Towards Relating the Evolution of the Gene Repertoire

in

Mammals to Tissue Specialisation

Shiri Freilich

Wolfson College

This dissertation is submitted to the University of Cambridge

for the degree

of Doctor of Philosophy

21 December 2006

To Leon, who was the wind blowing in my sails, in the deep blue sea of this journey of ours.

This Thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This Thesis does not exceed the specified length limit of 300 pages as defined by the Biology Degree Committee.

This Thesis has been typeset in 12pt font according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Summary: Towards Relating the Evolution of the Gene Repertoire in Mammals to Tissue Specialisation

The sequencing efforts of recent years have provided a rich source of data for investigating how gene content determines similarity and uniqueness in a species' phenotype. Work described in this PhD Thesis attempts to relate innovations in the gene repertoire along the mammalian lineage to the most obvious phenotypic characteristic of animals: the appearance of highly differentiated tissue types. Several different approaches, outlined below, have been followed to address some aspects of this problem.

Initially, a comprehensive study of the pattern of expansion of the complement of enzymes in various species was performed in order to obtain a better view of the principles underlying the expansion of the gene repertoire in mammals. Although several studies have described a tendency toward an increase in sequence redundancy in mammals, not much is known about the way such sequence redundancy reflects functional redundancy. Our analysis indicates that different kingdoms differ in the pattern of expansion of their enzymatic set. Whereas for unicellular species one can observe a strong correlation between the gene content and the enzymatic reaction repertoire, for multicellular species we observe a tendency toward an increased functional redundancy.

Furthermore, the relationship between expansion of gene families in mammals and differentiation of tissue types was studied by analysing mouse gene expression data from many tissues. We show that duplication events lead, on average, to a more specific expression of the duplicate proteins. Though, such tendency is observed only when the duplication event postdated the transition to multicellularity. The combined expression distribution of all members of a protein family implies a complementary expression between family members. Therefore, the analysis provides large-scale, corroborative support for the subfunctionalization theory, which states that the division of expression of an ancestor gene between its daughter duplicates promotes the retention of the duplicates in the genome.

Based on the assumption that the physiological role of a tissue is determined by the unique composition of genes expressed in that tissue, we characterised each tissue with regard to the distribution of function and phyletic origin of the expressed genes. A stronger tendency is observed for pre-metazoan genes (which are predominantly involved in

metabolic functions) to be globally expressed, versus a stronger tendency observed for metazoan-specific genes (which are predominantly involved in regulatory functions) to be specifically expressed.

Finally, the phyletic origin of the mammalian metabolic genes and the relationship between innovations in this metabolic set to the development of new, tissue-specific, pathways was studied.

Overall, the analysis demonstrates that tissue differentiation is at least in part achieved by tissue specialisation of metabolism – either by differentially expressing the ancient metabolic proteins, or by gaining new metabolic proteins participating in tissue specific functions.

Preface

This Thesis is the end result a challenging but also immensely rewarding Ph.D. work at the European Bioinformatics Institute (EBI). For this experience, there are several people to whom I would like to thank.

Firstly, I would like to express my deep gratitude to my supervisor, Prof. Janet Thornton. It was a great privilege working for her both scientifically and personally. I couldn't have wished for a better supervisor and I will miss her guidance. Throughout my research in the Thornton group I was fortunate enough to enjoy a most stimulating and amiable environment. I would like to thank the past and present members of the group for allowing me such an experience.

I would like to thank my family: My husband Leon who had encouraged me to apply, always took interest in my work, and was a true partner and a source of encouragement and support; my father Moshe, my mother Tzipor and my two brothers, Ofer and Ariel, whose love and support were my source of comfort throughout my stay away from home.

My Ph.D. had started in March 2002 and is ending in Summer 2006. The past four and a half years in Cambridge had been wonderful. I loved my research and I have acquired precious friends. During that time my son, Idan, was born.

The conclusion of this Thesis takes place in Israel with a different, sad, tone – the sounds of war. On this note, I would like to wish the good people from both sides of the border, better days, days of peace.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

This thesis describes several approaches taken in order to study how innovations in the gene repertoire in the mammalian genome are related to specialization of tissues. The work was done using sequence analysis and expression data analysis techniques. The introduction section provides a brief preface to a few relevant issues, and summarizes the state of knowledge at the time this Thesis began. Section 1.1 discusses the phylogenesis of mammalian lineage where a special focus is given to the appearance of the multicellular ancestor of animals; Section 1.2 discusses the evolution of the mammalian gene repertoire and describes the main functional innovation in different lineages leading to mammals as inferred from comparative genome studies; Section 1.3 provides a brief review of the ontogenesis process in mammals; Section 1.4 describes the available expression data from mammalian tissues and summarises the information from comparative studies between tissues; Section 1.5 provides a brief overview of concepts and methodologies which are relevant to the analyses performed as part of this Thesis.

1.1 Major events in the phylogenesis of mammals: From a single-cell protist to the appearance of vertebrates

1.1.1 The appearance of a multicellular ancestor of animals

The nature of the single-cell ancestor of animals is an issue that had intrigued biologists for more than a century. As early as 1866 the ancestor of animals was suggested to be a choanoflagellate-like species, a proposal made on the basis of the remarkable morphological similarity between feeding cells of sponges and choanoflagellate species (Lang et al. 2002). Choanoflagellates belongs to the protists group, a diverse collection of poorly characterized single-celled eukaryotes. Together with animals and fungi (but not plants) some protist species are classified into a single monophyletic group (Figure 1). This group is estimated to diverge from a eukaryotic ancestor common to plants, animals and fungi, at over a billion years ago, and possibly represents one of the earliest branching in the evolution of extant eukaryotes (Bonner 2003). A few morphological characteristics unite all members of this group, where the most obvious common traits are the presence of flattened

mitochondrial cristae, and a single basal flagellum in these cases where flagellum is present in motile reproductive stages (for example the male reproductive cells of animals). Within this group, molecular evidences indicate that choanoflagellate and metazoan are sister groups that form a monophyletic assemblage to the exclusion of exclusion of fungi (Lang et al. 2002; Steenkamp et al. 2006). The earliest metazoa – sponges – are most likely to have evolved from colonies of unicellular choanoflagellate (Kerszberg and Wolpert 1998).

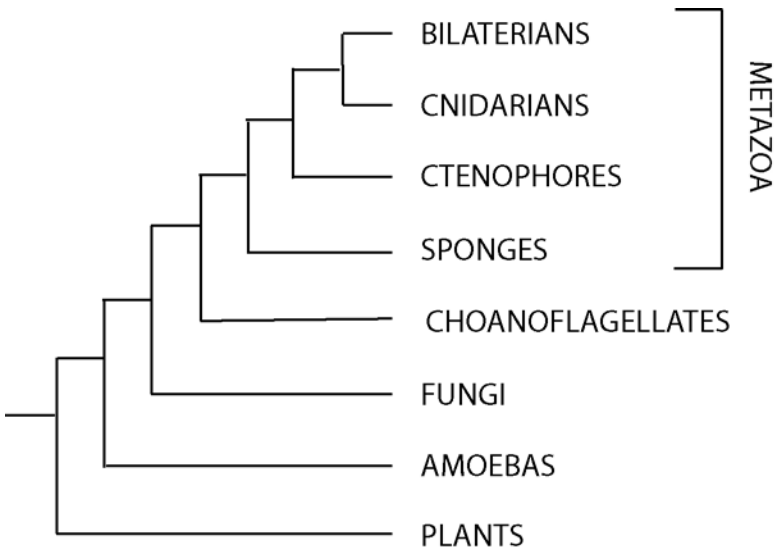

Figure 1 A phyletic tree showing the main metazoan phyla and their eukaryotic relatives. The tree is based on the analysis of 18S ribosomal DNA and is taken from (Raff 1996).

Why did unicellular eukaryotes give rise to multicellular organisms? What was the selective advantage to being multicellular? The first step in the transition to multicellularity was most likely the evolution of simple, undifferentiated cell clusters. Such organisms could have easily evolved as a result of mutations that encourage aggregation of cells, for example by a failure to separate following cell division (Kerszberg and Wolpert 1998). Advantages of big colonies can possibly be metabolic cooperation between heterotroph cells (Pfeiffer and Bonhoeffer 2003), or increased speed which enables the colony to compete more efficiently for nutrients and prey (Bonner 2003). When food was in short supply it is possible that some cells "gave up" their lives for other colony members (Kerszberg and Wolpert 1998).

Once multicellularity had occurred, how did the multicellular organization originate? The first steps in differentiation can be inferred from the study of simple multicellular organisms with two cell types, such as volvocine algae and cellular molds species (although the transition to two cell types in these species is independent of the transition in the ancestor of animals). In these species there is a close relation between the number of cells and the number of cell types (Bonner 2003). Increase in number of cells is therefore suggested to be a prime mover of the increase in complexity (as reflected by the number of cell types) as there is a better likelihood for the occurrence of beneficial mutations that by reaching spatial differentiation produce greater efficiency of the organism. There is evidence to show that eukaryotic cells needed to invent little new for the development of multicellular organisms. The unicellular ancestors already had, for example, a cell cycle in which there was temporal control of gene activity as well as the capability to exert forces for movement and changes in cell shape (Kerszberg and Wolpert 1998). In *Volvox* (a volvocine algae), genes involved in cellular differentiation include these that control the initial asymmetric cleavage that initiate the two distinct differentiation pathways that lead to differentiation of somatic cells (which are responsible for the movement of the colony) and germ cells (Bonner 2003). In the slime mold *Dictyostelium discoideum*, the development of each step in its life cycle is suggested to be adaptive and is accompanied by the appearance of a limited number of phenotypic characteristics. The basic assumption is that every aspect of slime mold development is governed by the advantage of dispersing its spores effectively (Bonner 2003). The amoebae first feed as separate isolated cells that engulf bacteria in the soil. Once they have cleared an area of food, they aggregate to form a multicellular cell mass that migrates in the form of a slug to a suitable spot and then differentiates into a small fruiting body consisting of a delicate cellulose stalk that encloses vacuolate dead amoebae and an apical spore mass in which each amoeba has become encapsulated in a resistant cellulose spore case (Bonner 2003). Assuming that the ancestral form was a species with solitary amoebae whose cysts were spread out in the soil, the first social stages probably arose by the appearance of an adhesive molecule whose properties were originally related to cyst formation. These primitive aggregates would have had a special selective advantage, such as their enhanced dispersal by nematodes. The next step postulated is the invention of chemotaxis, a more effective way of producing aggregations of cysts, where the important step would have been to produce an endogenous aggregation inducing factor such as cAMP.

Those aggregations that were large and therefore stuck up into the air would have had a selective advantage in dispersal, and through time ways of building greater height evolved. Subsequently, there was a coopting of the chemotaxis system to produce a polar migrating cell mass so that it could move away from the feeding ground, again presuming this would further propagation. Rising up into the air would obviously be more effective if the spores were lifted up on a stalk. This amounts to a selection for a division of labor, a differentiation into stalk cells and spores. Such differentiation began as cell sorting, and the most active amoebae, which would be the leanest ones, would lead the pack and form a primitive stalk, at the same time hoisting the replete amoebae up into the air to become spores (Bonner 2003).

1.1.2 Early metazoa

All metazoan phyla have evolved from a single ancestor, which appeared approximately 1000 million years ago (Muller 2001). Sponges (or Porifera) are the earliest extant metazoan, and are considered as living fossils of the ancestral metazoan (Figure 1). Most sponges are marine species and their body wall is perforated with pores (hence the name Porifera) through which water containing food particles is filtered. The sponges have a cellular grade of organization, with no discrete tissues or organs and only a few cell types (Raff 1996).

The next level of cellular organization is detected in cnidarians (sea anemones, hydras, and jellyfish) and ctenophores (comb jellies) phyla, which based on cladistic, morphological and molecular evidences, are located next to sponges on the phyletic tree (Figure 1). Cnidarians and ctenophores are diploblastic species - i.e., have primary two germ layers: endoderm and ectoderm. The diploblastic phyla form the sister group to the triploblastic bilaterians, which develop three germ layers: endoderm, ectoderm and mesoderm.

Figure 2 Marine flatworms (polycladids). The picture was taken from http://www.ucmp.berkeley.edu/platyhelminthes/platyhelminthes.html.

The origin of the bilaterian, triploblastic metazoa provided an innovation in metazoan body organization that allowed the evolution of the majority of animal body plans. The fossil evidence shows that this radical departure from the diploblastic animals occurred prior to Cambrian radiation. Flatworms (or platyhelminths) are the closest extant species to the ancestral bilateria (Figure 2). They are the first animals in the evolutionary tree to have a worm shape, with a head, and the capability to move in a specific direction, looking actively for food or for a mate (Martindale et al. 2002). They have no body cavity other than the gut, and since they lack an anus they use the same pharyngeal opening for both taking in food and expelling waste. The lack of a cavity constrains flatworms to be flat since they must respire by diffusion making a flattened shape necessary.

The majority of bilateral species already have a true body cavity – the coelom. The disparity of coelomate species is overwhelming and includes the arthropods, the annelids, and the mollusks phyla – all belong to the protostomes superphylum. The great success of these species in most niches marks the origin of the coelom as one of the key innovations in animal evolution. Deuterostomes is a relatively small superphylum within the coelomate, but it has the virtue of containing among its member phyla the vertebrates and other chordate. In chordate, the bilaterally symmetrical body plan of the ancestral deuterostome was transformed into a highly modified body plan. The deuterostome ancestor would have been bilaterally symmetrical and would have possessed a dorsal hollow nerve cord and gill clefts. The chordate lineage added a notochord and somites to produce a mobility system that allowed swimming by side-to-side motion (Raff 1996).

Vertebrates are a derived group within the chordates, which is more complex both morphologically and genetically in comparison to its chordate ancestor. Vertebrates

innovations include neural crest cells and their derivates, an elaborate segmented brain, and a well defined endoskeleton composed of cartilage and bones (Shimeld and Holland 2000).

Due to the growing number of sequencing projects, increasing parts of the eukaryote phylogeny are being covered by complete genomes. Up to date, fully assembled genome sequences are available for a variety of protists, fungi, plants and animal species (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi). Currently, all fully sequenced animals are bilateria (including nematodes, flies, and mammals). Though, on-going genome projects at different stages of progression cover wider parts of the metazoan kingdom. Such projects include the full sequencing of a larger collection of insects, nematodes, and chordates (including urochordates, birds, fishes, reptiles, amphibia, and additional mammalian species), as well as the sequencing of mollusca, cnidaria and arthropoda species (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi). Sequences retrieved from complete genomes were used for various analyses as part of this Thesis: chpater 2 concers a copmparative analysis of the complement of the enzymatic sets in fully sequenced species; in chpters 3, 4, and 5 sequences were used to determine the phyletic age of mammalian proteins; in chapter 6 sequences were used to estimate the phyletic age of mammalian metabolic reactions.

1.2 Innovations in the gene repertoire along the mammalian lineage

The accumulation of a growing number of fully sequenced unicellular and multicellular eukaryote species enables the investigation of the extent to which obvious differences between different lineages are reflected in differences among the protein complements encoded by their genome. Various comparative studies between eukaryote and prokaryote species, and among different eukaryotic groups indicate that eukaryotic protein inventions occurred largely through two important mechanisms (Apic et al. 2001; Aravind and Subramanian 1999; Copley et al. 1999; Waterston et al. 2002). The first is the combination of protein domains into new architectures. The second is lineage specific expansions of gene families that often accompanied the emergence of lineage specific functions and phylogenesis. When studying the major innovations in the transition to animal multicellularity, these two mechanisms had the major contribution to the emergence of new proteins. Only few domains are entirely metazoan-specific, and in many cases those domains

are largely based on stabilizing disulfide bonds or chelated metal ions, suggesting that they could have arisen easily *de-novo* (Aravind and Subramanian 1999).

Which are these proteins specific to metazoa? The functional assignment of proteins specific to multicellular animals indicates that a large fraction of them contain families of domains which mediate extracellular adhesion and families of domains that bind DNA and are predicted to regulate transcription (Aravind and Subramanian 1999; Chervitz et al. 1998; Copley et al. 1999; Rubin et al. 2000). Novel DNA-binding domains include metal chelating forms such as the nuclear hormone receptor family. Novel extracellular adhesion modules include C-type lectin domains (Aravind and Subramanian 1999; Rubin et al. 2000). The immunoglobulin domain and domains that take part in phosphotyrosine signal transduction are present in unicellular species and participate in intracellular communication, but are far more prominent in animals (Aravind and Subramanian 1999; Chervitz et al. 1998; Rubin et al. 2000). Massive expansions are also observed in domain families that participate in intracellular signaling such as cyclic nucleotide cyclases and calcium binding domains and in domains that target the ubiquitin-mediated degradation pathway. Other families have not been significantly expanded in metazoa but the architecture of their domains has noticeably changed. C2H2 finger domain is an example of a regulatory element whose incorporation into proteins diverges between unicellular and multicellular eukaryotes, so that animals contain more finger domains per protein in comparison to unicellular eukaryotes. In protein families that participate in signaling, such as the regulators of small GTPase signaling, very few domain architectures are conserved between metazoa and yeast (Aravind and Subramanian 1999).

The invention and proliferation of specific transcription factor families in animals correlates well with the increase in drastically different cell types and spatial complexity that is achieved through a developmental process. The increase and recruitment of extracellular interaction domains not only correlates with the need for adhesion in multicellular organisms but also with the need for intercellular contacts required for tissue development. Related to the intercellular communication is the rise in intracellular signaling that facilitates response to external stimulus. The novel juxtapositions of signaling domains during the origin of multicellular animals probably corresponded to the increase in intracellular signaling. Thus, it appears that the rise of animals' multicellularity can be correlated with the selection of protein families that allowed differentiation (the transcription factors) and cellular

communication (the signaling and adhesion specific domains) (Aravind and Subramanian 1999; Chervitz et al. 1998; Rubin et al. 2000).

How did the novel, metazoan-specific pathways evolve? The Notch pathway is an example of a key pathway integral for embryonic development in higher animals. Whereas the notch receptor and its ligand are estimated to be metazoan-specific, the intracellular components of the pathway (DNA-binding domain and chromatinic ATPase) are most likely descended from a unicellular ancestor. This suggests that the pathway did not arise all at once with the rise of animal's multicellularity but rather by recruiting pre-existing proteins and adding a few new inventions (Aravind and Subramanian 1999).

Similarly to the nature of the innovations accompanying the transition to multicellularity, most of the vertebrate specific proteins are the result of new domain combinations, and the increase in the number of proteins is to a large extent the result of expansion of protein families (Lander et al. 2001; Waterston et al. 2002). The number of domain combinations is higher in vertebrate compared to non-vertebrate metazoan and to unicellular eukaryotes. This difference is most prominent in the recent evolution of novel extracellular and transmembrane architectures in the mammalian lineage (Lander et al. 2001).

Families that are expanded in human relative to fly (*Drosophila melanogaster*) and worm (*Caenorhabditis elegans*) are in many cases involved in control of development, including many families of growth factors that control organogenesis. The detection of such homologous proteins between invertebrate and vertebrate, although in a low-copy number in invertebrates, indicates that developmental triggers of morphogenesis in vertebrates have evolved in a pre-vertebrate ancestor. Unlike the common origin of many growth factors, most prevalent transcription factor families are different in vertebrates and non-vertebrates (Lander et al. 2001), and numerous protein families of transcription factors were expanded on the vertebrate stem lineage, including the Hox and ParaHox families (Shimeld and Holland 2000). The innovations in the transcription factor repertoire have major biological implications because transcription factors are critical in animal development and differentiation. The emergence of major variations in the developmental body plans that accompanied the early radiation of animals could have been driven by lineage-specific proliferation of such transcription factors. Other families that were expanded include protein components of particular functional system such as cell death signaling system. The number of proteins involved in cytoskeleton, defence and immunity, and transcription and

translation is higher in human compared to invertebrate. These expansions are all clearly related to aspects of vertebrates' physiology (Lander et al. 2001). The future sequencing of a growing number of fully sequenced metazoan species will provide a better resolution for identifying the contribution of each lineage to the mammalian protein repertoire.

