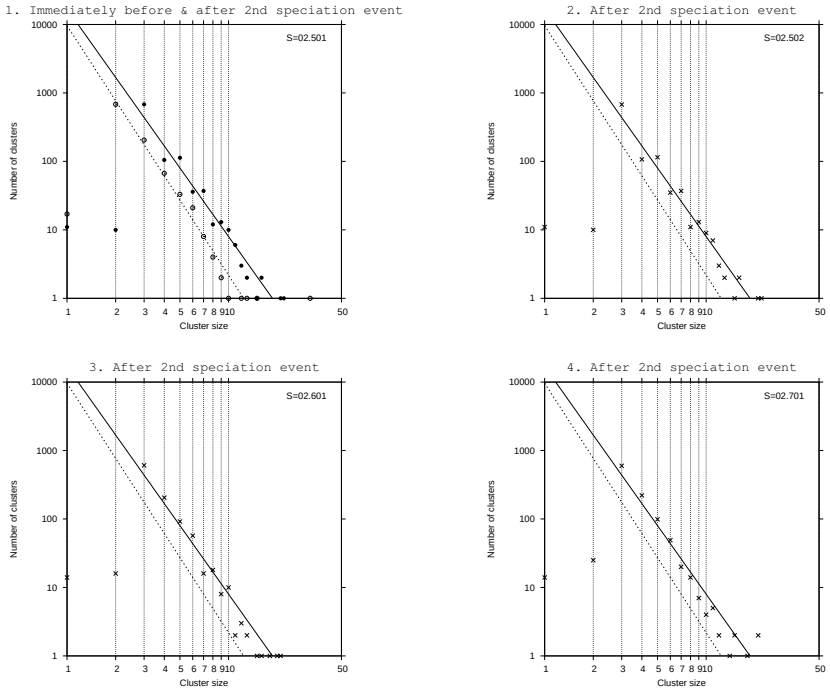

Figure 2: Second speciation event at $S = 2.5$

White dots and dotted line: data immediately *prior* to the speciation and fitted power-law equation
Black dots and solid line: data immediately *after* the speciation and fitted power-law equation
Crosses: data at later time points (higher S)

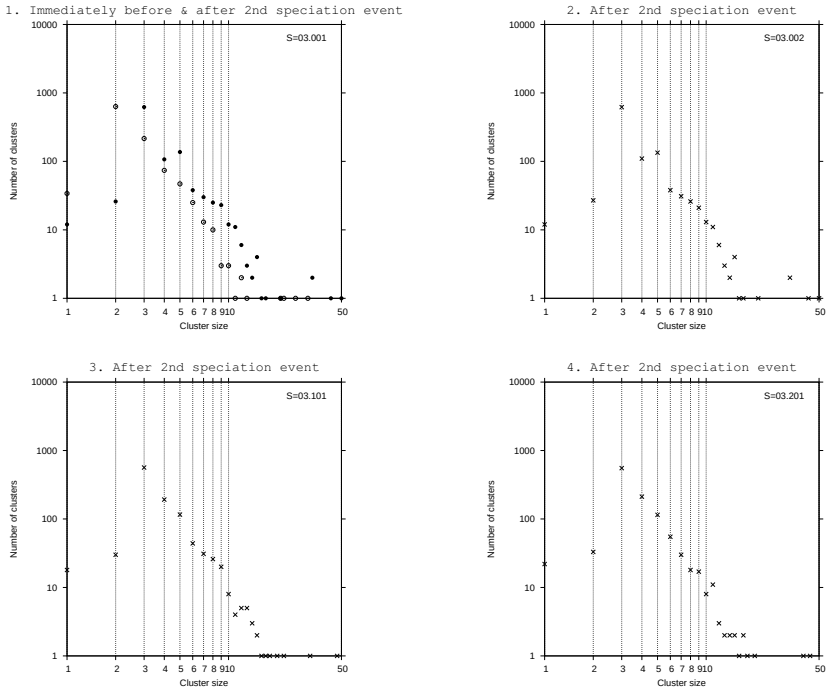

Figure 3: WGD *without hazard shift* followed by two speciations
WGD at $S = 2.0$, first speciation at $S = 2.5$, and second speciation at $S = 3.0$