1.3 Differentiation of tissues in vertebrates

An animal starts its life as a single cell – a fertilized egg. During development this cell divides repeatedly to give rise to many different cells in a final pattern of a well defined body structure. The embryonic development of vertebrates, as of other animals, involves dramatic cell movement. The pattern of cell movement is dictated by the pattern of gene expression, which determines cell surface properties and motility.

The three germ layers from which higher animals are constructed: endoderm, mesoderm, and ectoderm are formed in gastrulation, an early stage of embryogenesis, where cells from the exterior of the early embryo tuck into the interior to form a gut cavity. In vertebrates, the movements of gastrulation are organised by signals from the Organizer (the node in the mouse embryo, which corresponds to the dorsal lip of the amphibian blastopore). The signals from the Organizer specify the dorsoventral axis of the body. Subsequent development involves the rolling up of an ectoderm layer which forms the neural tube and neural crest. In the midline, a rib of specialized cells called the notochord elongates to form the central axis of the embryo. The long slabs of mesoderm on either side of the notochord become segmented into somites, which are masses of mesoderm distributed along the two sides of the neural tube that will eventually become dermis, skeletal muscle and vertebrae. The formation of the somites depends on periodic pattern of gene expression, which dictates the way the mass of cells will break up into separate blocks. Similarly, the left-right anatomical asymmetry of the vertebrate body is foreshadowed by left-right asymmetry in the pattern of gene expression in the early embryo. Migrant cells, such as those of the neural crest, break loose from their original neighbours and travel through the embryo to colonize new sites. Specific cell-adhesion molecules, such as cadherins and integrins, help to guide the migration and control the selective cohesion of cells in new arrangements (Alberts 1994).

In every segment along the anterior-posterior axis of the vertebrate embryo there is a serial pattern of expression of homeotic selector genes, called Hox genes. The products of these genes are transcription-factor proteins; all possess a highly conserved DNA-binding homeobox domain. Different Hox genes exert a different effect in each segment, largely due to variations in those parts of the protein that do not bind directly to DNA, but interact with other proteins in DNA-bound complexes. The different patterns in these complexes along the different parasegments of the embryo act to dictate which DNA-binding sites will be recognised and whether the effect on transcription at those sites will be activation or repression. In this way, the products of the Hox genes combine with other gene-regulatory proteins and modulate their actions giving each parasegment along the body axis its characteristic features. In mouse, as in most animals, the Hox genes are grouped in the genome, and they are arranged in a sequence that matches their sequence of expression along the axis of the body. Other homeobox-containing genes are scattered in the genome and are not clustered into Hox complexes. Many of them also participate in controlling the variations on the basic developmental theme. Different classes of neurons, for example, are often distinguished from one another by expression of specific genes of the large homeobox gene super-family (Alberts 1994; Raff 1996).

More than 200 cell types are traditionally recognised in the adult human body. Those collaborate with one another to form a multitude of different tissues, arranged into organs performing widely varied functions. The genome is normally identical in every cell; and the cells differ because they express different sets of genes. The final pattern of differentiated cell types is thus the outcome of a more hidden program of cell specialization – a program played out in changing patterns of expression of gene-regulatory proteins, giving one cell different potentialities from another (Alberts 1994).

1.4 Conservation and variation between adult mammalian tissues

The availability of various high-throughput techniques for the detection of gene expression allowed the construction of the expression profile of thousands of genes. Several public databases are now available that report the expression profiles of various adult mammalian tissues (for example: (Hsiao et al. 2001; Miki et al. 2001; Su et al. 2004)). Those

databases enable the detection of how obvious phenotypic differences between tissues are related to the repertoire of genes expressed in each tissue.

The initial clustering of tissues reveals that the expression profiles reflect the embryological origin of the tissues. For example, muscular tissues, nerve tissues, and digestive tissues fall into discrete groups (Miki et al. 2001). Genes that are expressed in all tissues were termed housekeeping genes (or maintenance genes) and are assumed to be required to maintain basic cellular functions, such as those activities which are critical to the successful completion of the cell cycle (Warrington et al. 2000). Maintenance genes encode proteins mediating a variety of basic cellular functions including intermediary metabolism, gene transcription, protein translation, proteins involved in cell signaling and communication, and proteins involved in shaping the structure and motility of the cell (Hsiao et al. 2001). Most of the ribosomal proteins are included in this set (Hsiao et al. 2001; Miki et al. 2001). In each tissue, genes unique to the tissues are expressed (Su et al. 2002; Warrington et al. 2000). Those genes are possibly critical for the specific physiological roles of the tissues. For example, genes whose expression is restricted to the testis, include genes that are known to be involved in testis-specific functions, such as SRY (sex determining region Y)-box5 (SOX5), testicular tektin 2 (TEKT2), and zona pellucida binding protein (ZPBP) (Su et al. 2002).

1.5 Methodology

This section provides a brief overview of concepts and methodologies which are relevant to the analyses performed as part of this Thesis.

1.5.1 BLAST and PSI-BLAST

The Basic Local Alignment Tool (BLAST) programs are widely used tools for rapidly searching protein and DNA databases for sequence similarity. The BLAST algorithm emphasizes speed over sensitivity making the algorithm practical on the huge genome databases currently available (Altschul et al. 1990; Altschul et al. 1997).

To run, BLAST requires two sequences as input: a query sequence and a sequence database. BLAST searches for high scoring sequence alignments between the query sequence

and sequences in the database using a heuristic approach. The central idea of the BLAST algorithm is that statistically significant alignment is likely to contain high-scoring pairs (HSP) of aligned words. BLAST first scans the database for a words of length "W" (typically of length three for proteins) that scores at least "T" when compared to the query using a substitution matrix. Any aligned word pair satisfying this condition is called a hit. In the second stage, BLAST tries to extend the match in both directions, while checking whether each hit lies within an alignment with score sufficient to be reported. If a high-scoring ungapped alignment is found, the database sequence is passed on to the third stage. In the third stage, BLAST performs a gapped alignment between the query sequence and the database sequence. The gapped alignment is based upon the observation that an HSP of interest may entail multiple hits on the same diagonal and within a relatively short distance of one another. Extension of alignment is invoked only when two non-overlapping hits are found within chosen distance A of one another on the diagonal. Any hit that overlaps the most recent one is ignored. The great majority of hits are dismissed after the calculation of looking up, for the appropriate diagonal, the coordinate of the most recent hit, checking whether it is within distance A of the current hit's coordinate, and finally replacing the old with the new coordinate (Altschul et al. 1990; Altschul et al. 1997).

The opening and continuing of gaps affect the score of the alignment. Gap scores are typically calculated as the sum of the gap opening penalty and the gap extension penalty. The final score of an alignment is calculated as the sum of substitution and gap scores. The expected (E) value of an alignment describes the number of different alignments with scores equivalent to, or better than the alignment's score that are expected to occur in a database search by chance. The E-value takes into account the size of the database. The lower the E-value, the more significant the score is. Statistically significant alignments are then displayed to the user.

Position-Specific Iterative BLAST (PSI-BLAST) is one of the BLAST programs, which is used for finding distant relatives of a protein (Altschul et al. 1997). PSI-BLAST performs three distinct operations. First, it constructs a multiple alignment from BLAST output. To produce the multiple alignment the program simply collects all database sequence segments that have been aligned to the query with E-value below a threshold. The query is then used as a template for constructing a multiple alignment. Then, it processes this alignment into a position-specific score matrix. In constructing the matrix score, not only a

column's observed frequencies are important, but also the effective number of independent observations it constitutes: a column consisting of a single valine and a single isoleucine carries different information than one consisting of five independently occurring instances of each. Finally, PSI-BLAST uses this matrix to search the database, and a larger group of proteins is found. This larger group is used to construct another profile, and the process is iterated. By using a motif (or profile) rather than sequence, PSI-BLAST is much more sensitive in picking up distant evolutionary relationships than the standard protein-protein BLAST. In many instances, PSI-BLAST is able to automatically uncover biological interesting similarities that elude simple database searches. Multiple iterations of the PSI-BLAST are sometimes required to recognize the more distantly related protein family members.

1.5.2 Estimation of synonymous and nonsynonymous substitution rates (d_N/d_S)

Estimation of synonymous and nonsynonymous substitution rates is important in understanding the dynamics of molecular sequence evolution (Kimura 1981). As synonymous (silent) mutations are largely assumed invisible to natural selection, while nonsynonymous (amino-acid replacing) mutations may be under strong selective pressure, comparison of the rates of fixation of those two types of mutations provides a powerful tool for understanding the mechanisms of DNA sequence evolution. For example, variable nonsynonymous/synonymous rate ratios among lineages may indicate adaptive evolution or relaxed selective constraints along certain lineages. Likewise, models of variable nonsynonymous/synonymous rate ratios among sites may provide important insights into functional constraints at different amino-acid sites and may be used to detect sites under positive selection (Yang and Nielsen 2000).

Methods for estimating the numbers of synonymous and nonsynonymous substitutions between two DNA sequences involve three steps. First, the numbers of synonymous (*S*) and nonsynonymous (*N*) sites in the sequences are counted. Second, the numbers of synonymous and nonsynonymous differences between the two sequences are counted. Third, a correction for multiple substitutions at the same site is applied to calculate

the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site between the two sequences. Two major features of DNA sequence evolution - transition/transversion rate bias, and base/codon frequency bias – are takes into account at this stage (Yang and Nielsen 2000).

1.5.3 Functional annotation methods

Functional annotations are essentially based on the expansion of the relatively small number of experimentally determined functions to large collections of proteins. The majority of the annotation procedures use sequence similarity in order to transfer functional annotations between proteins (Valencia 2005). Such sequence-based procedures for functional annotations use a system of rules for transferring the annotation from the most similar sequence with an annotated function. Along the direct transfer of annotation between homologues proteins, improved approaches use annotations at domain level (Bateman et al. 2004; Mulder et al. 2005); or a more comprehensive analysis of protein families including the identification of orthologous sequences (Tatusov et al. 2003; von Mering et al. 2003).

While the exploration of sequence relationships is still the mainstream tendency, the exploration of other sources of information other than sequence similarity for the prediction of protein function are becoming more widely used. Such approaches use, for example, neighborhood relationships on the genome in order to predict function (reviewed in (Valencia 2005)). Alternatively, the availability of large-scale protein interaction networks, obtained by high-throughput proteomics or by several prediction methods, has made it possible to predict function on the basis of the relationships within these networks (Marcotte 2000). Such 'nonhomology' methods analyze patterns such as domain fusion, conserved gene position and gene co-inheritance and coexpression to identify protein–protein relationships (Huynen et al. 2000; Marcotte et al. 1999; van Noort et al. 2003).

It can be said that all approaches to function annotation face formidable practical problems related to the accuracy of the input experimental information, the reliability of current systems for transferring information between related sequences, and the reproducibility of the links between database information and the original experiments reported in publications. In addition to these technical difficulties, lies the deeper problem of

the evolution of protein function in the context of protein sequences and structures. Nevertheless, despite the limitations of the current approaches, the annotation of protein function at large scale is essential for performing comprehensive analyses of genomic data (Valencia 2005).

1.5.4 Gene Ontolgy (GO)

Genome sequencing projects have driven the development of functional classification schems. One of the most widely used classification schemes is the Gene Ontology (GO) classification scheme (Ashburner et al. 2000). This classification scheme separates gene function into three independent ontologies: biological process, molecular function and cellular component. One particular protein may be described by many different categories within the classification scheme. For example, a particular protein may function in several different biological processes, contain domains with diverse molecular function and participate in multiple interactions. Each ontology is therefore a network of nodes, able to handle data at different levels of completeness (Ashburner et al. 2000).

1.5.5 Definition and identification of orthologous and paralogous proteins

Orthology and paralogy are central concepts of comparative genome analysis. The term orthology describes genes in different species that originate from a single gene in the last common ancestor of these species. The term paralogy refers to genes that have diverged via a gene duplication event within a genome (Hulsen et al. 2006). Paralogy can exist between genes in different species, since gene duplication events occur both before and after speciation. The term 'in-paralogues' indicate paralogues that arose through a gene duplication event after speciation, while 'out-paralogues' arise following a gene duplication preceding speciation. Out-paralogues can never be orthologues, while in-paralogues can form a group of genes that together are orthologous to a gene in another species (O'Brien et al. 2005). Orthologous genes are more likely to have a functional similarity than paralogous genes, which have often undergone changes in substrate or ligand specificity. The high level of functional conservation between orthologous proteins makes orthology highly relevant for protein function prediction. It is also widely used in genome analysis, where the

information about a protein in one species is used for the functional annotation of the orthologous protein in another species (Hulsen et al. 2006).

Traditionally, orthology relationships for individual gene families have been predicted by carefully constructed multiple alignments and by reconstructing phylogeny. However, for genome-scale investigations, current methods do not yet automatically generate multiple alignments of unfailing quality. Instead, orthology across whole genomes has been determined automatically using reciprocal best hits in all-against-all comparisons of amino acid sequences (Clamp et al. 2003; Remm et al. 2001; Waterston et al. 2002), where two sequences are identified as orthologues if they find each other as the highest scoring alignments. This procedure is most reliable for data retrieved from relatively simple genomes such as those of prokaryote species, but works less well where gene sets are retrieved from more complex species where the genome includes lineage-specific expansions of gene families. When duplications have independently occurred in one or both of the lineages where orthology is inferred, the resulting orthologous genes are in one-to-many or many-to-many relationships respectively. For each set of orthologues, relying solely on reciprocal best hits will only identify one pair out of all orthology relationships (Goodstadt and Ponting 2006). Several methods approaching the difficulties of identifying ortholgy in heiger species had been suggested, including the widely used Ensembl (Clamp et al. 2003) and Inparanoid (Remm et al. 2001) methods for predicting orthology and paralogy. Both approaches start with reciprocal best-hitting protein sequence pairs. In Ensembl, the classification into orthologous cluster considers not only sequence similarity but also the genomic location of the sequences. In Inparanoid, additional orthologues are added to the original orthologous pair only if their proteins are more sequence-similar to the initial orthologue from the same species.

A major limitation of the Inparanid and the Ensembl approaches is the underlying assumptions that protein similarity accurately reflects evolutionary distance and that paralogues evolve at equal rates. PhyOP (phylogenetic orthology and paralogy) - a new approach to predict ortholgy and paralogy has been recently designed to meet this requirement (Goodstadt and Ponting 2006). Unlike Inparanoid and Ensembl, PhyOP predicts orthology using a distance metric based not on amino acid substitutions, but rather on dS, - the number of synonymous nucleotide substitutions per synonymous site. This

approach is applicable for detecting orthology between highly related species, where for over long evolutionary distances the method becomes increasingly less appropriate because of saturation at synonymous sites.

1.5.6 Microarray data acquisition

1.5.6.1 Acquisition of microarry data

A DNA microarray is a collection of spots of short DNA sequences termed probes, which are attached to a solid surface for the purpose of monitoring expression levels. A single DNA microarray might contain thousands of probes, enabling the simultaneous detection of the expression of many genes. In order to detect expression, microarrays exploit the preferential binding (hybridization) of complementary nucleic acid sequences. The process of hybridization of nucleotides to complementary nucleotides allows two DNA chains to form very stable double-stranded hybrids. With a single mismatch, the bond will be very much weaker. The longer the sequence, the less the impact of a single mismatch, but under the conditions used in many microarray experiments, a single mismatch can cause a significant destabilization of the hybrid. The trick in making use of the microarray then is having correct oligonucleotide sequences on the array with which to detect specific gene fragments (Watson and Akil 1999).

When one wants to monitor the expression of genes in a given sample using DNA microarray, than the initial step is the extraction of mRNAs from the sample. Then, the entire population of mRNA is copied into cDNA, while incorporating a fluorescent label. These labeled cDNAs represent the mRNA populations from which they were derived in a stable and detectable form. The resultant labeled cDNA populations are then applied to the DNA array so that each specific cDNA in the sample can bind to the cognate gene-specific DNA element on the array, allowing its abundance in the sample to be measured. The next task is to scan the array with a laser confocal microscopic system capable of detecting a weak fluorescence signal over each location on the array. The fluorescent signal over each of the locations on the arrays is captured and quantitated. After quantitating the signal and correcting for several types of errors and accounting for controls, the value for each location is inserted into a database cell associated with the gene/mRNA represented by the probe on

the array. In effect, a mRNA concentration value is now available for each gene product in each area of each organism. (Watson and Akil 1999).

1.5.6.2 The Affymetrix Gene Chip

While DNA arrays have been produced both by individual scientists and by biotechnology corporations, gene chips are an exclusively commercial product. Gene chips is a phrase coined by Affymetric, Inc. in California, which uses a unique procedure to build high-density oligonucleotide microarrays (Lipshutz et al. 1995). Gene Chip microarrays are manufactured for several species including mouse and human. The GeneChips are widely used tool for the monitoring of expression. Since parts of this Thesis concern the analysis such data, the following paragraphs contain a brief description of their design and its implications on the procedure of data acquisition.

The Affymetrix microarrays are divided into several hundred thousand probe cells grouped in pairs. Each pair has a Perfect Match (PM) and a MisMatch (MM) cell. The PM cell contains oligonuceotide sequence designed to hybridize to a specific target gene. The corresponding MM cell contains an altered version of the PM version designed to increase the sensitivity of the method by capturing non-specific hybridization. A group of probe pairs designed to measure the abundance of a specific target molecule is known as a probe set. Each probe set typically consists of between 16 and 20 probe pairs, where each probe is composed of about 25 nucleotides (Rajagopalan 2003).

The representation of a gene by several reporters requires that the fluorescent intensity determined for each probe cell will be converted to a measure of relative transcript abundance. A few different types of analysis methods are available based on different statistical approaches (Rajagopalan 2003). In MAS5 (Microarray Suite Software 5), an analysis approach applied to data analysed in this Thesis, Affymetrix has implanted algorithm for absolute and comparison analyses that are based on non-parametric statistical techniques (Liu et al. 2002). In the absolute analysis, detection calls are defined for each probe pair (the corresponding PM and MM pair) on the basis of discrimination score evaluating the relative difference between the call in each of the two probes. The detection call for a probe set is made via a Wilcoxon signed rank test to determine whether the mean value of the discrimination score over all probe pairs in the probe set is greater than 0.015. The signal of

expression level of the target molecule associated with each probe set is defined as the average value of PM in the set (Liu et al. 2002).

1.6 Outline of Thesis

Work described in this PhD Thesis attempts to relate innovations in the gene repertoire along the mammalian lineage to the appearance of highly differentiated tissue types. Several different approaches have been followed to address some aspects of this problem. **Chapter 2** describes a comprehensive study of the pattern of expansion of the complement of enzymes in various species that was performed in order to obtain a better view of the principles underlying the expansion of the gene repertoire in mammals. **Chapter 3** presents a study of the relationship between expansion of gene families in mammals and specialisation of tissue types that was performed by analysing mouse gene expression data from many tissues. The relationship between the physiological role of a tissue and the unique composition of the genes expressed in that tissue is discussed in **Chapter 4** where a few examples of genes whose specific expression define tissue-specific processes are described. In **Chapter 5** a more robust approach was taken in order to characterise each tissue with regard to the distribution of molecular function and phyletic origin of the genes expressed in the tissue. **Chapter 6** concerns the relationship between innovations in the mammalian metabolic repertoire and the development of new, tissue-specific, pathways. **Chapter 7** ties together the main observations from this Thesis.