White dots: data *prior* to the speciation
Black dots: data immediately *after* the speciation
Crosses: data at later time points (higher S)

D. Paper IV

Minireview

Evaluating dosage compensation as a cause of duplicate gene retention in *Paramecium tetraurelia*

Timothy Hughes*, Diana Ekman†, Himanshu Ardawatia*‡, Arne Elofsson† and David A Liberles‡

Addresses: *Computational Biology Unit, Bergen Center for Computational Science, University of Bergen, 5020 Bergen, Norway. †Department of Biochemistry and Biophysics, Stockholm University, 10691 Stockholm, Sweden. ‡Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA.

Correspondence: David A Liberles. Email: liberles@uwyo.edu

Published: 22 May 2007

Genome ***Biology*** 2007, **8:**213 (doi:10.1186/gb-2007-8-5-213)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/5/213

Abstract

The high retention of duplicate genes in the genome of *Paramecium tetraurelia* has led to the hypothesis that most of the retained genes have persisted because of constraints due to gene dosage. This and other possible mechanisms are discussed in the light of expectations from population genetics and systems biology.

Many genomes display extensive gene duplication, which may result from either small-scale duplications or from duplication of the whole genome. What determines whether both copies of a duplicate gene are retained in the genome, and their subsequent evolutionary fate, is still a matter of debate. Aury *et al*. [1] have recently characterized gene duplication in the ciliate *Paramecium tetraurelia*, a unicellular eukaryote, which appears to have undergone multiple rounds of whole-genome duplication with a high level of retention of the duplicate copies. They suggest that this high level of retention is due to constraints arising from gene dosage, rather than other proposed mechanisms. Here we discuss these results in relation to the various models proposed for gene duplication and retention.

When duplication of a gene, or genome, occurs in an individual organism, it will only become part of the species genome if it becomes 'fixed' in the population (that is, becomes part of the genome of all members of the population). If the initial duplication event is evolutionarily neutral, the duplicated genes will become fixed in the population with a probability dependent on the inverse of the effective population size. It has been suggested, however, that the initial duplication event is likely to be deleterious for gene duplicates with functional regulatory regions, because of the metabolic cost of producing extra protein [2]. This would reduce the probability of fixation.

Given that fixation probably occurs much more quickly than the resolution of the fates of the duplicate copies, most work has considered fate determination as an independent step that occurs after the random process of fixation. Once fixation occurs, if there is purely neutral evolution at the protein level, one copy of a duplicated gene will quickly become a pseudogene, leaving a single ancestral copy with an ancestral function. While relaxation of selective constraint is generally thought to occur after gene duplication, negative selection, which discards changes, apparently returns quickly. Negative selection on parts of the gene may also be coupled to positive selection for the evolution of new functions or levels of expression. Relaxation of selective constraint (or a combination of negative and positive selection) that quickly gives way to stronger negative selection has been observed both in *Paramecium* [1] and in computer simulations of the evolution of gene duplicates [3].

Models that aim to explain the retention of duplicated genes include the subdivision of expression profiles or functions of the ancestral gene between the duplicates (subfunctionalization) [4]; the acquisition of new functions by one or both

duplicated copies (neofunctionalization) [5]; selection to increase robustness by maintaining a highly conserved back-up copy [6]; and selection for increased gene dosage or for dosage-compensation effects, as suggested for *Paramecium* (see also [7]).