Chapter 2: The complement of enzymatic sets in different species

2.1 Overview

This chapter describes a comprehensive analysis of the complement of enzymes in a large variety of species. Since enzymes are a relatively conserved group, there are several classification systems available that are common to all species and link a protein sequence to an enzymatic function. Therefore, enzymes are an ideal functional group to study the relationship between sequence expansion, functional divergence and phenotypic changes. By using information retrieved from the well annotated SWISS-PROT database together with sequence information from a variety of fully sequenced genomes and information from the EC functional scheme we aimed to estimate the fraction of enzymes in genomes and to determine the extent of their functional redundancy in different domains of life. We found that prokaryote and eukaryote species are different both in the fraction of enzymes in their genomes and in the pattern of expansion of their enzymatic sets. We observe an increase in functional redundancy accompanying an increase in species complexity. The increased functional redundancy in eukaryotic species reflects a general trend and is observed for the majority of the reaction types.

This work has been published in (Freilich et al. 2005).

2.2 Introduction

Linking information on the size of a protein group in a species with information on its functional diversity provides an insight into the ways in which genome expansion affects the functional repertoire in different species and in different domains of life. Here, the enzyme complement from various species is considered in detail. Several factors make the set of enzymes a natural candidate for such a study: enzymes are a relatively conserved group (Doolittle et al. 1996; Peregrin-Alvarez et al. 2003) that has been studied extensively, therefore, there are several unique classification systems available. The Enzyme Commission (EC) scheme is a universal reaction classification system common to all species (Tipton and Boyce 2000). Such a system is a fundamental requirement for studying function divergence.

Metabolic pathway databases - such as KEGG (Kanehisa et al. 2002), WIT (Overbeek et al. 2000) or EcoCyc (Karp et al. 2000) - relate an enzymatic reaction to a higher level cellular process, enabling us to make the link between genotypic and phenotypic changes.

Previous studies have shown that metabolic genes have a roughly constant fraction of the gene content of the genome in each domain of life, i.e., a constant and not-necessarily common fraction for bacteria, archaea and eukaryota (van Nimwegen 2003). Here, the first goal is to examine whether the fraction of the enzymes in the genome is constant not only within but also across the three domains of life. The analysis concerns all enzymes in a species, i.e., all proteins classified under the EC scheme (including proteins involved in micro and macromolecule metabolism, signalling and degradation). The second goal is to characterise the pattern of expansion of enzymatic sets in different species and different domains. In particular, we examine the extent to which two different modes contribute to expansion: the broadening of the reaction repertoire of an organism (more enzymatic reactions) or an increase in its functional redundancy (more proteins performing the same function).

2.3 Methods

2.3.1 Construction of the query list

In order to ensure that the enzyme set for each species is as complete and validated as possible, we listed all the highly curated UniProt (Apweiler et al. 2004) enzymes. The list is composed of SWISS-PROT (Boeckmann et al. 2003) (Release 41.0) proteins assigned with an EC number. In order to exclude assignments based on sequence similarity, annotations including 'hypothetical protein', 'by similarity', 'putative' or 'probable' were filtered from the list, as well as proteins for which all references are common to more than 20 entries. TrEmbl (Boeckmann et al. 2003) (release 23) entries with the "experimental" evidence code were included in the list. The final enzyme list known as the 'query list' is composed of 23,431 proteins. The distribution of the enzymes between the three domains of life is the following: 14128 Eukaryota (1869 human), 7653 Bacteria, 574 Archaea, and 1076 viruses. The comprehensiveness of the query list composition aims to cover all possible domain combinations.

Each enzyme in the list was used as a query for a PSI-BLAST search against fully sequenced genomes, including 63 bacteria species, 16 archaea species and 6 eukaryota species. The PSI-BLAST search was performed in Inpharmatica Ltd by Ruth V. Spriggs, Richard A. George, Bissan Al-Lazikani and Mark Swindells.

2.3.2 Construction of enzymatic sets for each species

Homologues to each of the SWISS-PROT sequences, in the constructed enzymatic query list, were retrieved from the BiopendiumTM (version 13, created 21 March 2003) (Swindells et al. 2002) using PSI-BLAST (Altschul et al. 1997) to three iterations (E-value cut-off for inclusion in next iteration 0.003). Homologues were only accepted if (1) a homologue is recognized with E-value <= 0.001; (2) the alignment covers at least 80% of the query protein; (3) a homologue was found within a pre-defined set of completed genomes. The set of completed genomes includes genomes from the NCBI completed genomes (15 August 2002), Ensembl human (Birney et al. 2004) (release 11.31), and Ensembl mouse (release 11.3).

Since Ensembl fly and worm genomes were not included in BiopendiumTM version 13, these genomes were constructed by comparing all fly/worm sequences in the BiopendiumTM against the latest available Ensembl genomes (6 August 2003). All sequences from either C.elegans or D.melanogaster were retrieved from the BiopendiumTM and processed. The most complete resulting genomes were 98.8% complete for C.elegans (at 90% identity and 80% overlap) and 99.1% complete for D.melanogaster (at 95% identity and 80% overlap).

2.3.3 Data retrieved from KEGG database

The number of enzymes in a species was downloaded at Jan 2004 from http://www.genome.ad.jp/kegg/docs/upd_genes.html.

The number of reactions (fully assigned EC numbers) per species was extracted from the enzyme file in the LIGAND section by automated text parsing (downloaded at February 2004).

The classification of genes to reactions was extracted from the 'ligand' section in KEGG. The classification of genes to orthologue clusters was extracted from the 'ko' section in KEGG (downloaded on August 2004).

2.3.4 Determination of the representation of a species or a group of species in the query list

In order to examine how well a species is represented in the query list we counted all proteins in a species matching an enzyme from the query list with more than 80% identity. We used the 80% cut-off rather than counting directly the number of enzymes per species in the query list in order to transfer annotations between closely related species. For example, if an enzyme in the query list was identified in one strain of *E.coli* the use of the 80% cut-off will enable us to transfer the annotation to other *E.coli* strains. The average fraction of these highly curated enzymes (as determined by the 80% cut-off) over all species was calculated, as well as the z-score (number of standard deviations from the mean in each species, Table 1).

In order to study the level of representation in the query list for each domain of life we repeated the same procedure, done for each species, for each group of species (e.g., bacteria, archaea and eukaryota). The fraction of highly curated enzymes per group was calculated by summing the number of such enzymes in all species classified to the group and dividing it by the summed number of genes. As the data set includes many more bacteria species (63) than archaea (16) and eukaryota (6) we divided the bacterial species into the lower level classification of sub-families, repeating the same procedure (Table 2).

Table 1 Size and fraction of enzyme sets in species. E - eukaryota; B – bacteria; A – archaea.

Species		Proteome size	Permissive set			Conservative set			Representation in experimental studies***
			Number of enzymatic proteins	Fp *	$\frac{(Fp - \overline{Fp})}{\sigma p}$ **	Number of enzymatic proteins	Fc *	$\frac{(Fc - \overline{Fc})}{\sigma c}$ **	
E	H.sapiens	24847	5027	0.20	-2.56	2608	0.10	-0.39	1.14
	M.musculus	22345	4532	0.20	-2.55	2442	0.11	-0.32	1.11
	D.melanogaster	13525	3369	0.25	-1.77	1257	0.09	-0.59	-0.19
	C.elegans	19556	3519	0.18	-2.94	1042	0.05	-1.25	-0.52
	S.cerevisiae	6333	1790	0.28	-1.20	1023	0.16	0.55	2.49
	S.pombe	5000	1435	0.29	-1.12	654	0.13	0.04	0.50
B	E.coli	4279	1684	0.39	0.69	1077	0.25	2.05	4.55
	E.coli_O157	5324	1761	0.33	-0.38	1075	0.20	1.22	3.44
	E.coli_O157J	5361	1748	0.33	-0.46	1070	0.20	1.18	3.40
	S.typhimurium	4553	1720	0.38	0.43	1048	0.23	1.69	3.68
	Y.pestis	4083	1397	0.34	-0.18	821	0.20	1.21	1.36
	Buchnera	574	337	0.59	3.99	267	0.47	5.60	0.54
	H.influenzae	1714	745	0.43	1.39	482	0.28	2.54	0.57
	P.multocida	2015	898	0.45	1.58	535	0.27	2.28	0.17
	X.fastidiosa	2832	812	0.29	-1.13	392	0.14	0.16	-0.61
	X.campestris	4181	1535	0.37	0.24	573	0.14	0.14	-0.49
	X.axonopodis	4312	1567	0.36	0.18	579	0.13	0.09	-0.51
	V.cholerae	3835	1310	0.34	-0.19	706	0.18	0.92	-0.17
	P.aeruginosa	5567	2146	0.39	0.56	863	0.16	0.44	0.12
	N.meningitidis	2079	738	0.35	0.04	426	0.20	1.27	-0.12
	N.meningitidis_A	2065	726	0.35	-0.02	414	0.20	1.20	-0.15
	R.solanacearum	5116	1752	0.34	-0.18	661	0.13	0.01	-0.56
	H.pylori	1576	538	0.34	-0.19	177	0.11	-0.27	-0.31
	H.pylori_J99	1491	534	0.36	0.09	176	0.12	-0.18	-0.29
	C.jejuni	1634	657	0.40	0.84	200	0.12	-0.10	-0.32
	R.prowazekii	835	351	0.42	1.15	128	0.15	0.41	-0.22
	R.conorii	1374	364	0.26	-1.50	133	0.10	-0.53	-0.44
	M.loti	7275	2603	0.36	0.08	674	0.09	-0.60	-0.53
	S.meliloti	6205	2352	0.38	0.45	734	0.12	-0.17	-0.18
	A.tumefaciens	5402	2101	0.39	0.62	630	0.12	-0.20	-0.37
	A.tumefaciens_C	5299	2093	0.39	0.72	624	0.12	-0.18	-0.38
	B.melitensis	3198	1212	0.38	0.45	470	0.15	0.31	-0.35
	C.crescentus	3737	1480	0.40	0.74	453	0.12	-0.12	-0.58
	B.subtilis	4112	1528	0.37	0.32	694	0.17	0.67	1.76
	B.halodurans	4066	1470	0.36	0.15	571	0.14	0.20	-0.40
	S.aureus_N315	2625	1041	0.40	0.75	408	0.16	0.45	0.04
	S.aureus_Mu50	2748	1033	0.38	0.39	400	0.15	0.28	-0.05
	S.aureus_MW2	2632	1013	0.38	0.55	397	0.15	0.37	-0.07

	L.monocytogenes	2846	1160	0.41	0.93	463	0.16	0.57	-0.43
	L.innocua	3043	1144	0.38	0.39	443	0.15	0.28	-0.48
	L.lactis	2267	860	0.38	0.45	347	0.15	0.41	0.23
	S.pyogenes	1697	671	0.40	0.73	289	0.17	0.69	-0.07
	S.pyogenes_M18	1845	675	0.37	0.22	295	0.16	0.52	-0.17
	S.pneumoniae	2094	838	0.40	0.81	327	0.16	0.46	0.02
	S.pneumoniae_R6	2043	823	0.40	0.85	323	0.16	0.49	0.02
	C.acetobutylicum	3848	1482	0.39	0.55	389	0.10	-0.46	-0.46
	C.perfringens	2723	1055	0.39	0.59	321	0.12	-0.18	-0.51
	T.tengcongensis	2588	937	0.36	0.16	333	0.13	0.00	-0.66
	M.genitalium	484	207	0.43	1.28	56	0.12	-0.22	-0.36
	M.pneumoniae	689	227	0.33	-0.40	64	0.09	-0.59	-0.49
	M.pulmonis	782	335	0.43	1.29	73	0.09	-0.59	-0.60
	U.urealyticum	614	246	0.40	0.81	51	0.08	-0.76	-0.54
	M.tuberculosis	3927	1482	0.38	0.42	411	0.10	-0.40	-0.19
	M.tuberculosis_ CDC1551	4187	1459	0.35	-0.07	407	0.10	-0.52	-0.20
	C.glutamicum	3040	1056	0.35	-0.09	341	0.11	-0.27	-0.08
	S.coelicolor	7897	2816	0.36	0.06	623	0.08	-0.83	-0.43
	F.nucleatum	2067	765	0.37	0.29	216	0.10	-0.40	-0.63
	C.trachomatis	895	331	0.37	0.29	102	0.11	-0.24	-0.25
	C.muridarum	916	334	0.36	0.20	102	0.11	-0.29	-0.29
	C.pneumoniae	1054	342	0.32	-0.48	105	0.10	-0.48	-0.53
	C.pneumoniae_ AR39	1112	343	0.31	-0.76	105	0.09	-0.57	-0.54
	C.pneumoniae_ J138	1069	345	0.32	-0.51	105	0.10	-0.51	-0.54
	B.burgdorferi	1638	395	0.24	-1.90	69	0.04	-1.44	-0.41
	T.pallidum	1036	309	0.30	-0.93	79	0.08	-0.87	-0.53
	Synechocystis	3167	1149	0.36	0.17	407	0.13	-0.00	-0.12
	Anabaena	6129	1908	0.31	-0.71	487	0.08	-0.82	-0.39
	D.radiodurans	3182	1192	0.37	0.37	358	0.11	-0.27	-0.61
	A.aeolicus	1560	634	0.41	0.91	220	0.14	0.21	-0.58
	T.maritima	1858	745	0.40	0.82	227	0.12	-0.11	-0.05
A	M.jannaschii	1785	661	0.37	0.30	189	0.11	-0.38	-0.01
	M.acetivorans	4540	1380	0.30	-0.83	293	0.06	-1.07	-0.43
	M.mazei	3371	1102	0.33	-0.44	250	0.07	-0.91	-0.39
	M.thermoautotrophicum	1873	665	0.36	0.04	195	0.10	-0.41	0.02
	M.kandleri	1687	544	0.32	-0.52	136	0.08	-0.80	-0.52
	A.fulgidus	2420	812	0.34	-0.29	202	0.08	-0.75	-0.60
	Halobacterium	2622	766	0.29	-1.03	175	0.07	-1.03	-0.52
	T.acidophilum	1482	590	0.40	0.77	122	0.08	-0.77	-0.42
	T.volcanium	1499	574	0.38	0.51	118	0.08	-0.83	-0.56
	P.horikoshii	1801	599	0.33	-0.34	148	0.08	-0.77	-0.26
	P.abyssi	1769	645	0.36	0.20	177	0.10	-0.48	-0.21
	P.furiosus	2065	701	0.34	-0.23	185	0.09	-0.65	-0.15
	A.pernix	1840	551	0.30	-0.91	117	0.06	-1.08	-0.60
	S.solfataricus	2977	880	0.30	-0.97	198	0.07	-1.03	-0.28
	S.tokodaii	2826	779	0.28	-1.31	189	0.07	-1.03	-0.55
	P.aerophilum	2605	658	0.25	-1.71	153	0.06	-1.16	-0.66

* Number of enzymes/proteome size (fraction of enzymes)
**Standard deviation distance from mean (mean fraction of enzymes in all 85 species).

$\overline{Fp}$ = 0.35, $\overline{Fc}$ = 0.13, σp = 0.058, σc = 0.060.

***Standard deviation distance from mean value for level of representation (see material and methods)

Table 2 Level of representation of different species groups in SWISS-PROT.

	Species group (no. of species)		Total number of genes in the group*	Total number of query-list highly related enzymes ** in the group	Fraction	Standard deviation distance from mean***
Within domains	Eukaryota (6)		91606	5376	0.06	0.95
	Bacteria (63)		184396	7214	0.04	0.10
	Archaea (16)		37162	471	0.01	-1.05
Within bacterial subfamilies	proteobacteria	Gammaproteobacteria (12)	48630	4745	0.10	2.77
		Betaproteobacteria (3)	9260	128	0.01	-0.36
		Epsilonproteobacteria (3)	4701	78	0.02	-0.26
		Alphaproteobacteria (7)	33325	429	0.01	-0.40
	Firmicutes	Bacillales (7)	22072	786	0.04	0.45
		Lactobacillales (5)	9946	309	0.03	0.28
		Clostridia (3)	9159	62	0.01	-0.62
		Mollicutes (4)	2569	20	0.01	-0.59
	Actinobacteridae (4)		19051	363	0.02	-0.16
	Fusobacterales (1)		2067	5	0.00	-0.79
	Chlamydiales (5)		5046	54	0.01	-0.48
	Spirochaetales (2)		2674	27	0.01	-0.50
	Cyanobacteria (1)		9296	699	0.08	1.93
	Deinococci (1)		3182	11	0.00	-0.75
	Aquificales (1)		1560	7	0.00	-0.71
	Thermotogales (1)		1858	52	0.03	0.17

*Relates only to species from the 85 species participating in the analysis
**Query-list highly related enzymes – enzymes with more than 80% identity to a protein in the query list
*** Mean – mean value of the enzymatic fraction in the three domains or in the bacterial subfamilies. Within domains: mean = 0.037, standard deviation = 0.023; within bacterial subfamilies: mean = 0.023, standard deviation = 0.027.

2.4 The size and fraction of the enzymatic set in different species

2.4.1 The full complement of enzymes in species

To identify the fraction of proteins which are enzymes we started from the complete list of highly curated enzymes in SWISS-PROT (Boeckmann et al. 2003) and performed a PSI-BLAST (Altschul et al. 1997) search against 85 fully-sequenced genomes: 63 bacteria, 16 archaea and 6 eukaryota species. That is – we infer enzyme function if the sequence is sufficiently similar to one of the validated enzymes in the query list. To explore the consequences of this assumption, for every species in the analysis we define three sets of proteins:

The conservative set: all proteins in a species matching an enzyme from the query list with more than 40% identity. Previous studies have shown that enzymes exhibiting more than 40% sequence identity share in most cases the same function (Todd et al. 2001). Therefore, using a 40% identity cut-off enables us to transfer the functional annotation from the query protein to its hits with reasonable confidence.

The permissive set: all proteins recognising an enzyme from the query list after three PSI-BLAST iterations with an E-value cut-off of 10-3. The use of PSI-BLAST identifies more distantly related homologues, often with low sequence identity (< 20%). Such distant relatives have often evolved new functions (Todd et al. 2001), these sets will include enzymes and non-enzymes.

The KEGG database predicted species' enzymatic set. The KEGG (Kyoto Encyclopaedia of Genes and Genomes) database aims to contain the complete information available for functionally annotated enzymes in fully sequenced genomes (Kanehisa and Goto 2000; Kanehisa et al. 2002). The annotations are retrieved from well-established sources including SWISS-PROT, GenBank and the original genome projects. KEGG also introduces its own predictions based on orthologous relationships.

In Figure 3, the sizes of these different sets were plotted against proteome size. For prokaryote species all three estimates indicate that the expansion of enzymes correlates with proteome expansion (Table 3). A linear model accurately describes the expansion of the

permissive sets as indicated by a correlation coefficient of R2 = 0.98. The correlation coefficients for the KEGG sets and the conservative sets are lower, but still significant at 0.90 and 0.79 respectively. The actual protein composition of the conservative and the KEGG sets in *E.coli* almost completely overlaps. In addition, the reaction composition (i.e., the repertoire of EC reactions) is more than 80% identical for 74 out of 80 prokaryote species and more than 75% identical for the remaining six species. We find that the KEGG prediction for the number of enzymes is usually an intermediate between the permissive prediction and the conservative prediction suggesting that the cut-off used for the conservative set is stricter than the one used for the KEGG assignments.