Selection that depends on gene dosage can involve two different mechanisms. Selection for increased gene dosage involves a positive selection pressure to increase expression from a locus that is already highly expressed and has little mutational capacity to increase its expression or concentration-dependent activity. The dosage-compensation model, on the other hand, invokes a negative selection pressure to retain the function and expression levels of both copies in order to preserve the correct stoichiometry - the appropriate amounts or activity of the proteins in relation to each other or other proteins. Subfunctionalization is a nearly neutral model, with neither positive nor negative selection on gene function during the initial period of preservation, whereas neofunctionalization involves positive selection for the generation of new functions in the retained genes. Selection for redundancy, like that for dosage compensation, is characterized by negative selection. Several of these processes can act at different levels of biological regulation: for example, neofunctionalization and subfunctionalization can occur through changes in protein expression, changes in protein function, or changes in alternative or constitutive splicing. Dosage compensation, on the other hand, is a model in which conservation acts simultaneously on all of these processes.

Genome duplication favors the retention of duplicate genes

From examination of a variety of genomes, tandem and segmental gene duplications are known to occur at very high rates (on average 0.01 per gene per million years), similar in magnitude to the rate of mutation per nucleotide site [8,9]. Following such duplications, the average half-life of a gene copy is of the order of a few million years, with only a small fraction of duplicates surviving beyond a few tens of millions of years (TH and DAL, unpublished observations). Following whole-genome duplication, on the other hand, a large proportion of duplicate genes is retained after tens of millions of years (as in *Xenopus laevis* [10]) or even hundreds of millions of years (in teleost fish [11]). For teleost fish, the rate of retention has been reported to be much higher for the products of whole-genome duplication than for those of small-scale duplication [11].

One possible explanation for these differences is that gene fate is shaped by different evolutionary forces, depending on whether a gene is duplicated in a whole-genome event or not. In a whole-genome duplication, unlike a smaller-scale duplication, the entire network of interacting partners is duplicated together (Figure 1). It is unclear to what degree this build-up of pleiotropic constraints is a limitation as duplicates diverge, and this question needs to be addressed, potentially using protein structural models. The dosage-compensation model would predict that the build-up of pleiotropic constraint is difficult to resolve without deleterious effects, thus introducing a strong negative selection initially against the loss of genes or interactions. This would lead to gene retention and initial conservation of sequence and expression after whole-genome duplication.