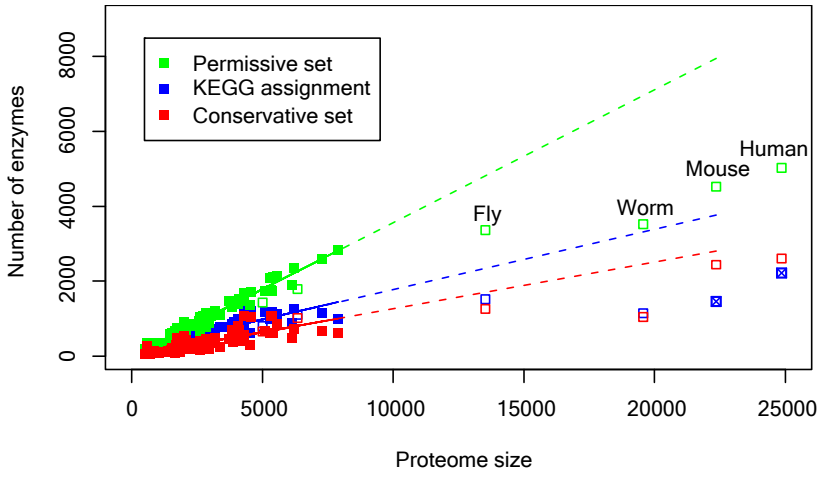

Figure 3 Number of enzymes per species versus proteome size. Full squares – prokaryote species, Empty squares – eukaryote species, Crossed squares – incomplete proteome for human and mouse in KEGG compared to these proteomes in the Biopendium™ (proteome size here is the one retrieved from the Biopendium™). The straight lines represent the regression line calculated for each enzyme set of the prokaryote species. The dotted lines are the extensions of the lines calculated for prokaryote species.

Table 3 Regression and correlation coefficients of different enzyme sets (Figure 3)

	Prokaryote species		Eukaryote species	
	Correlation coefficient (R^2)	Regression coefficient* ± Std. error	Correlation coefficient (R^2)	Regression coefficient* ± Std. error
Permissive set	0.98	0.36 ± 0.007	0.98	0.17 ± 0.017
KEGG assignments	0.90	0.16 ± 0.009	0.77	0.05 ± 0.019
Conservative set	0.79	0.13 ± 0.011	0.84	0.08 ± 0.026

*For a linear regression line (y = ax + b) the regression coefficient is the constant a; it is the slope of the regression line.

For the four multicellular species – worm, fly, mouse and human - the trend is less obvious and the low number of analysed species prohibits definitive conclusions. If the regression line observed from prokaryote species is extended to eukaryote species (Figure 3), the number of enzymes in the multicellular metazoa is lower than predicted. This suggests a slower rate of expansion for the enzymatic sets, relative to the increase in proteome size in multicellular species. The decrease is especially obvious for the permissive set.

The expansion of the eukaryotic permissive and conservative sets appears to correlate with genome size, with the exception of the worm genome that has almost the same number of enzymes as the smaller fly genome. The reasons for this might be specific expansions of non-enzymatyic protein families in C.elegans, such as the nematode-specific expansions of the G-protein-coupled receptor (GPCR) superfamily (O'Halloran et al. 2006). A decrease in the number of enzymes in worm compared to fly can also be observed in the KEGG sets. The low number of mouse and human enzymes in KEGG might be due in part to the partial representation of the mouse and human proteomes in the KEGG database. The total number of mouse and human proteins in KEGG (downloaded in February 2004) is approximately 2/3 of the current ENSEMBL estimate (August 2004), so the apparent low values are misleading.

2.4.2 Comparison of the permissive and conservative enzymes

The expansion patterns of the conservative set and the KEGG set as measured by the slope of the regression line are similar (Table 3) and probably reflect the similarity between the two annotation procedures. In contrast, the expansion of the permissive set is distinguishable from the former two and can be better described by a linear model (higher correlation coefficient). The main difference between the sets is that whilst the first two sets only include close relatives to proteins annotated in SWISS-PROT, the permissive enzyme sets also include more distant relatives (whose function may have diverged). We wanted to examine how the composition of the original query set influences the final hit list in different species under different cut-offs. For each species we plotted Nc (number of proteins in the conservative set) against Np (number of proteins in the permissive set, Figure 4A) and Fc (fraction of the conservative set) against Fp (fraction of the permissive set, Figure 4B). As in

all species Fp is bigger than Fc ($\overline{Fp}$ = 0.35, $\overline{Fc}$ = 0.13) we normalised the relative fraction of each set by comparing it to the mean fraction in all species. For each species we calculated the z-score (the standard deviation - σ - distance from the mean): $(Fp - \overline{Fp})/\sigma p$ (σp = 0.058) and $(Fc - \overline{Fc})/\sigma c$ (σc = 0.060) (Figure 4C, Table 1). The diagonal line represents an equal relative fraction of enzymes in the conservative and the permissive sets. In *Chlamydophila pneumoniae*, for example, the conservative set covers about 10% and the permissive set covers about 30% of the genome, which corresponds to similar z-score ~0.5 for the two sets (Table 1).

Species plotted above the diagonal line are those in which the relative fraction of enzymes was higher in the conservative set than in the permissive set. Most species that lie well above the diagonal are eukaryota and gammaproteobacteria (Figure 4C). Eukaryota and gammaproteobacteria are the most widely studied species as indicated by the composition of our query list (Table 2; see section 2.3.4 for further explanation about determination of group's relative representation in the query list). These species are over-represented in the query list. Archaea species on the other hand almost always lie under the diagonal line, possibly due to their low representation in the experimental data. This suggests that the permissive set is less sensitive to the biases in our experimental knowledge, which may explain the improved linear correlation for this set (Figure 3). The permissive set was used to calculate enzymatic fraction, but functional analysis was confined to the conservative set.

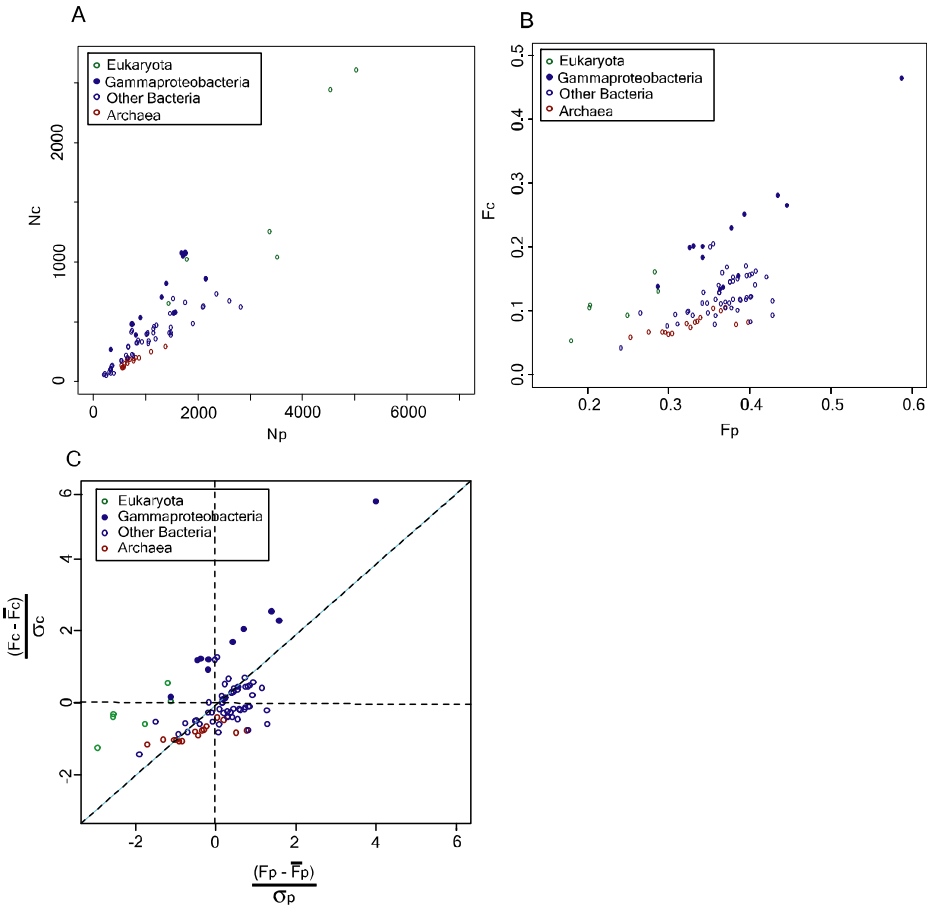

Figure 4 (A) The number of enzymes in a species calculated using different cut-offs. (B) The fraction of enzymes in a species calculated using different cut-offs. (C) The standard deviation from the mean value for the genomic fraction of enzymes in all 85 species examined (z-score). Np, Nc: the number of enzymes in the permissive and the conservative set, respectively; Fp, Fc: the fraction of the permissive and the conservative set, respectively, in each genome; $\overline{Fp}$, $\overline{Fc}$: the mean fraction of the permissive and the conservative set, respectively, in all 85 species; σp, σc: standard deviation of the permissive and the conservative set, respectively.

2.4.3 The enzyme fraction of the genome in different species and domains

Since the permissive enzymatic sets are less biased towards the common model organisms, these sets were used to calculate the enzymatic fraction (Fe) in different species. The distribution of Fp for species from the three domains of life is shown in Figure 5. $\overline{Fp}$ is the average Fp of all 85 species in the analysis. In most archaeal and bacterial species (67 out of 79 species) enzymes and enzymes-related proteins occupy approximately 30 – 40% of the genome and lie within one standard deviation of the mean value.

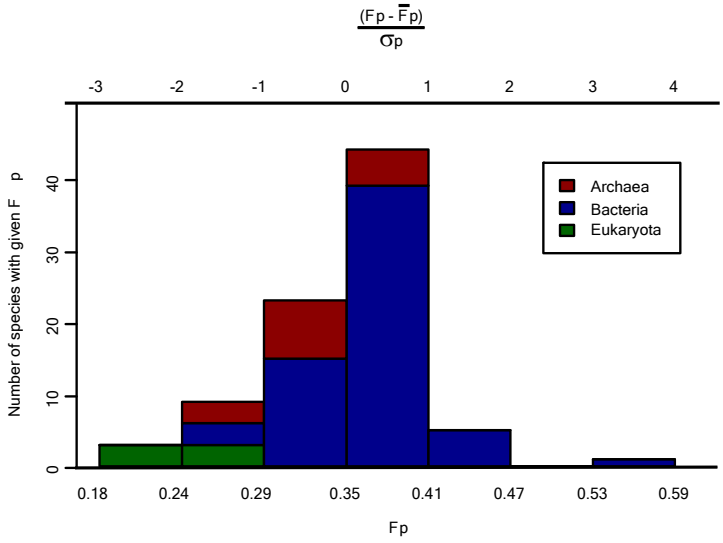

Figure 5 Distribution of the standard deviation from the mean of the fraction of enzymes in 85 species. The values calculated are derived from the permissive enzymatic sets. Eukaryote species: $\overline{F}$ = 0.23, σ = 0.05; Bacteria species: $\overline{F}$ = 0.37, σ = 0.05; Archaea species: $\overline{F}$ = 0.33, σ = 0.04.

All six species in which the fraction of enzymes is significantly higher than the mean (more than 1 standard deviation) are bacteria. These bacteria are phyletically diverse and include one alphaproteobacteria (*Rickettsia prowazekii*), three gammaproteobacteria (*Haemophilus influenzae, Pasteurella multocida, Buchnera*) and two mollicutes (*Mycoplasma genitalium, Mycoplasma pulmonis*). Five are pathogenic bacteria and one is a symbiont (*Buchnera*). Four (*R.prowazekii, Buchnera, M.genitalium, M.pulmonis*) are intracellular, obligatory pathogens or symbionts with a small genome. Intracellular pathogens and endosymbionts bacteria were previously shown to have a relatively high fraction of enzymes (Cases et al. 2003). The high metabolic fraction was suggested to be the result of a trend, which occurred in many species independently, in favour of a massive loss of regulatory proteins in intracellular species functioning in a relatively stable environment (Cases et al. 2003). Phyletic diversity can also be observed in the six prokaryote-species with a small fraction of enzymes (more then 1 standard deviation below the mean). These include three bacteria (alphaproteobacteria, gammaproteobacteria and a spirochaetales) and three archaea-species (one euryarchaeota and two thermoprotei species). The phyletic diversity of these extreme groups supports the notion that the metabolism of an organism better reflects its specific adaptation rather than its phyletic history (Aguilar et al. 2004; Cases et al. 2003) .

All eukaryota species in this sample have a relatively low fraction of enzymes (Figure 5) occupying 18 – 29% of their genome (Table 1). The two yeast species and the fly are between one to two standard deviations from the mean. The lowest fraction of enzymes was recorded for three of the metazoan-species: worm, mouse and human, where enzymes constitute only 18 – 20% of the genome. This observation is compatible with previous studies showing that the fraction of proteins involved in metabolism decreases when species complexity grows (Andrade et al. 1999; Tamames et al. 1996).

We observe here that the increase in the number of enzymes and enzyme-related proteins is correlated with proteome expansion in most prokaryote species (Figure 3). The fraction of enzymes and enzyme-related proteins is approximately constant and ranges between 30 to 40%. The trend is observed in both archaea and bacteria species, where extreme values seem to reflect species-specific adaptations rather than a phyletic trend. For example, a relatively high fraction of enzymes was found in species that had a massive loss of regulatory proteins. In contrast, a relatively low fraction of enzymes was found in eukaryote species – a lineage whose phylogenesis involved a massive recruitment of regulatory proteins (Andrade et al. 1999; Aravind and Subramanian 1999; Chervitz et al. 1998). Therefore, compatible with studies preformed within domains (Ranea et al. 2004; van Nimwegen 2003), the observations from this analysis - exploring together Archaea, Bacteria, and Eukaryota - suggest that the rate of enzymes expansion in a species is approximately constant, and that differences in the fraction of enzymes mainly reflect dramatic changes in the size of other functional categories.

2.5 Enzymes recrutiment and functional diversification

2.5.1 The use of the EC scheme

There are two explanations for the expansion of the number of enzymes as proteome size increases: either a larger variety of reactions has evolved or there are more proteins catalysing the same reactions (isoenzymes), with other differences, such as the control of expression, driving their evolution and retention. In order to estimate the level of functional diversification accompanying gene recruitment, we used the Enzyme Commission (EC) scheme (Tipton and Boyce 2000) in which each reaction has been assigned a unique

four digit identifier. Enzymatic reactions are divided into six classes represented by the first digit in the EC number: oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. The second digit refers to the subclass, which generally contains information about the type of compound or group involved in the reaction. This subclass definition differs between classes but in most cases it describes the donor group. The third digit further specifies the type of reaction involved, in most cases describing the type of acceptor involved. The fourth digit is a serial number that is usually used to identify the substrate for individual enzymes within a sub-subclass (Tipton and Boyce 2000).

One of the drawbacks in using the EC scheme for determining functional divergence is that some EC numbers describe generic reactions where the compounds are not fully specified. For example, all protein tyrosine-kinases are classified as EC 2.7.1.112 whilst all proteases lie in EC 3.4 but have multiple third and fourth digit depending on their catalytic mechanism and target (Barrett 1994; Tipton and Boyce 2000). However, the focus of the EC scheme on reaction type rather than mechanism (Tipton and Boyce 2000) is in most cases advantageous for this kind of study.

Here, in order to estimate the level of functional diversity accompanying gene recrutment we plotted the number of enzymes in a species against the number of reactions (number of distinct EC numbers assigned in a species). The reaction ratio, R, equals the total number of reactions divided by the number of enzymes in the organism. If R equals 1 then the increase in the enzyme number is entirely due to having more reactions. A smaller ratio implies an increase in the number of enzymes performing the same reaction.

Since function has been experimentally determined for only a small number of proteins (Andrade and Sander 1997), calculating the number of reactions per species requires inferring the function from sequence. Several recent analyses have used EC numbers to study how function changes as homologues diverge (Devos and Valencia 2000; Todd et al. 2001; Wilson et al. 2000). For single and multi domain proteins, variation in the EC number was found to be rare above 40% sequence identity (Todd et al. 2001). Here, based on this observation, we estimated the number of reactions in the conservative sets by transferring the EC number from a protein in the query list to all hits sharing more than 40% sequence identity. For each species we counted the number of distinct EC reactions. As using functional assignments inferred from sequence can at best only provide an estimation of the

number of reactions, we also examined the KEGG assignments obtained by considering orthologous relationships (Kanehisa and Goto 2000; Kanehisa et al. 2002).

The reaction ratio for each organism was calculated, counting both the number of distinct EC reactions down to the fourth level (R4) and distinct EC reactions down to the third level (R3).

2.5.2 Functional diversty of different species and domains

The two functionally annotated data sources (KEGG and the conservative set) provide similar observations regarding the diversification of enzyme function (Figure 6A, B), showing that eukaryotes and prokaryotes species differ in their expansion pattern. For prokaryotes, plotting the number of reactions against the number of enzymes shows that the two sets are essentially identical, both with a correlation coefficient ~ 0.99 (Table 4). The regression line obtained (for prokaryotic organisms only, conservative set – Figure 6A) is given by the following equation:

Number of reactions = 0.78 * Number of Enzymes + 24

Table 4 Regression and correlation coefficients of the distribution of the number of different reactions against the number of enzymes.

	Prokaryote species		Eukaryote species	
	Correlation coefficient (R^2)	Regression coefficient ± Std. error	Correlation coefficient (R^2)	Regression coefficient ± Std. error
KEGG assignments (4 digits)*	0.98	0.64±0.016	0.85	0.32±0.100
Conservative set (4 digits) **	0.99	0.78±0.011	0.99	0.24±0.019
Oxidoreductases (1) ***	0.98	0.72 ±0.016	0.99	0.26±0.013
Transferases (2) ***	0.99	0.76±0.014	0.95	0.17±0.028
Hydrolases (3) ***	0.97	0.76±0.020	0.99	0.30±0.020
Lyases (4) ***	0.98	0.74±0.015	0.58	0.24±0.166
Isomerases (5) ***	0.98	0.67±0.014	0.96	0.12±0.018
Ligases (6) ***	0.99	0.85±0.017	0.83	0.22±0.075

* Figure 6B; ** Figure 6A; ***

Figure 7;

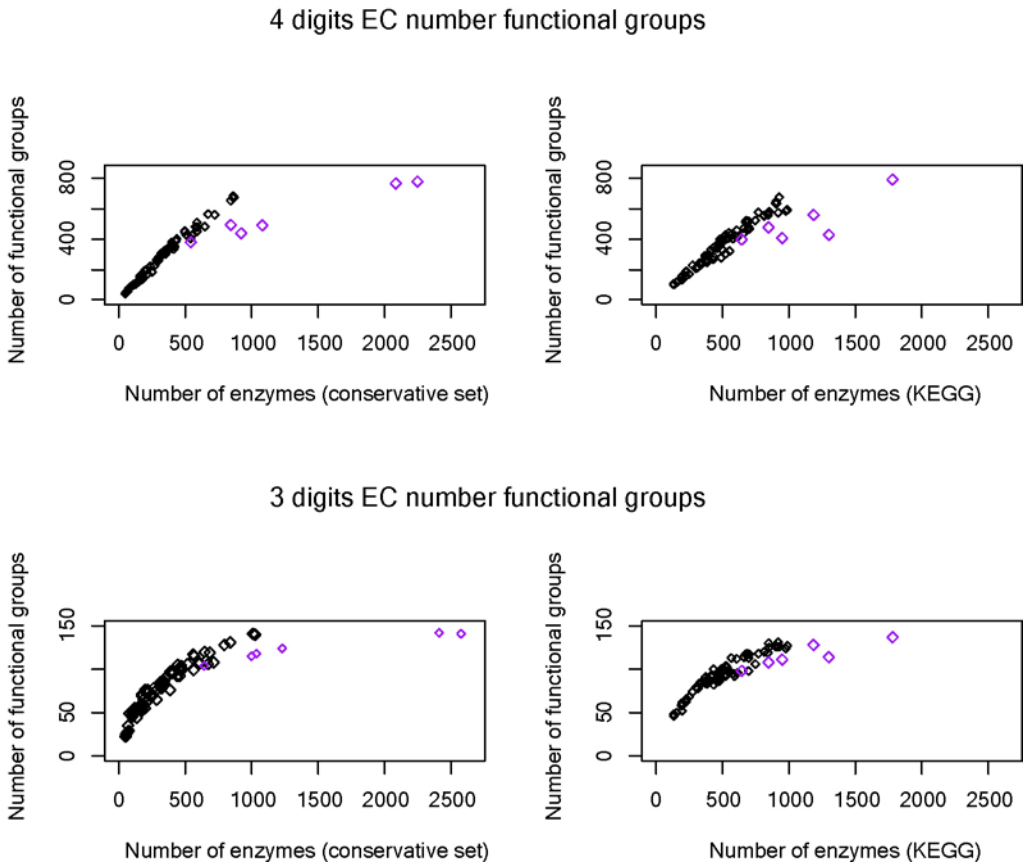

Figure 6 Number of reactions versus number of enzymes. Reactions are described either by four digit assigned EC numbers (A, B) or the three digit assigned EC numbers (C, D). (A, C) Number of enzymes per species is the number of proteins with a full EC assignment (A) or a 3-digit EC assignment (C). The EC assignments are inferred for proteins sharing 40% sequence identity with a highly curated enzyme from the query list. (B, D) Number of enzymes is the number of proteins assigned by KEGG with a full EC number. Black squares – prokaryote species; purple squares – eukaryote species. The straight lines represent the regression line calculated for each set of the prokaryote species. Total number of reactions refers to the number of different reactions in all 85 species examined.

Therefore the number of reactions is approximately 4/5 the number of enzymes, i.e., most enzymes perform a single reaction and most reactions are performed by a single enzyme. From the equation, equality between the number of enzymes and the number of reactions is achieved for approximately 110 enzymes. A hypothetical species with 110 enzymes is expected to be able to catalyse 110 reactions without having a single isoenzyme. Below 110 enzymes the number of reactions is predicted to exceed the number of enzymes due to multi-functional enzymes. In Chlamydia muridarum for example, the number of enzymes in the conservative set is 88, the number of reactions predicted from the equation is 92, and the actual number of reactions is 89.

The linear dependency between the number of enzymes and reactions found for prokaryotic species does not apply for eukaryotic genomes, as the number of reactions in eukaryotic species is lower than that predicted by the prokaryotic regression line, indicating a higher functional redundancy. This observation is compatible with previous studies reporting a general trend, where higher organisms have larger sizes for many of their protein families (Enright et al. 2003; Krakauer and Plotkin 2002; Muller et al. 2002; Tatusov et al. 1997). While both data sets agree, the KEGG assignments suggest a slightly lower increase in reaction ratio in eukaryotic species, probably because of the requirement for an orthologous relationship for functional inference. Thus in prokaryotic species the expansion in the number of enzymes mainly reflects the broadening of the reaction repertoire, whilst in eukaryotes the expansion is to a greater extent the result of increased reaction redundancy.

When studying the distribution of R3 reactions (reactions detailed down to the third level), a plateau is observed after reaching an enzymatic set size of around 500 proteins (Figure 6C, D). The number of functional groups in prokaryote species is similar to the number in similar size eukaryotic species (size relates to estimated number of enzymes). The plateau is the result of the low functional diversity between species with regard to the three-digit EC number reactions. True to August 2004, the EC scheme contains 4327 known R4 reactions, which map to only 236 R3 and 63 R2 reactions. The conservative data set includes 1805 R4 reactions that are mapped to 187 R3 reactions. As none of the species examined has more than 800 R4 reactions, there is still diversity in the reaction composition between species, while such diversity is unlikely for the R3 reactions, where species only have up to 142 reactions. Human (777 R4 reactions, 141 R3 reactions) and E.coli (682 R4 reactions, 140 R3 reactions) for example, share only about 40% (on average) of their R4 reactions versus 80% of their R3 reactions. Even species with a low number of enzymes share a very similar set of reactions (not shown). Beyond approximately 500 enzymes, species contain almost all R3 reactions and functional diversification is achieved by diversification in the R4 EC number reaction composition. It is important to remember however that the inference of function herein is conservative; there is no doubt that in reality most species are likely to have more reactions than predicted here.

2.5.3 Functional diversity of different reaction classes

We examined whether the increased functional redundancy observed in eukaryote species merely reflects the massive recruitment of enzymes in a limited number of broadly defined reaction types (e.g., protein tyrosine-kinases) or whether it reflects a more general trend. Firstly, the enzymes were divided into the six reaction classes of the EC scheme and the number of enzymes and reactions (R4) in each class was plotted against the size of the proteome (Figure 7). Since the results are essentially the same for the KEGG and the conservative data sets, only the conservative set is presented and the data in Figure 7 are derived from the data in Figure 6A. The plot shows that the enzymes are unequally distributed between the six classes and their rates of expansion are radically different. A massive expansion of the transferases and hydrolases functional classes can be observed in metazoa (human, mouse, worm and fly).

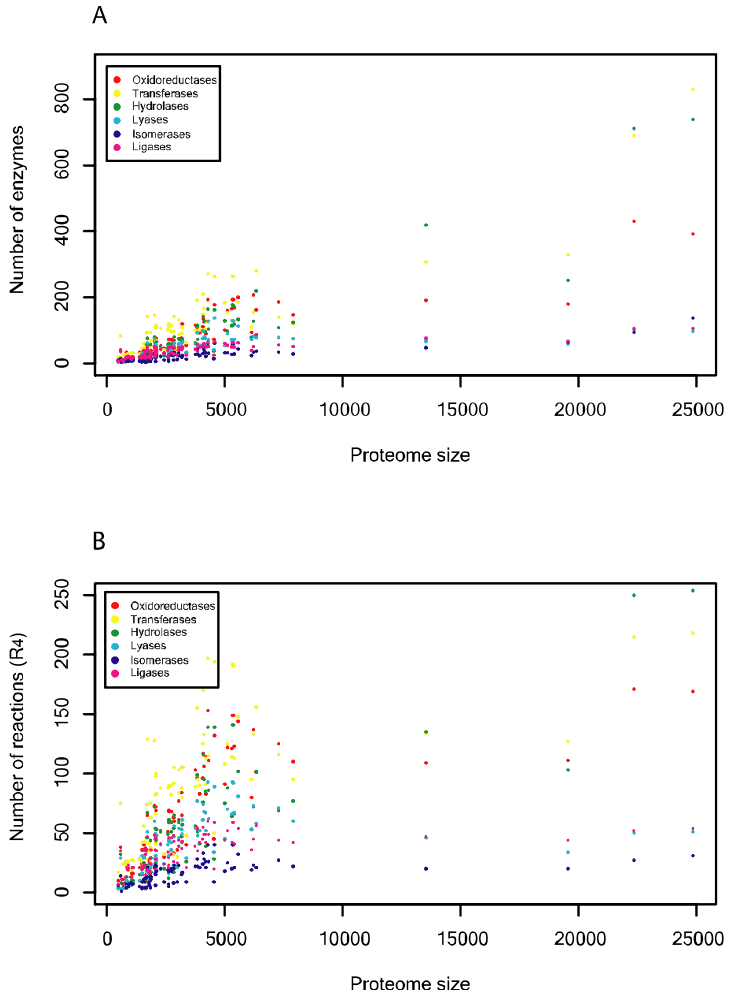

Figure 7 Number of enzymes (A) and number of R_4 reactions (B) against proteome size. The number of reactions and number of enzymes are derived from the conservative set.

Next, the number of reactions in each class was plotted against the number of enzymes in that class (i.e., the number of oxidoreductase reactions in a species against the number of oxidoreductase enzymes in that species) (Figure 8). In prokaryotes, oxidoreductases, transferases and hydrolases seem to be more abundant than lyases, isomerases and ligases. For all functional classes a clear linear dependency exists between the number of enzymes and the number of reactions with a correlation coefficient ranging from 0.97 to 0.98. The slope of the linear regression line is very similar in all plots and ranges between 0.72 and 0.76 in oxidoreductases, transferases, hydrolases and lyases. The slowest rate of functional diversification (regression coefficient of 0.67) was observed for isomerases and the highest rate (regression coefficient 0.85) was observed for ligases.

In eukaryotic species the correlation coefficient is lower in most reaction classes, compared to prokaryote species. Unlike prokaryote species, the distribution pattern is not constant and differs between the reaction classes. The rate of expansion of reactions varies and ranges from 0.12 (isomerases) to 0.30 (hydrolases). In one of the reaction classes (lyases) there is no clear linear dependency between the number of reactions and the number of enzymes (correlation coefficient = 0.58). In three reaction classes (oxidoreductases, transferases, and hydrolases) there is a massive recruitment of enzymes in mammals although there is no significant increase in the number of reactions. A more moderate increase in the number of mammalian enzymes is observed for the isomerase and ligase reaction classes where the number of reactions is again smaller than the one predicted by the regression line describing the distribution in prokaryote species. Lyases are the only reaction class with a smaller number of enzymes in mammals than in prokaryote species. Other eukaryote species also exhibit an increase in functional redundancy compared to prokaryote species, although more moderate than in mammals. The yeast Schizosaccharomyces pombe is the least functionally redundant of the eukaryote species examined.

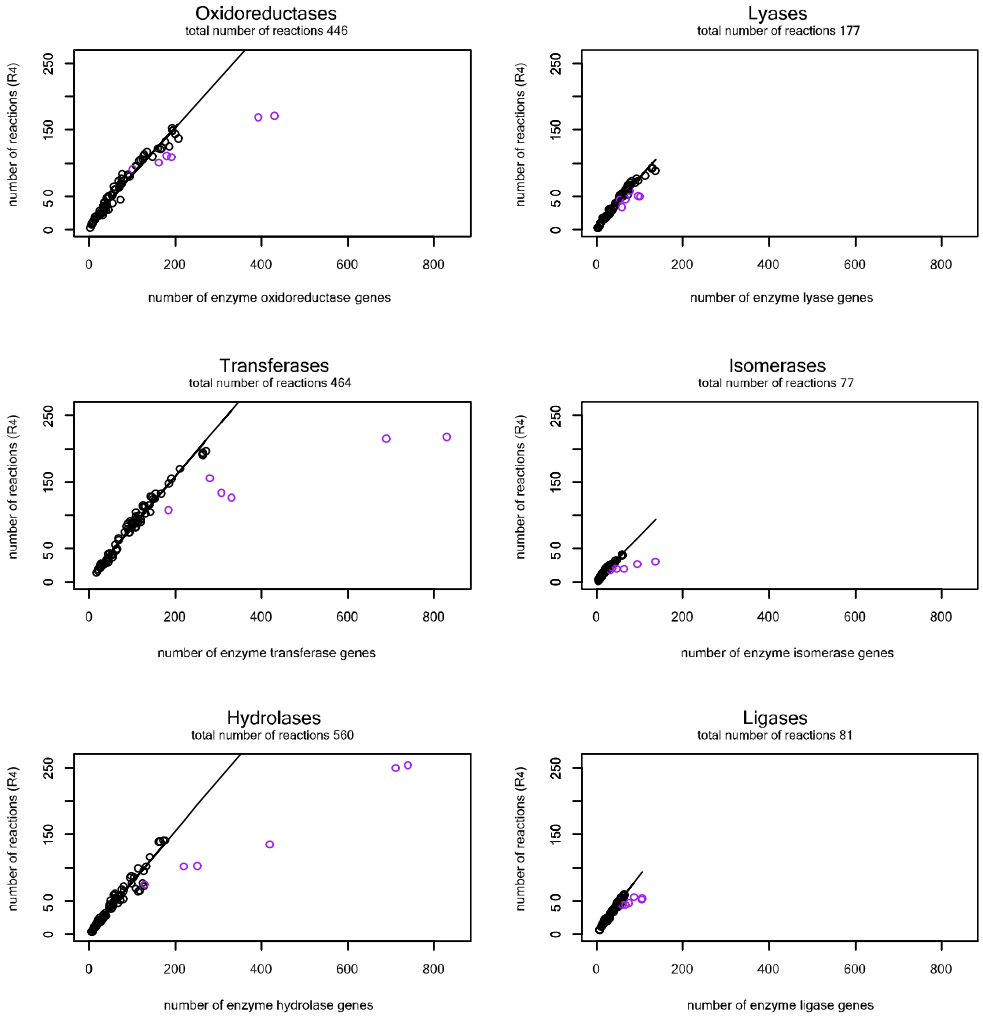

Figure 8 Number of R_4 reactions versus the number of enzymes assigned to the reaction class. The number of reactions and number of enzymes are derived from the conservative set. The straight lines represent the regression line calculated for each set of the prokaryote species. Total number of reactions refers to the number of different reactions in all 85 species examined.

2.5.4 The extent of functional redundancy – how many reactions in a species are redundant?

In order to quantify how general the functional redundancy is (e.g., how many reactions per species have multiple enzymes), we studied the distribution of the number of enzymes per reaction in five species representing different levels of complexity (Figure 9). The species examined, ordered from the most to least complex, are the multicellular metazoa human and fly, Saccharomyces cerevisiae (unicellular eukaryote), and two prokaryote species – the free living Escherichia coli and the intracellular obligate symbiont Buchnera. In all functional classes, the fraction of reactions with only a single enzyme assigned increases

when species' complexity decreases. Human has the biggest fraction of reactions that are redundant (58%), i.e., the same reaction is performed by more than a single enzyme, followed by fly (44%), yeast (37%), E.coli (25%) and Buchnera (10%).

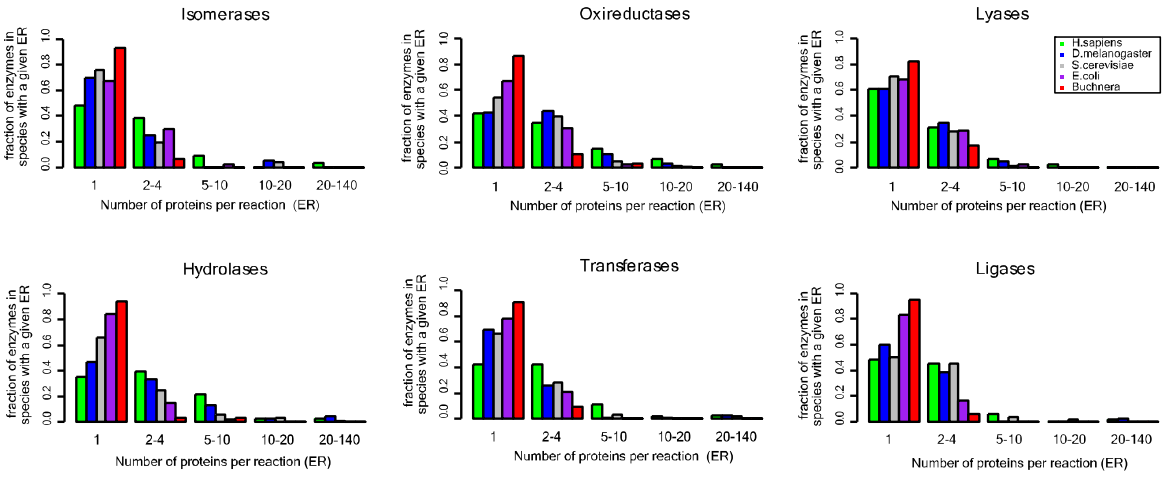

Figure 9 For each class, the fraction of all enzymes in the species with a given 'number of enzymes per reaction' (ER). The total number of enzymes (conservative set) is as follows: *H.sapians* – 2608, *D.melanogaster* – 1257, *S.cerevisiae* – 1023, *E.coli* – 1077, *Buchnera* – 267.

Buchnera, the small genome intracellular obligate symbiont, exhibits a lower functional redundancy compared to the free-living *E.coli*. Both *E.coli* and *Buchnera* belong to the Enterobacteriaceae subdivision of the gammaproteobacteria. The evolution of the *Buchnera* genome, as of many other obligate parasites and symbionts, involved a massive reduction in the size of the genome (Moran 2003). The number of genes in *Buchnera* (~500) is approximately 1/10 of the number of genes in the closely related *E.coli* genome. Almost every *Buchnera* gene has a clear orthologue in *E.coli*, indicating that *Buchnera* provides a good approximation of a minimal *E.coli* genome (in the context of an intracellular environment (Moran 2002)). Since almost all of its reactions seem to be assigned to a single gene product it seems that *Buchnera* has lost even the limited functional redundancy observed in *E.coli*. The lack of functional redundancy might be related to the fact that *Buchnera* is an intracellular species functioning in a relatively stable environment.

Increased functional redundancy is observed in species with increased complexity. In eukaryote species, and especially in metazoa where there are many different cell types and

environments, many reactions are catalysed by more than a single enzyme. The increase in functional redundancy might also reflect the lack of physical constraint on the size of the eukaryotic genome where the selective pressure to lose redundant genes is less strong than in bacterial genomes (Lynch and Conery 2003).

The most redundant reaction class in human is the hydrolase reaction class where 65% of the reactions are catalysed by more than a single enzyme, followed by the transferase and oxidoreductase functional classes where 58% of the reactions are redundant (Figure 9). These correspond to the classes, which are expanded massively (Figure 8). In the isomerase and ligase reaction classes, where one observes a moderate increase in the number of enzymes, 52% of the reactions are redundant. In the lyase reaction class, where the proteins/enzymes ratio in eukaryotes is very similar to the ratio in eukaryotes (Figure 8), only 39% of the reactions are redundant.

The reactions that are assigned more than 20 enzymes per species were listed in Table 5. Only 22 such reactions were identified for the five model species. For many of the reactions listed the high number of proteins assigned is due to a broad specificity of the EC reaction or to having several proteins working in concert as sub-units of a protein complex. Therefore the proteins in many of these reactions are not true isoenzymes, and these big reaction clusters do not usually represent a true functional redundancy. Yet, a comparison of the species distribution of the reactions is of interest, as in many cases it reflects unique species-specific or lineage-specific adaptations. For example, Microsomal P450 has 34 proteins assigned in human. Multicellularity was suggested to be a driving force for P450 duplication as it is a natural choice for making and degrading mammalian signalling molecules like retinoic acid, thromboxane A2, steroids, and ecdysone (Nelson 1999). Similarly 22 human proteins are assigned to cyclic-nucleotide phosphodiesterase. Cyclic-nucleotide phosphodiesterase is a regulator of the cAMP signalling pathway - a central pathway in learning and memory (Bolger et al. 1993). Two reactions that are massively expanded only in the fly – glutathione transferase and cholinesterase – are primarily responsible for metabolic resistance to insecticides (Ranson et al. 2002). Several transferase and hydrolase reactions were massively expanded in all eukaryotic genomes examined; many of them are reactions involved in signalling and degradation – functional classes that have been massively expanded in metazoa. Only a single reaction was differentially expanded in

bacteria – 30 proteins were assigned to the PEP-dependant phosphotransferase enzyme II in E.coli, a participant in the chemotactic pathway in motile bacteria (Lux et al. 1995).

Table 5 Reactions assigned to more than 20 proteins per species in one of the five model species in Figure 9.