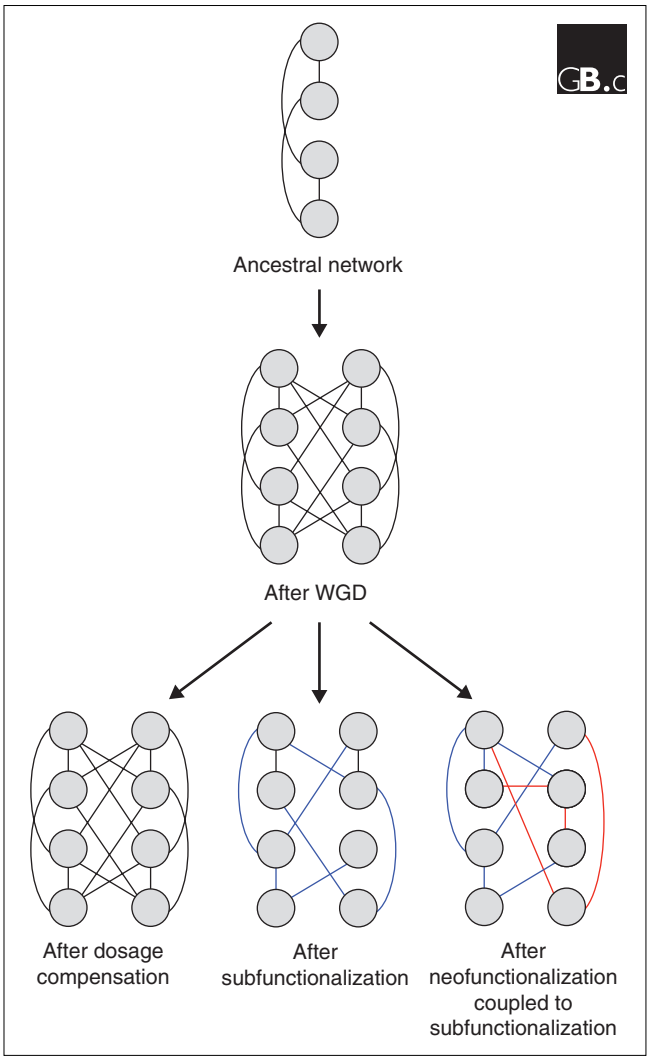

Figure 1
Possible outcomes for gene retention after whole-genome duplication. An ancestral network of interacting proteins is shown. Following a whole-genome duplication event, all of the proteins together with their interactions are duplicated. Over time, depending upon the evolutionary forces that are operating on the genome, different interactions are retained, gained or lost. Under the dosage-compensation model (bottom left), all interactions are retained. Under the subfunctionalization model (bottom center), redundant interactions become nonredundant (blue). When this is combined with the neofunctionalization model (bottom right), new interactions are also gained (red). In this figure, all of the duplicated copies have been retained as functional genes, but that is not the most likely outcome with increasing evolutionary time.

Gene duplication in the *Paramecium* genome

With the sequencing of the genome of *P. tetraurelia* by Aury *et al.* [1], it was found to contain 39,642 genes, more genes than many other completely sequenced genomes. Furthermore, these genes can be grouped into families whose members are very closely related in sequence. Phylogenetic analysis of these gene families points to a recent whole-genome duplication in *P. tetraurelia*, in addition to several older genome duplications. The most recent duplication occurred long enough ago for negative selection to have set in, however.

Aury *et al.* [1] find that duplicate genes for signaling proteins and transcription factors are preferentially retained in the genome, as are duplicated genes for proteins known to form multicomponent complexes, with a positive correlation between retention and the number of components in the complex. A similar correlation between retention and complexity was observed for genes involved in metabolic pathways. More highly expressed genes were also more likely to have been retained.

Interestingly, the co-retained duplicates did not always originate from the same whole-genome duplication. In regard to complex-forming proteins, genes that were co-retained after the most recent whole-genome duplication were not found to be those preferentially retained in the older duplications. In all, Aury *et al.* [1] found that patterns of retention across whole-genome duplications were affected by gene function, and showed a preference for retention of duplicated genes that had not retained a duplicate in an older whole-genome duplication.

The authors conclude that dosage compensation to maintain the stoichiometry of protein complexes and metabolic pathways and keep them functioning correctly plays an important part in the retention of duplicate genes after a whole-genome duplication. From consideration of the traces of the preceding whole-genome duplications they also propose that over time there is a slow progressive loss of duplicates, as gene-expression levels become adapted for stoichiometric reasons, for example.

The dosage-compensation model predicts that duplicates of genes for proteins that do not form complexes or do not have concentration-dependent roles in metabolism will be rapidly lost. In the case of duplicated genes encoding interacting proteins, it predicts strong selection for retention, but if one of the interacting duplicates is lost from the genome, the model predicts that the loss of the remaining duplicate will now be positively selected for. The first part of this prediction is qualitatively satisfied by the observations from the *P. tetraurelia* genome of the retention of genes for complex-forming proteins. On the other hand, the retention patterns and differing profiles of nonsynonymous (K_a) and synonymous (K_s) substitutions (K_a/K_s profiles) for duplicates of different ages do not seem to support dosage compensation as the driving force for keeping them in the genome. Selection as a result of dosage compensation thus appears to be complex and may have a role in modulating other evolutionary mechanisms. The apparent burst of either positive selection or relaxation of selective constraint in the period shortly after genome duplication implies that selective mechanisms other than dosage compensation are also acting.