Class	Enzyme	Enzyme name	Species	Number of assigned proteins	Description[20]
Oxidoreductases	1.6.5.3	NADH dehydrogenase (ubiquinone)	H.sapiens	46	Mitochondrial protein complex.
	1.6.99.3	NADH dehydrogenase (cytochrome c reductase)	H.sapiens	45	Protein complex.
	1.9.3.1	cytochrome-c oxidase	H.sapiens	21	Mitochondrial complex.
	1.14.14.1	unspecific monooxygenase, microsomal P450	H.sapiens	34	A group of heme-thiolate proteins (P-450), acting on a wide range of substrates.
Transferases	2.5.1.18	glutathione transferase	D.melanogaster	26	A group of enzymes with broad specificity
	2.7.1.27	erythritol kinase	S.cerevisiae	24	
			D.melanogaster	44	
			H.sapiens	76	
	2.7.1.37	protein kinase	S.cerevisiae	25	A broad specificity group of enzymes that are under review by the NC-IUBMB. Signalling.
			D.melanogaster	51	
			H.sapiens	132	
	2.7.1.69	PEP-dependant phosphotransferase enzyme II	E.coli	30	Comprises a group of related enzymes.
	2.7.1.112	protein-tyrosine kinase	D.melanogaster	28	The reaction includes all enzymes acting as protein-tyrosine kinases. All phosphorylated proteins, regardless of their function and nature, are commonly considered as the substrate in the reaction. Signalling.
			H.sapiens	84	
	2.7.7.6	DNA-directed DNA polymerase	S.cerevisiae	29	Protein complex
			H.sapiens	22	
	2.7.7.49	RNA-directed DNA polymerase	H.sapiens	115	Protein complex
Hydrolases	3.1.1.8	cholinesterase	D.melanogaster	24	Acts on a variety of choline esters and a few other compounds.
	3.1.2.15	ubiquitin thiolesterase	H.sapiens	27	Degradation.
	3.1.3.16	phosphoprotein phosphatase	D.melanogaster	22	A group of enzymes removing the serine- or threonine-bound phosphate group from a wide range of phosphoproteins. Signalling.
			H.sapiens	36	

	3.1.3.48	protein-tyrosine-phosphatase	H.sapiens	61	Dephosphorylates O-phosphotyrosine groups in phosphoproteins. Signalling.
	3.1.4.17	3',5'-cyclic-nucleotide phosphodiesterase	H.sapiens	22	Regulator of the cAMP signalling pathway[29]
	3.4.21.1	chymotrypsin	D.melanogaster	26	Broad specificity of substrates. Degradation.
			H.sapiens	52	
	3.4.21.4	trypsin	D.melanogaster	35	Broad specificity of substrates. Degradation.
			H.sapiens	37	
	3.4.25.1	proteasome endopeptidase complex	D.melanogaster	24	Degradation
	3.6.3.14	H+-transporting two-sector ATPase	S.cerevisiae	28	Mitochondrial complex.
			D.melanogaster	31	
			H.sapiens	42	
Isomerases	5.2.1.8	peptidylprolyl isomerase	H.sapiens	74	Broad specificity of substrates. Involved in protein folding.
Ligases	6.3.2.19	ubiquitin--protein ligase	D.melanogaster	22	All protein-lysine, are commonly considered as the substrate in the reaction. Degradation.
			H.sapiens	30	

2.6 Caveats

While performing such a large-scale analysis, several important limitations must be acknowledged:

2.6.1 The use of the EC scheme for estimating functional redundancy.

Proteins sharing the same EC number can be related in several ways. They may be proteins working in concert (e.g., in protein complexes), duplicated genes that are predicted to have the same function and genes that do not share sequence similarity but still catalyse the same reaction. The two last groups might represent specific adaptations to different regulatory modes. We wanted to verify that the increase in functional redundancy observed in higher species is indeed mostly the result of family expansion rather than the result of having more reactions that describe proteins working in concert. Therefore, the KEGG database was used in order to estimate the fraction of reactions that are assigned to more than one protein and the fraction of reactions that are assigned to more than a single orthologous cluster (Figure 10).

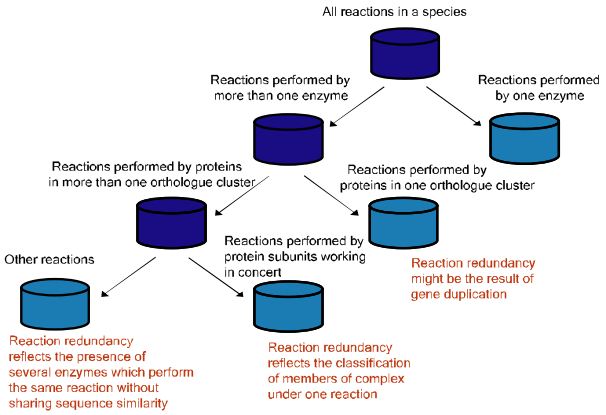

Figure 10 Description of possible ways in which enzymes that perform the same reaction in a species are related**.**

In 71% to 76% of the multi-protein reactions in the five eukaryotes examined (***H.sapiens, M.musculus, C.elegans, S.cerevisiae, S.pombe)*** all enzymes are classified into a single orthologue cluster, suggesting that the functional redundancy is, in most cases, the result of a gene duplication event (Table 6). A text search was done on those enzymes classified to multi-protein, multi-orthologue-cluster reactions to look for annotations indicating that they are part of a complex (e.g., complex, sub-unit, component, chain). The high fraction of multi-protein reactions that are classified to one orthologue cluster together with the low fraction of 'complex-reactions' in eukaryote species imply that the increase in functional redundancy observed is mainly the result of enhanced gene duplication, rather than an artefact caused by the increase in number of reactions representing proteins working in complex. We further tested the data in order to verify that we do not miss out proteins working in complex. The analysis, which is reported in Appendix B, indicates that the large majority of complexes are indeed identified.

Table 6 Distribution of different reaction groups in species. Number in brackets represents the % out of all reactions with multi enzymes. Data were extracted from the KEGG database (see material and methods). HSA - *H.sapiens,* MMU - *M.musculus,* CEL - *C.elegans,* SCE - *S.cerevisiae,* SPO - *S.pombe,* ECO - *E.coli,* BSU - *B.subtilis,* PAE - *P.aeruginosa,* STM - *S.typhimurium.*

		HSA	**MMU**	**CEL**	**SCE**	**SPO**	**ECO**	**BSU**	**PAE**	**STM**
All reactions		816	610	416	491	408	656	531	582	663
Reactions with one enzyme		540	428	270	334	296	513	393	400	525
Reactions with multi enzymes	one orthologue cluster	206 (74%)	131 (72%)	111 (76%)	119 (75%)	79 (70%)	61 (42%)	86 (62%)	114 (62%)	58 (42%)
	multi orthologue clusters where reactions do not describe proteins working in complex	49 (18%)	35 (19%)	17 (12%)	16 (10%)	10 (9%)	43 (30%)	22 (16%)	23 (13%)	39 (28%)
	multi orthologue clusters where reactions describe proteins working in complex	21 (8%)	16 (9%)	18 (12%)	22 (14%)	23 (21%)	39 (27%)	30 (22%)	45 (25%)	41 (30%)

In a similar way we analysed the distribution of various EC groups in the conservative set. The reaction distribution was studied in five species: ***H.sapiens***, ***D.melanogaster***, ***S.cerevisiae***, ***E.coli*** and ***Buchnera***. Clustering of proteins into sequence groups was done according to their similarity to the query protein which identified them, i.e., proteins that were recognised by the same query protein (40% identity, 80% overlap) are clustered together. Similar results to those observed from the KEGG database are obtained using the conservative set. Finally, we studied the sequence similarity between protein pairs in 156 reactions in human assigned to two proteins. 79% of the protein pairs share more than 40% sequence identity.

An additional challenge when using the EC scheme for determining functional divergence is that some EC numbers describe generic reactions where the compounds are not fully specified (e.g., protein-kinases). We verified that the increased functional redundancy observed in eukaryotic species reflects a general trend rather than lineage specific expansion of proteins assigned to a limited number of broadly defined reaction

types. Firstly, we show that increased functional redundancy is observed for most of the functional classes (i.e., not limited to protein-kinases or protein-peptidases). Secondly, we show that in most reaction classes extensive recruitment of enzymes is accompanied by a general increase in the number of redundant reactions (i.e., increase in the number of reactions to which more than a single enzyme is assigned). The above analyses demonstrate that, although the EC scheme is not ideal for studying functional redundancy in a few broadly defined reactions such as protein-kinases and peptidases, the functional redundancy reported in the analysis is much more general.

2.6.2 Inferring function from sequence.

As function has been experimentally proven for only a small number of proteins (Andrade and Sander 1997), calculating the number of reactions per species requires the inference of function from sequence. Here, based on previous studies of function-sequence dependency in enzymes (Todd et al. 2001), we transferred annotations from a query protein to hits sharing at least 40% sequence identity, when the query protein has an 80% overlap with the hit. Such annotation transfer has been found to be accurate in the large majority of examined enzyme-pairs, although not in all. As at the moment there are no genomes, which are experimentally annotated in full, we are forced to rely on sequence-based annotations. In a study of the sequence-function dependency in single and multi domain proteins Hegyi and Gerstein have shown that in multi-domain proteins, in a case of complete coverage along the full length of both proteins, function is conserved in 90% of the protein pairs (Hegyi and Gerstein 2001). We repeated our analysis while conditioning annotation transfer in mutual full coverage (80% of the length) between query and hit. The observations under the above limitation are consistent with the results obtained by conditioning annotation transfer only by full coverage (80%) of the query protein (results not shown). As shown, the results from the sequence-inferred database are also consistent with the results obtained when using the functionally annotated KEGG database (Kanehisa and Goto 2000; Kanehisa et al. 2002).

2.6.3 The scarcity of fully sequenced eukaryote species.

As only a limited number of eukaryote and metazoa genomes are available to date, and this number was even more limited while performing the analysis (2004), it is not yet clear how general are the trends identified here. Future sequencing of additional unicellular and multicellular eukaryote species will provide a better understanding of the extent to which differences between eukaryote and prokaryote species can be related to the transition from the prokaryote to the eukaryote cell or to the transition from unicellularity to multicellularity.

2.7 Conclusions and Discussion

The chapter describes a comprehensive analysis of the complement of enzymes in a large variety of species. We aimed to estimate the fraction of enzymes in genomes and to determine the extent of their functional redundancy in different domains of life.

A few major trends emerge from the analysis:

- **Enzymatic fraction:** The fraction of enzyme-related proteins is approximately constant and ranges between 30 to 40% of the genome in most prokaryote species. The trend observed is common to both archaea and bacteria species where extreme values seem to reflect species specific adaptations rather than a phyletic trend. The relatively low fraction of enzymes found in eukaryote species might be related to a massive recruitment of regulatory proteins (Andrade et al. 1999; Aravind and Subramanian 1999; Chervitz et al. 1998). We therefore suggest that the rate of enzyme expansion in all domains is approximately constant, and that the differences in the fraction of enzymes mainly reflect dramatic changes in the size of other functional categories.
- **Increased functional redundancy in eukaryote species:** In eukaryotic species the enzymes/reactions ratio is higher than in prokaryotic species and therefore the functional redundancy in these species must increase. Whilst enzymatic sets grow when proteome size increases in all species, eukaryotic and prokaryotic species differ in their pattern of expansion. In prokaryotic species the expansion of the enzymatic set mainly reflects the broadening of the reaction repertoire, whilst the expansion of

the eukaryotic set, and especially of multicellular eukaryotes, is to a larger extent the result of an increase in reaction redundancy. An increased sequence redundancy accompanying an increase in the number of genes was previously reported (Enright et al. 2003; Krakauer and Plotkin 2002; Muller et al. 2002; Tatusov et al. 1997). The quantitative assessment performed here for the enzymatic set of each species, confirms and quantifies this general trend. Whereas 58% of the enzymatic reactions in human are found to be redundant, less than 10% of the enzymatic reactions in the intracellular symbiont bacteria *Buchnera* are performed by more than a single enzyme.

Two fundamental differences between unicellular prokaryotes and multicellular eukaryotes can contribute to different adaptation strategies regarding the addition of new enzymes. First, genomes of eukaryotic species lack physical constraints on their size and therefore the selective pressure to lose redundant genes is less efficient than that in bacterial genomes (Lynch and Conery 2003). Second, unlike unicellular species, cells of multicellular species function in a relatively stable environment and their metabolic diversity is therefore more likely to represent spatial adaptations.

This chapter concerns the relationship between the expansion of a functional group (enzymes) and changes in the size and composition of its functional repertoire. The results suggest that the recruitment of new enzymes in metazoa mostly contribute to an increase in functional redundancy. The next chapter concerns the relationship between family expansion and mammalian spatial complexity.

Chapter 3 Relationship between duplication events and differentiation of expression in mouse tissues

3.1 Overview

Gene duplications have been hypothesized to be a major factor in enabling the evolution of tissue differentiation. This chapter concerns the relationship between duplication events, the time they took place and the expression breadth of not only the duplicated genes but also the cumulative expression breadth of the gene family to which they belong. The analyses described in this chapter indicate that only duplicates that arose through post-multicellularity duplication events show a tendency to become more specifically expressed, while such a tendency is not observed for duplicates that arose in a unicellular ancestor. Unlike the narrow expression profile of the duplicated genes, the overall expression of gene families tends to maintain a global expression pattern. The work presented here supports the view suggested by the subfunctionalization model, that expression divergence, following gene duplication, promotes the retention of a gene in the genome. The global expression profile of the gene families suggests the division of expression between the family members, whose expression becomes specialized. As specialization of expression is coupled with an increase rate of sequence divergence it can facilitate the evolution of new, tissue-specific, functions.

This work has been published in (Freilich et al. 2006).

3.2 Background

Gene duplication events were long ago suggested to contribute to the attainment of the complex body organisation in metazoan species (Ohno 1970). A possible mechanism through which gene duplication can contribute to tissue differentiation is described in the recent model of subfunctionalization (Lynch and Force 2000). According to this model, two daughter genes can accumulate degenerative mutations resulting in the division of the ancestral function, and hence promote the retention of both duplicate copies in the genome.

Division of the expression of the ancestral gene between its daughter duplicates, through the accumulation of mutations in the promoter region is one mode of function division. Several examples for subfunctionalization of expression were reported for individual genes (Adams et al. 2003; Force et al. 1999; Prince and Pickett 2002). The findings from several studies that use microarray expression information to explore several aspects of the relationship between gene duplication and expression divergence are consistent with these predicted from the subfunctionalization model. Expression divergence between duplicate genes was shown to increase with evolutionary time when studying both temporal (differentiation modes in yeast, (Gu et al. 2002)) and spatial (human tissues - Makova and Li 2003; plant tissues - (Blanc and Wolfe 2004)) expression divergence patterns, where the divergence of expression occurs relatively shortly after the duplication event. Duplication events of mammalian genes tend to lead towards a tissue-specific expression pattern (Huminiecki and Wolfe 2004).

By using expression information from various mouse tissues, we explore several aspects of the relationship between duplication events and specialisation of expression that have not yet been characterised. Firstly, we studied the relationship between duplication events and the expression breadth of the duplicated gene. Previous analysis has shown general trend for increased tissue-specificity as family size increases (Huminiecki and Wolfe 2004). But, since both tissue-specific expression of a gene and the presence of closely related duplicate genes have been independently demonstrated to be associated with a relatively high divergence rate (Duret and Mouchiroud 2000; Kondrashov et al. 2002; Winter et al. 2004; Zhang and Li 2004), we verified that a relationship between expression breadth of a gene and its number of duplicates is not simply derived from the mutual relationship of expression breadth and number of duplicates with the divergence rate. Secondly, we explored the relationship between duplication events and expression breadth of the duplicated genes in the context of the time when a duplication event took place (Figure 11A). As spatial expression divergence following gene duplication is more likely in a tissue-differentiated environment, duplication events that took place in the unicellular ancestor are likely to affect expression breadth differently than duplication events that took place after the transition to multicellularity. Finally, we explored the relationship between duplication events and the cumulative expression breadth of the duplicated gene family (Figure 11B).

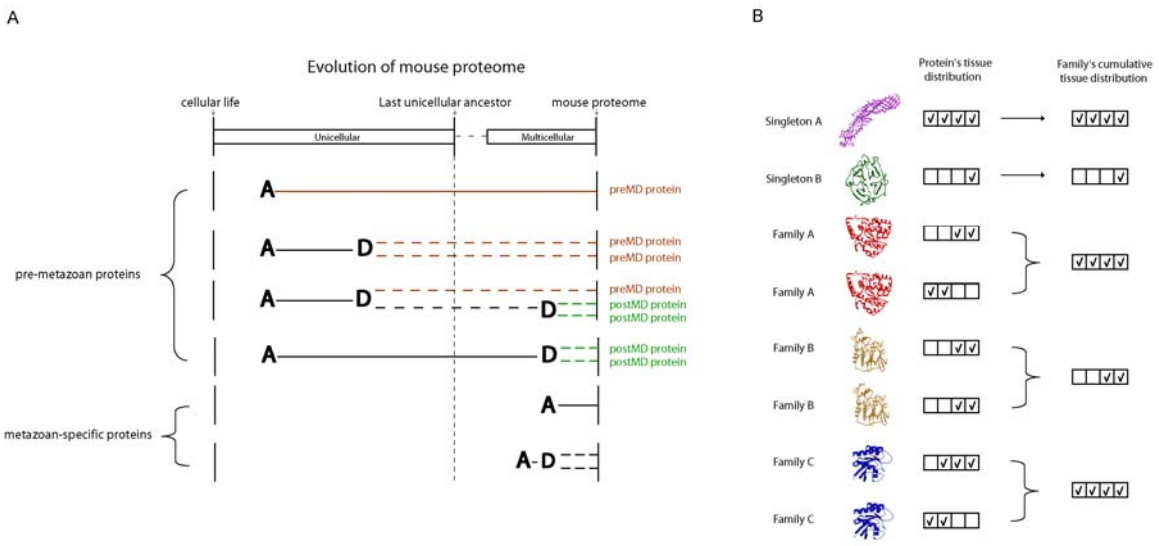

Figure 11 A schematic illustration of concepts described in the text. (A) Proteins illustrating different aspects of phyletic age/time of duplication in the mouse proteome, when the calibrated time is the transition from unicellularity to multicellularity. A – Appearance of a protein in the mouse proteome. D – Duplication event, leading to the retention of both copies in the mouse proteome. The appearance of a novel protein relates to events where protein contains a novel combination of domains or to events where a protein sequence was changed beyond the recognition of traditional sequence search algorithms and therefore there is a high likelihood that the protein performs a new function. Pre-metazoan mouse proteins are proteins that have descended from a protein present in the unicellular ancestor of mouse; metazoan-specific proteins are proteins that are unique to the multicellular lineage of metazoa. As all duplications of metazoan-specific proteins are bound to take place after the transition to multicellularity, proteins from this group are not classified into groups of time of duplication (preMD/postMD). (B) Building a cumulative expression profile for protein families. The cumulative expression profile of each family was built by recording all tissues in which at least a single family-member is expressed. Singleton proteins, by definition, are single-member families and the cumulative distribution is identical to the protein distribution. Family A is an example of complementary expression with no expression overlap between the duplicate proteins; Family B is an example of identical expression; Family C is an example of complementary expression with partial expression overlap. The protein cartoons used in this figure are only illustrative.