Following the most recent whole-genome duplication in *P. tetraurelia*, species radiation occurred, resulting in the *P. tetraurelia* complex of 15 sibling species. Aury *et al.* [1] propose that this burst of speciation is a side-effect of the whole-genome duplication, occurring as a result of differential gene loss in different populations, leading to inviable hybrids and reproductive isolation by Dobzhansky-Muller incompatibility [12]. Such a proposition is consistent with the loss of proteins not under dosage-balance constraint under the dosage-compensation model and in our opinion is most consistent with speciation accompanied by neofunctionalization or subfunctionalization.

In evaluating alternative explanations of the retention profiles for duplicates in the paramecium genome, effective population size may be an important consideration. Effective population size (together with mutation rate) as a modulator of the strength of selection has been implicated as an important switch between subfunctionalization as a purely neutral process and neofunctionalization or, potentially, dosage compensation as mechanisms involving selection [4,8,9]. *Paramecium* has been shown to have a relatively large effective population size, making mechanisms that involve selection possible [13]. However, it has been shown that binding interactions as well as regulatory modules can subfunctionalize in the preservation of duplicate genes [3,14], and so the subfunctionalization model for gene duplicate retention may also be consistent with a dependence on the number of interacting protein partners, where the probability of subfunctionalization might be expected to be proportional to the number of ways of subfunctionalizing the interactions with partners. This is a different mechanism of gene retention from dosage compensation, but this characteristic of subfunctionalization has not been evaluated to show that it has the same potential to retain duplicate genes in such high numbers as dosage compensation appears to be able to do. Eventually, quantitative models characterizing these various processes can be tested against the data to extend our understanding of the process of gene retention.

Where does dosage compensation fit in?

Dosage compensation may indeed affect the short-term retention rate of duplicate genes after whole-genome duplication. Over longer time frames, however, proteins involved in complexes and pathways are not preferentially retained in the duplicate pairs originating from whole-genome duplications, neither in *P. tetraurelia*, as indicated

by Aury *et al.* [1], nor in yeast [15] (except for ribosomal proteins [16]). In fact, whereas 17% of highly connected proteins (hubs) in the yeast protein-protein interaction network belong to a pair originating from the relatively ancient whole-genome duplication that has occurred in *Saccharomyces cerevisiae*, only 5% of the party hubs, which are coexpressed with their interaction partners, are part of such a pair [15]. Homologous complexes in yeast appear to have been created through stepwise partial duplications and not through whole-genome duplication [17].

The results of Aury *et al.* [1] do suggest that after more recent whole-genome duplication events, the duplicate proteins belonging to complexes and pathways are initially retained to a greater extent than other proteins. According to this view, although dosage sensitivity is not sufficient for the long-term fixation of duplicates in the genome, it may be important in the first phase following the whole-genome duplication. One might postulate dosage compensation as a mechanism for holding duplicated genes in the genome for some time, to give an opportunity for eventual neofunctionalization (as has been suggested for subfunctionalization [3]). However, even in the period immediately following duplication, stoichiometric issues will be dependent on the interplay between expression and sequence as well as selective pressures for concentration dictated by metabolism and systems-level constraints. Further modeling work is needed to understand the mechanism, as the suggestions by Aury *et al.* [1] and alternative suggestions (such as subfunctionalization of binding interactions) are part of an ongoing synthesis to understand the process of gene duplication and its relationship to the evolution of gene function.

Considering the case of metabolic networks, the patterns of retention or modification have been observed to be influenced by network structure, topology and function, and the positioning of duplicate genes at key points in the network. Genes coding for enzymes involved in directing higher metabolic fluxes are subject to greater evolutionary constraints as a gene duplication event would increase the flux through an enzyme-catalyzed reaction. It has been observed in *S. cerevisiae* that genes encoding highly connected enzymes in metabolic pathways have a higher likelihood of maintaining duplicates [18]. Thus, duplication of genes encoding enzymes carrying high metabolic fluxes are more likely to be retained compared to genes encoding enzymes carrying lower metabolic fluxes.

Enzymes in a pathway can evolve with different functional requirements, which can lead to mismatches in the enzyme activities upon duplication [19]. This means that upregulation of individual enzymes can increase or decrease the flux capacity of the pathway and by different amounts. Hence, if only certain proteins increase the performance of the pathway, the duplicates of the other proteins in the pathway will not provide extra fitness to the organism. This also has implications for the retention of duplicate copies based upon an entire pathway being duplicated, indicating that the negative selective pressure for retention of each duplicate in a pathway would not be equally strong. Interestingly, it has been argued that the neutral expectation for biological networks involves a more complex network than that minimally required for function, without necessarily invoking robustness as a driving force for this non-minimal network [20].

The findings by Aury *et al.* [1] lend further support to the idea that dosage compensation can play a role in the retention of duplicated genes in a genome. Whole-genome duplication events in additional lineages representing different time points will enable a fuller testing of this and other hypotheses, as well as their functional implications for systems biology.

References

1. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, *et al.*: **Global trends of whole genome duplications revealed by the ciliate *Paramecium tetraurelia*.** *Nature* 2006, **444:**171-178.
2. Wagner A: **Energy constraints on the evolution of gene expression.** *Mol Biol Evol* 2005, **22**:1365-1374.
3. Rastogi S, Liberles DA: **Subfunctionalization of duplicated genes as a transition state to neofunctionalization.** *BMC Evol Biol* 2005, **5:**28.
4. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations.** *Genetics* 1999, **151:**1531-1545.
5. Ohno S: *Evolution by Gene Duplication.* New York: Springer-Verlag, 1970.
6. Kuepfer L, Sauer U, Blank LM: **Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*.** *Genome Res* 2005, **15:**1421-1430.
7. Withers M, Wernisch L, dos Reis M: **Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome.** *RNA* 2006, **12:**933-942.
8. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290:**1151- 1155.
9. Lynch M, Conery JS: **The origins of genome complexity.** *Science* 2003, **302:**1401-1404.
10. Hughes MK, Hughes AL: **Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*.** *Mol Biol Evol* 1993, **10:**1360-1369.
11. Blomme T, Vandepoele K, de Bodt S, Simillion C, Maere S, van de Peer Y: **The gain and loss of genes during 600 million years of vertebrate evolution.** *Genome Biol* 2006, **7:**R43.
12. Orr HA: **Dobzhansky, Bateson, and the genetics of speciation.** *Genetics* 1996, **144:**1331-1335.
13. Snoke MS, Berendonk TU, Barth D, Lynch M: **Large global effective population sizes in *Paramecium*.** *Mol Biol Evol* 2006, **23:** 2474-2479.
14. Braun FN, Liberles DA: **Retention of enzyme gene duplicates by subfunctionalization.** *Int J Biol Macromol* 2003, **33:**19-22.
15. Ekman D, Light S, Bjorkman AK, Elofsson A: **What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*?** *Genome Biol* 2006, **7:**R45.
16. Papp B, Pal C, Hurst LD: **Dosage sensitivity and the evolution of gene families in yeast.** *Nature* 2003, **424:**194-197.
17. Pereira-Leal JB, Teichmann SA: **Novel specificities emerge by stepwise duplication of functional modules.** *Genome Res* 2005, **15:**552-559.
18. Vitkup D, Kharchenko P, Wagner A: **Influence of metabolic network structure and function on enzyme evolution.** *Genome Biol* 2006, **7:**R39.
19. Salvador A, Savageau MA: **Evolution of enzymes in a series is driven by dissimilar functional demands.** *Proc Natl Acad Sci USA* 2006, **103:**2226-2231.
20. Soyer OS, Bonhoeffer S: **Evolution of complexity in signaling pathways.** *Proc Natl Acad Sci USA* 2006, **103:**16337-16342.