3.2 General description of the data

We identified pairs of close homologous mouse proteins by performing an all-against-all BLAST search (Altschul et al. 1997) for the entire proteome of mouse (detailed in section 3.3.5). These protein pairs are termed here duplicate-pairs as the genes which encode them are likely to have arisen through the duplication of a common ancestor gene. For each protein we counted the number of its duplicate-pairs. Out of 31,535 mouse proteins, we identified 14,384 duplicate proteins (proteins for which at least a single duplicate-pair is detected) and 3,961 singleton proteins (proteins where no close homologues are detected). The remaining 13,190 proteins were not classified into either category and were not further

analysed. Duplicate proteins were grouped into 2,738 protein families using a single-linkage algorithm, i.e., if protein A and protein B form a duplicate-pair and protein B and protein C form a duplicate pair, all three proteins are clustered into the same family.

In order to study whether pre-multicellularity and post-multicellularity duplication events lead to different expression breadth of their resulting duplicates, we studied the relationship between expression breadth and number of duplicate-pairs in three groups of mouse proteins (illustrated in Figure 11A): metazoan-specific proteins, post-Multicellularity Duplicates (postMD), and pre-Multicellularity Duplicates (preMD). Metazoan-specific proteins are 'novel' proteins that have emerged in a metazoan species (as a result of a domain shuffling (Soding and Lupas 2003; Waterston et al. 2002), for example). Since these proteins have emerged after the transition to multicellularity, all the duplication events of their encoding genes must have taken place in a multicellular organism. Pre-metazoan mouse proteins are 'ancient' proteins that have descended from a unicellular ancestor of mouse. Duplication of genes encoding pre-metazoan proteins, unlike duplications of genes encoding metazoan-specific proteins, can either predate or postdate the transition to multicellularity. The postMD protein group consists of pre-metazoan mouse proteins that have duplicates that have arisen through a post-multicellularity gene duplication event. The preMD protein group consists of pre-metazoan mouse proteins that do not have any such 'recent' duplicates, and therefore all of the duplicates of such proteins, if any, have arisen through a pre-multicellularity gene duplication event.

The classification of mouse proteins into the three groups (preMD, postMD and metazoan-specific) was done by comparing the mouse proteins to the proteins of an estimated last unicellular ancestor of mouse. To estimate the proteome of this last unicellular ancestor, we used a combination of the complete protein sequences from a variety of unicellular species (three unicellular eukaryotic species including Fungi and Alveolata, one bacterial species, detailed in section 3.3.8). The varied origin of these protein sequences reduces the impact of species-specific gene loss on the classification of preMD/postMD proteins. Proteins with no detectable homologues in any of these unicellular species, or in any other unicellular species out of more than 200 fully sequenced species examined (detailed in section 3.3.7), were classified as metazoan-specific proteins (***E***-value < 1e-3, similar to the definitions used in (Waterston et al. 2002)). Proteins with detectable homologues were further classified as preMD or postMD using the Inparanoid programme

(O'Brien et al. 2005) **(**detailed in section 3.3.8). The Inparanoid clustering procedure identifies orthologues between two species, and allows the identification of duplicates that arose through post-speciation duplication events and duplicates that arose through pre-speciation duplication events. Since the speciation between mouse and its last unicellular ancestor marks the appearance of a multicellular ancestor of animals, the construction of orthologous groups between the proteins of mouse and its last unicellular ancestor enables us to distinguish between pre- and post-multicellularity duplicate-pairs. The time-gap that exists between the speciation of mouse from its last unicellular ancestor and the speciation from the unicellular species analysed here, suggests that some of the proteins classified as postMD might have been the result of duplication events that took place in a unicellular species. However the estimated short length of this time-gap relative to the period of time since the appearance of a multicellular ancestor of animals (Hedges et al. 2004), together with the major role of duplication events and lineage-specific gene expansions known to be involved in shaping the metazoan gene repertoire supports the hypothesis that a substantial part of the duplicates of postMD proteins arose in a muticellular ancestor. Therefore this group is different from the preMD group where none of the duplicates of the proteins arose from a post-multicellularity duplication event. 15,394, 3,699, and 2,231 mouse proteins were classified as metazoan-specific, postMD and preMD proteins respectively (Table 7).

Expression information was retrieved from 22 adult mouse tissues (Affymetrix U74Av2 GeneChip). As the sequence similarity between duplicate proteins can lead to promiscuity of their reporting probes we limited the analysis to probes which uniquely report a single sequence (detailed in section 3.3.3). In cases where a probe set is mapped to several splice variants, only the longest transcript is further analysed. After this filtration the dataset contained expression information for 4914 mouse proteins. For each protein, we recorded its expression breadth according to the Absent/Present call in each tissue. In order to avoid re-counting similar tissues, the tissues were grouped into 13 clusters and only a single representative member of each cluster was used for analyses comparing expression breadth. All analyses were repeated using an additional microarray expression dataset from mouse tissues (Novartis GNF1M GeneChip) that will be referred herein as the additional dataset (detailed in section 3.3.1-4). The number of proteins in the groups analysed is listed in Table 7 (for the main dataset) and Table 10 (for the additional dataset).

Table 7 The total number of proteins in the different groups of phyletic age/time of duplication analysed. The numbers in brackets are the numbers of proteins in the group for which expression data were available.

	Complete dataset	PreMD subset	PostMD subset	Metazoan-specific subset
Number of proteins in the mouse proteome	31535 (4914)	2231 (740)	3699 (618)	15394 (1915)
Number of proteins that are either singleton or duplicate proteins*	18345 (2731)	811 (291)	2495 (431)	9390 (1060)
Number of singleton proteins	3961 (603)	667 (226)	0 (0)	1960 (792)
Number of duplicate proteins	14384 (2128)	144 (65)	2495 (431)	7430 (268)

* Proteins that did not match the criteria to be either singletons or duplicates were discarded

3.3 Methods

3.3.1 Expression data

We used microarray data from hybridizations of RNA from mouse tissues to Affymetrix U74Av2 GeneChip using the standard protocol. The data are available from ArrayExpress (www.ebi.ac.uk/arrayexpress/ (Rudd et al. 2001), accession ID = E-HGMP-2). Absence/Presence flags were generated using the Microarray Suite 5.0 package (Affymetrix MAS 5.0) with its default settings, as described in (Freilich et al. 2005). The detection algorithm implemented in MAS5 uses probe pairs intensities to generate a detection call for the transcripts. Each probe pair in a probe set is a factor in determining whether the measured transcript is detected (Present), marginal or not detected (Absent). The detection calls are calculated as detailed in the MAS5 manual (http://www.affymetrix.com/Auth/support/downloads/manuals/data_analysis_fundamentals_manual.pdf (Rohmer et al. 1979)). We used the default parameters (detection P value <

0.04) and have treated Marginal calls as undetected transcripts. In the text, this dataset is referred as the main dataset.

As an additional data source we used NOVARTIS microarray data from hybridizations of RNA from mouse tissues to Novartis GNF1M GeneChip (Su et al. 2004). Absence/Presence flags were generated using the Bioconductor implementation of the MAS5 algorithm with its default settings. The data are available from http://expression.gnf.org (Rahmann 1995). In the text, this dataset referred as the additional dataset. The two datasets are compared in section 3.3.4.

3.3.2 Construction of tissue-clusters

Main dataset: The Absence/Presence calls from 22 adult male mouse tissues were used to build 13 tissue-clusters by constructing a tree and then cutting it into clusters (binary distance measure, average agglomeration method). The tree was cut at a height that allowed as large as possible variety of tissue-clusters, and yet clustered together highly similar tissues (such as the two testis samples or different parts of the colon). The tissues and the tissue-clusters are listed in Table 8.

Additional dataset: Similarly, the Absence/Presence calls from 47 adult male mouse tissues were used to build 20 tissue-clusters, listed in Table 9.

In order to avoid re-counting of similar tissues we used a single representative of each tissue-cluster for analyses comparing the expression breadth of proteins in mouse tissues. The analyses were repeated using different tissue compositions and compatible results are obtained.

3.3.3 Identifying non-promiscuous probe sets and mapping probe sets to mouse proteins

For both chips (main and additional), the individual probes' sequences were aligned against all mouse transcripts predicted in the EnsEmbl (Hubbard et al. 2002) release 30.33f. The alignment procedure allows a single discrepancy with either the PM (perfect match) or MM (mismatch) sequence. Probes were filtered out if (1) they are not perfectly aligned with any transcript; (2) they are aligned with more than a single transcript (promiscuous probes). Only probe sets with all probes perfectly match to a single gene and no match any other

gene were mapped to proteins. Proteins represented by more than a single probe set were discarded in order to avoid redundancy. In cases where a probe set is mapped to several splice variants, only the longest transcript is further analyzed. A single and unique probe set therefore represents each of the proteins in our dataset. The main dataset contains the expression information for 4914 proteins, and the additional dataset contains the expression information for 13045 proteins.

Table 8 Tissue list of the main data set. Cluster representatives are indicated using bold face.

Tissue	Cluster ID
Antrum	1
Appendix	2
Bladder	3
Gall bladder	3
Lung	3
Brain	4
Eye	4
Cecum	5
Duodenum	5
Ileum	5
Jejunum	5
Distal colon	6
Proximal colon	6
Heart	7
kidney	8
Liver	9
Muscle	10
Spleen	11
Thymus	11
Testis	12
Testis	12
Vas deferens	13

Table 9 Tissue list of the additional data set. Cluster representatives are indicated using bold face.

Tissue	Cluster ID
Amygdala	1
Cerebellum	1
Cerebral cortex	1
Dorsalrootganglion	1
Dorsalstriatum	1
Frontal cortex	1
Hippocampus	1
Hypothalamus	1
Olfactory bulb	1
Preoptic	1
Retina	1
Lower and upper spinalcord	1
Substantianigra	1
Trigeminal	1
b220.bcell	2
cd4.Tcell	2
cd8.Tcell	2
Lymph node	2
Thymus	2
Heart	3
Testis	4
Adipose tissue	5
Adrenal gland	5
Bladder	5
Lung	5
Prostate	5
Kidney	6
Liver	7
Salivary gland	8
Digits	9
Epidermis	9
Snout epidermis	9
Tongue epidermis	9
Bone	10
Large intestine	11
Small intestine	11
Spleen	12
Pituitary	13
Medial olfactory epithelium (MOE)	14
Vomeral-nasal organ (VMO)	14
Thyroid	15
Stomach	16
Pancreas	17
Brown fat	18
Skeletal muscle	19
Bone marrow	20

3.3.4 Comparison between the main and the additional data set.

Of the 4914 proteins mapped to a probe-set from the U74Av2 GeneChip, and 13045 proteins mapped to a probe-set from the GNF1M Genechip, there are 3935 proteins in common. We studied the expression pattern (Absence/Presence calls) of these proteins in 14 tissues reported by the two data sets. The number of protein expressed in each tissue for both sources of data and the overlap set are shown in Figure 12. Most proteins that are reported as expressed by the GNF1M GeneChip are also reported as expressed by the U74AV2 GeneChip, but not vice versa. These differences between the GeneChips may be explained by the differences in their sensitivity due to differences in their probe-sets. In U74Av2 GeneChip most probe-sets are composed of 16 individual reporters, while in GNF1M GeneChip most probe-sets are composed of only 11 individual reporters. Those design differences affect the sensitivity of the MAS5 Presence/Absence call.

Based on the analysis performed and the design of the reporter sets, we believe that the U74Av2 arrays are more sensitive when only reporting absence or presence of expression. Therefore, we chose to report the results from this data source (main dataset). We have repeated all analyses with the GNF1M arrays as shown in section 3.9.

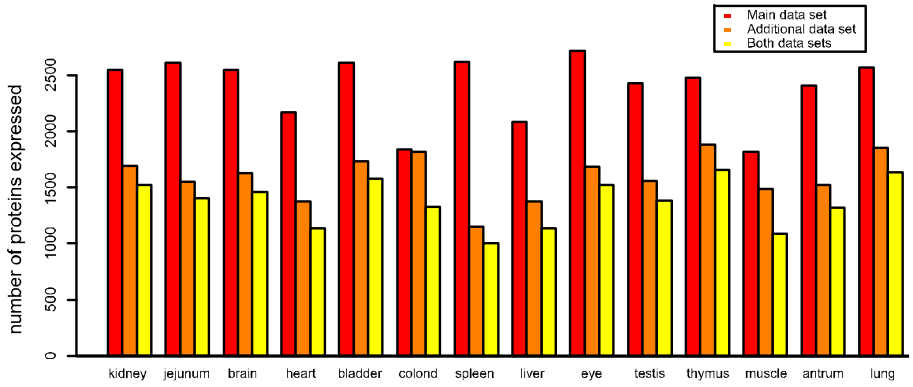

Figure 12 Number of proteins expressed in a tissue, out of 3935 proteins common to the main and additional data set.

3.3.5 Identification of singleton and duplicate proteins

We conducted an all-against-all BLAST (Altschul et al. 1997) self-search for the entire proteome of mouse (EnsEmbl release 30.33f). A singleton protein was defined as a protein that (1) did not hit any protein other than itself or its splice-variants with *E*-value <= 0.1; (2) recognized itself with *E*-value <= 1e-20. Proteins that recognize themselves with a low *E*-value (possibly as a result of a short sequence or low complexity which is masked by the BLAST search) will recognize their homologues with a low *E*-value. In order not to classify these proteins as singletons we applied this self-recognition condition.

Two proteins were regarded as duplicates if they meet the following criteria: (1) *E*-value <= 1-e10; (2) A mutual coverage of 80% between query and hit; (3) the proteins are not alternative forms encoded by the same gene. For each protein we counted the number of its duplicate-pairs, i.e., the number of proteins matching these criteria. When counting the number of pairs, all the homologues of the protein that are encoded by the same gene (i.e., splice-variants) are counted only once, as they arose through a common gene duplication event For example, if protein A has 3 homologues - C, D and E - and C and D are splice-variants of the dame genes, A is considered to have 2 duplicate-pairs.

Only proteins classified as either singletons or duplicates using the strict criteria above were further analyzed.

3.3.6 Retrieving evolutionary rate between mouse proteins and their orthologues in rat

Mouse and rat orthologous pairs and their calculated evolutionary rates (dN/dS values) were retrieved from EnsEmbl (downloaded 24.04.05). Those cases where a mouse protein matched more than a single rat orthologue were discarded from the analysis, unless one of the orthologous pairs was annotated as BRH (Best Reciprocal Hit). Evolutionary rate (dN/dS) values were obtained for 2279 proteins including 485 singleton proteins (out of 603) and 1794 duplicate proteins (out of 2128) that are represented in the main expression dataset. The mean rate and standard deviation for all pairs are 0.13 and 0.15, respectively.

3.3.7 Assignment of proteins into categories describing their time of origin in the mouse genome

Mouse proteins were classified as pre-metazoan (descendants from a unicellular ancestor of mouse) or metazoan-specific according to the results of a BLAST search (Altschul et al. 1997) against 221 fully sequenced species (including 22 eukaryote species, 6 of them metazoan). Genomes were downloaded from the COGENT (Janssen et al. 2003) database (version 228). The cut-off used to infer homology was BLAST E-value < 1e-3.

Proteins were only assigned to a single category. The classification process is hierarchical: first proteins with hits to more than ten prokaryote species and/or at least a single hit to a non-metazoan eukaryote are classified as pre-metazoan (14957 proteins). Mouse proteins recognising only metazoan proteins are classified as metazoan-specific (15394 proteins). 266 mouse proteins for which homologues were not inferred in any of the species (including mouse), probably due to short sequence or low-complexity, were not included in any of the categories.

3.3.8 Identification of preMD and postMD proteins

We used the Inparanoid programme (O'Brien et al. 2005) in order to classify pre-metazoan mouse proteins according to the estimated time (before or after the transition to multicellularity) when their extant duplicates in the mouse proteome have arisen. Briefly, the Inparanoid programme takes as input protein sequence information from two species (A and B) and clusters them into orthologous groups. Each group contains two main orthologues (protein A' from species A and protein B' from species B) which are reciprocal best hits. Proteins from species A that are more similar to A' than to any other protein from species B, and are more similar to A' than A' is similar to B are clustered together with A' in the same orthologous group. These proteins are considered to be in-paralogues of protein A', i.e., proteins that arose through a duplication of the gene encoding protein A' that took place after the speciation of species A from species B. Out-paralogues of protein A' are these proteins that arose through a duplication of the gene encoding protein A' that took place before the speciation of species A from species B. The request that in-paralogues of A' are

more similar to A' than A' is similar to B' reduces the probability that as a result of species-specific gene loss out-paralogues are classified as in-paralogues.

Firstly, we wanted to identify a group of mouse proteins that do not have any duplicates from duplication events that postdate the transition to multicellularity (preMD proteins). Such proteins will therefore not have any in-paralogues when species A is mouse and species B is its last unicellular ancestor. Secondly, we wanted to identify a group of mouse proteins for which all their duplicates arose through duplication events that postdate the transition to multicellularity (postMD proteins). Such proteins will therefore have only in-paralogues (but not out-paralogues) when species A is mouse and species B is its last unicellular ancestor. As a reference to the genome of the last unicellular ancestor of mouse we used the combined sequences of the complete proteomes of *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Escherichia coli* and *Plasmodium falciparum*.

The combined unicellular protein sequences (downloaded from the Inparanoid database, 20.06.05), together with the mouse sequences, were used as input for the Inparanoid programme. Inparanoid was run with default parameters. The clustering procedure had identified unicellular orthologues for 8305 mouse proteins. Of these, 5886 proteins have mouse in-paralogues and 2419 proteins do not.

The Inparanoid programme provides predictions for the in-paralogues a protein has, but not for its out-paralogues. Here, we used the Inparanoid clustering only in order to classify a protein as preMD or postMD, but not in order to count its number of duplicate-pairs since duplicates which are out-paralogues will not be recognised. Because the classification of protein as preMD or postMD was done separately from the count of its duplicate-pairs, we filtered the groups to include only proteins where all their duplicate-pairs are classified into an Inparanoid orthologous group (not necessarily the same group as some of the duplicate-pairs can be out-paralogues). This filtration omits, for example, proteins that have out-paralogues that as a result of a species-specific gene loss were not classified to any orthologous cluster. After the filtration the dataset contained 3920 proteins for which the Inparanoid procedure had identified in-paralogues (out of 5886) and 2231 proteins for which no such in-paralogues were identified (out of 2419). For those 2231 proteins without in-paralogues, we can therefore only identify duplicate-pairs that have a different unicellular orthologue, i.e.; those duplicate-pairs are predicted to arise through a pre-speciation duplication event. These 2231 proteins are termed here preMD proteins.

From the 3920 proteins with in-paralogues, we filtered out those proteins which not all of their duplicate-pairs are classified into the same Inparanoid orthologous group, leaving 3699 proteins in which all their duplicate-pairs are also in-paralogues, i.e., arose through a post speciation duplication event. These 3699 proteins are termed here postMD proteins.

3.4 The relationship between the number of homologues and the expression breadth of a protein

3.4.1 The expression breadth of a protein is negatively correlated with the number of its duplicate-pairs

A tendency of mammalian genes from big families to be specifically expressed was previously reported (Huminiecki and Wolfe 2004). We tested whether such tendency exists in our dataset by studying the dependence between expression breadth of a protein and the number of its duplicate-pairs. Expression information from the 13 representative tissues was available for 603 singleton proteins and 2128 duplicate proteins (Table 7). A significant negative correlation is observed between the number of duplicate-pairs and the expression breadth of a protein (kendall's tau = –0.20, *P* value < 2.2e-16). The expression is plotted against the number of duplicate-pairs in Figure 13A, which demonstrates the large variation between individual proteins. In order to illustrate the correlation observed for the raw data, we collected the proteins into bins according to the number of their duplicate-pairs (small bins were merged with neighbours so that each bin included at least 100 members). When plotting the mean expression in each bin against its ranking order (Figure 13B) one can observe a tendency for proteins with many duplicate-pairs to be more specifically expressed. The results are repeatable when using the additional dataset (Figure 16).

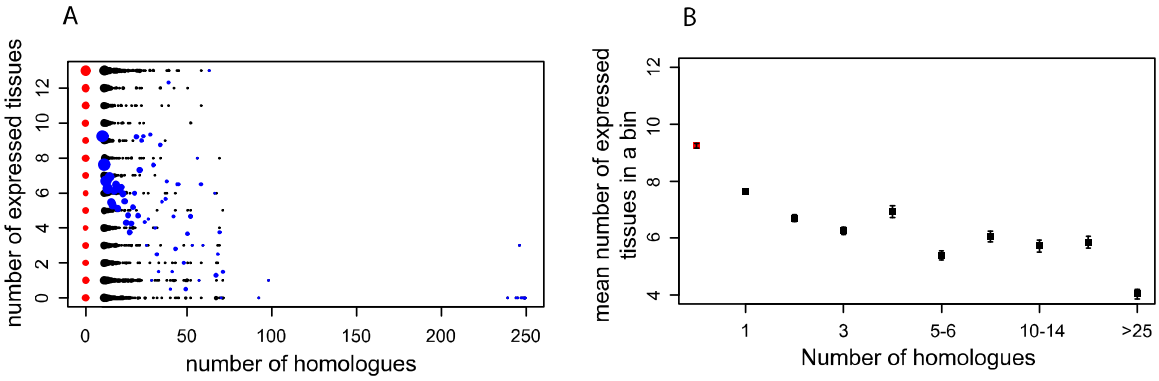

Figure 13 Expression breadth versus the number of duplicate-pairs. Red dots – singleton proteins, black dots – duplicate proteins. The tissues tested are the 13 cluster-representing tissues. (A) The size of the dots represents the number of proteins that have the same number of duplicate-pairs and the same expression breadth. The blue dots represent the average expression breadth of proteins with the same number of duplicate-pairs. Sample size: 2731 proteins; kendall's tau = -0.20; *P* value = 7.5e-56; 95% confidence intervals: (-0.22, -0.17) (B) Proteins are ordered according to their number of duplicate-pairs and collected into bins of at least 100 proteins. Each point represents a bin. Error bars indicate the standard deviation from the mean, obtained by bootstrapping.

3.4.2 The expression breadth is negatively correlated with the number of duplicate-pairs, independently of the correlation between expression breadth and rate

A possible explanation for the dependence between the number of duplicate-pairs and the expression breadth is the mutual correlation of both factors with the rate of evolution. Several studies show a correlation between the expression breadth of mammalian proteins and their recent rate of evolution, as inferred by comparison to an orthologue from another mammalian species (Duret and Mouchiroud 2000; Winter et al. 2004; Zhang and Li 2004). Other studies report a dependence between recent duplication events and the rate of evolution of the duplicated proteins (Kondrashov et al. 2002). For comparison with these studies we studied the dependencies between rate and expression breadth and between rate and number of duplicate in a subset of 2279 proteins from our data set that have a rat orthologue (detailed in section 3.3.6). Compatible with these studies, we found a correlation between rate and expression breadth (kendall's tau of –0.11, P value = 1.4e-14) and between rate and the number of duplicate-pairs (kendall's tau of –0.06, *P* value = 1.5e-5). Both correlations are weaker than the correlation reported here between expression breadth and number of duplicates.

We wanted to study whether the relationship between expression breadth and number of duplicates can be explained purely in terms of both factors' mutual correlation with the recent rate of evolution. Such test was suggested and performed by Tim Massingham. The first step was to estimate the dependence between expression breadth and number of duplicate-pairs using a standard contingency table. Next, the contingency table statistic was compared to those formed by randomizing the data such that only the correlation between the expression and number of duplicate-pairs that is due to rate is saved. The randomization was performed by grouping the proteins into bins of a similar recent rate of evolution and shuffling the numbers of duplicate-pairs within each group, forming a new set of data which have identical correlation between expression and rate as the original data and retaining a similar correlation between number of duplicate-pairs and rate. The contingency table statistic for the original data was compared to those of 10,000 sets of randomized data and it exceeded them all (test statistic 284, compared to maximum of 174 from 10,000 sets of data randomised to be consistent with the null hypothesis). Therefore, there is a relationship between the number of duplicates and the expression breadth that cannot be explained by the mutual correlation with the recent rate of evolution

3.5 Only post-multicellularity duplication events lead to expression specificity

The previous analysis indicates that, on average, a duplication event leads to a narrower expression profile (Figure 13). Can the narrowing of expression be a factor in promoting the retention of duplicate genes in the genome? If this was true, only duplication events that took place in a tissue-differentiated environment would be expected to lead to a decrease in the expression breadth of the duplicated genes. In contrast, duplication events that occurred in the unicellular ancestor would not, as the retention of the duplicate genes in the genome could not be due to tissue specification.

In order to investigate the effect of pre-multicellularity duplication events on expression breadth, we studied the correlation between expression breadth and the number of duplicate-pairs in the preMD subset - a subset of mouse proteins whose duplicates have all arisen through pre-multicellularity duplication events. In agreement with the hypothesis, preMD proteins illustrated no tendency towards an increase in tissue specificity accompanying a rise in the number of duplicate-pairs (kendall's tau = 0.001, *P* value = 0.51,

Figure 14A). When grouping the preMD proteins into bins the mean expression breadth in all bins remains approximately constant and high (about 10 tissues – the same as for singleton proteins), indicating a global expression for the majority of preMD proteins (Figure 14B).

For comparison to the preMD subset, we studied the correlation between expression breadth and number of duplicate-pairs in two subsets of mouse proteins whose duplicates (at least in part) have arisen through a post-multicellularity duplication event – the postMD and the metazoan-specific subsets. In agreement with the predicted tendency of post-multicellularity duplication events to lead to tissue specific expression, we detect a significant negative correlation between expression breadth and number of duplicate-pairs in both subsets (postMD subset: kendall's tau = -0.15, *P* value = 1.2e-6; metazoan-specific subset: kendall's tau = -0.28, *P* value < 2.2e-16). Although the correlations are not high, and although there is a great variability in the distribution of the data (Figure 14C,E), the significance of the correlations indicates that in both dataset there is a tendency of proteins with many homologues to be specifically expressed. This tendency is emphasised when binning the data according to the number of duplicate pairs and plotting the bins against their average expression breadth (Figure 14D,F). The proteins in these two subsets differ in their estimated phyletic age: unlike the 'novel' metazoan-specific mouse proteins, postMD proteins are estimated to be 'ancient' pre-metazoan proteins. The detection of negative correlation in one of the pre-metazoan protein subsets (the postMD subset) together with the inability to detect such correlation in the second pre-metazoan protein subset (the preMD subset) emphasises the importance of the time of duplication (D in Figure 11A), rather than phyletic age (A in Figure 11A), in shaping the relationship between duplication events and expression breadth. Such a relationship is only evident in the two subgroups where duplication events have postdated the transition to multicellularity.

The postMD and the metazoan-specific subsets differ in the mean expression breadth of bins that have a similar number of duplicate-pairs (Figure 14D,F). The average higher expression breadth of the postMD proteins possibly reflects the effect of the phyletic age of a protein on its expression breadth, where metazoan-specific proteins tend to be more tissue-specific than pre-metazoan proteins (Freilich et al. 2005).

The findings of this analysis indicate that only duplication events that postdate the transition to multicellularity tend to lead to the development of a tissue specific expression

pattern. This supports the prediction of the subfunctionalization model that tissue specialisation is a factor in the retention of duplicate genes in the genome of multicellular organisms. The same analysis was repeated using the additional dataset and the results obtained are compatible with those reported here: only duplicates that have arisen through post-multicellularity duplication events show a tendency to be more specifically expressed (Figure 17).

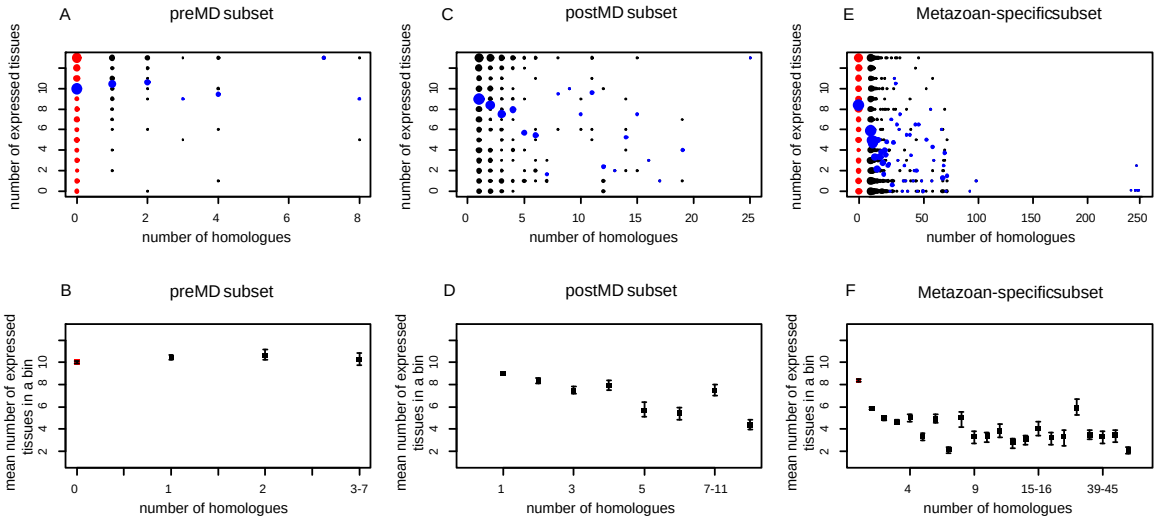

Figure 14 Expression breadth versus the number of duplicate-pairs. Red dots – singleton proteins, black dots – duplicate proteins. The tissues tested are the 13 cluster-representing tissues. (A,C,E) The size of the dots represents the number of proteins that have the same number of duplicate-pairs and the same expression breadth. The blue dots represent the average expression breadth of proteins with the same number of duplicate-pairs. preMD subset (A): 291 proteins, kendall's tau = 0.001, *P* value = 0.51, 95% confidence intervals: (-0.10, 0.10); postMD subset (C): 431 proteins, kendall's tau = -0.15, *P* value = 1.2e-6, 95% confidence intervals: (-0.23, -0.08); Metazoan-specific proteins subset (E): 1060 proteins, kendall's tau = -0.28, *P* value = 9.7e-43, 95% confidence intervals:
(-0.33, -0.24); (B,D,F) Proteins are ordered according to their number of duplicate-pairs and collected into bins of at least 10 proteins. Each point represents a bin. Error bars indicate the standard deviation from the mean, obtained by bootstrapping.

3.6 Cumulative tissue distribution of protein families is not correlated with family size

We wanted to study the expression breadth of protein families to test whether we can find a correlation not only between expression breadth of an individual protein and its number of duplicate-pairs, but also between the size of the family and the cumulative tissue distribution of the entire family (as illustrated in Figure 11B). Two possible scenarios for the

relationship between the size and the expression breadth of protein families can explain the tendency of proteins with many homologues to be specifically expressed:

- Complementary expression: a gene duplication event leads to tissue specialization of either one or both daughter genes, yet, the two duplicates together cover the expression range of the ancestral gene.
- Identical expression: retention of a duplicate gene in the genome is more likely when its expression is tissue-specific. Both duplicate genes will have the same specific expression pattern as the ancestor gene. The olfactory receptor family, one of the largest mammalian protein families, is one example where many members are specifically expressed in one type of cell, the olfactory epithelium (Hellman and Chess 2002; Young and Trask 2002).

We studied the relationship between size and expression breadth in 1249 protein families where expression information was available for at least a single family-member. The families vary both in size and in the fraction of the members for which expression information is available. For each protein family a cumulative expression profile was created by summing all tissues in which at least a single-family member is expressed (Figure 11B). No negative correlation is observed between the cumulative expression coverage of a protein family and its size (kendall's tau = -0.03, *P* value for a less one-sided test = 0.02). This is unlike the significant negative correlation observed between the expression breadth of an individual protein and its number of duplicate-pairs (Figure 13). Since for many of the families (43%) expression information is available for only a single member, we repeated the analysis using a subset of 189 protein families where expression information is available for at least ¾ of family members. Again, no negative correlation is observed when using this high-coverage subset, and the positive values of both confidence intervals exclude the possibility of negative correlation (kendall's tau = 0.07, *P* value = 1, 95% confidence interval: [0.01,0.14]). Therefore, although we have only partial expression information for the large majority of families, our data imply that increasing the size of a family does not affect, on average, the cumulative tissue distribution of a family. In Figure 15 we binned protein families according to their size and calculated the average cumulative expression distribution for each bin. As shown in the figure, the average cumulative distribution of protein families does not decrease when families increase in size and even when using the complete, low-

coverage, dataset (black dots) the average cumulative expression in all bins is approximately identical to the average expression breadth of singleton proteins. A better coverage of family members in the expression data is most likely to strengthen this observation, as indicated by the use of a high-coverage subset (green dots). The same observation applies when using the additional dataset (Figure 18).

Taken together the expression breadth of proteins (Figure 13) with the expression breadth of protein families (Figure 15), our results support the complementary expression model where a duplication event leads to a tissue specialization of one or both copies while the original tissue-distribution of the protein family remains constant.

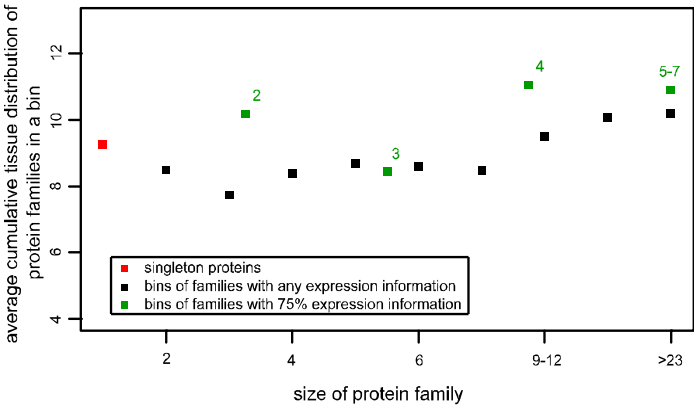

Figure 15 Average cumulative expression coverage in bins of protein families, ordered by size of family. Proteins families where expression information is available for at least a single member (black) are grouped into bins of at least 35 proteins (total number of families – 1249). Protein families where expression information is available or at least 75% of the family members (green) are grouped into bins of at least 10 proteins (total number of families – 189). Each point represents a bin. Values in the X-axis describe the size of a family with any expression information (black dots). The size of families with 75% expression information (green dots) is the value on top of each green dot.

3.7 Caveats

Several limitations of this analysis must be acknowledged:

- The analysis performed is based only on approximately one-fifth of mouse proteins (proteins that are included in the main expression data) and the expression values are based on a single replicate. However, we were able to repeat the analyses with an additional dataset and obtain compatible results.
- In both databases the expression data are retrieved from tissues, rather than from cell lines. A high-coverage, multi-replicates expression data from a wide

collection of mammalian cell lines would be ideal for performing such an analysis, but is not yet available.

- The sequence similarity search used in order to estimate the phyletic age of a protein might not be sensitive enough to detect distant homologues of fast evolving proteins due to their lower similarity with distant homologues. Fast-evolving pre-metazoan proteins can therefore be misclassified as metazoan-specific proteins (together with genuine metazoan-specific proteins such as those formed by domain shuffling (Soding and Lupas 2003; Waterston et al. 2002), for example). Although such misclassification of fast evolving proteins is possible, for several reasons it is not likely to affect the results of the analysis performed here. Mainly, dependence between the number of duplicate-pairs and expression breadth was not only detected for metazoan-specific proteins, but was also detected for a group of conserved proteins, which have recognised homologues in distant species (the postMD subgroup). The detection of dependence in the postMD group diminishes the likelihood that the reported dependence in the metazoan-specific group is derived from the misclassification of fast evolving proteins. To further preclude the possibility that biases in the distribution of evolutionary rate effect the analysis we tested the dependence between expression breadth and the number of duplicate-pairs for both the fastest and slowest evolving proteins within the metazoan-specific subset. A significant negative correlation between expression breadth and number of duplicate-pairs were found both in a group of 248 metazoan-specific proteins where mouse-rat evolutionary rate (d_N/d_S) < 0.05 (group 1) and in a group of 49 proteins where rate < 0.005 (group 2). Correlation was also detected in the group of the 248 fastest evolving proteins (group 3, evolutionary rate > 0.23). Correlation for group 1: kendall's tau = -0.28, *P* value = 5.11e-11; Correlation for group 2: kendall's tau = -0.39, *P* value 3.27e-5; Correlation for group 3: kendall's tau = -0.29, *P* value 6.7e-12. Finally, dependence between expression breadth and number of duplicates, demonstrated for the complete dataset, was shown here (section 3.4.2) to be independent of the recent rate of evolution.

3.8 Conclusions and discussion

By analysing the relationship between the expression breadth of a protein and its number of duplicate-pairs, we have shown that proteins have a tendency to become more specifically expressed after their encoding genes are duplicated. Such a tendency is not observed for the subset of proteins whose duplicates arose through events that predate the transition to multicellularity. Therefore, our analysis supports the view that expression divergence, following gene duplication, acts as a stabilising factor to retain a duplicate gene in the genome of multicellular species. The fact that we do not observe tissue specification of duplicates from pre-multicellularity duplication events suggests that these proteins had undergone a different type of subfunctionalization such as specialisation of their temporal expression or biochemical functions. Unlike the tendency towards specific expression of their protein members, protein families tend to maintain a global expression pattern, therefore implying that the specification of expression between family members is complementary. The findings of this large scale analysis are consistent with the predictions of the subfunctionalization model, which states that the division of the expression (among other functions) of an ancestor gene between its daughter duplicates promotes the retention of a gene in the genome (Lynch and Force 2000). However, given the lack of information about the expression pattern of a pre-duplication ancestor gene, the analysis performed here can only provide evidence for the current complementary expression between family members, but does not illuminate the expression pattern of the ancestral state. Other studies have indicated that the expression of duplicate genes is labile and often not consistent with the ancestral state (Gu et al. 2004; Huminiecki and Wolfe 2004).

Does the specification of expression following duplicatio event lead to the evolution of new, tissue-specific functions? Few reported examples link specification of expression between duplicate genes to the specification of their function, leading to th emergence of new tissue types. One such example is the duplication of an ancestral *opsin* gene into two paralogues, *c-opsin* and *r-opsin*. The paralogues are found, respectively, in the ciliary and rhabdomeric photoreceptor sister cell types, leading to differences in the light-sensitivity of those cells. It has been suggested that this duplication event, that took place in an early metazoan ancestor, had allowed the diversification of these two cell types from a precursor photoreceptor ancestor cell (Arendt et al. 2004). In mammals, ciliary photoreceptor cells

have became the main visual photoreceptor cells (rods and cones), whereas rhabdomeric photoreceptor cells are thought to give rise to cells involved in photoperiodicity regulation (Arendt et al. 2004).

How does the specification of expression lead to the evolution of new, tissue-specific functions? Several lines of evidence indicate that specifically expressed genes diverge at higher rate (Duret and Mouchiroud 2000; Winter et al. 2004; Zhang and Li 2004), possibly due to less strict functional constraints (Hastings 1996). Revealing functional divergence between homologous can shed light on the way the duplication of certain genes has co-evolved with the development of new tissues. As such understanding is only just beginning, robust approaches are needed in order to link the repertoire of genes expressed in a tissue to a unique physiological role. The next chapter concerns the characterization of a subset of tissue-specific singleton genes in order to reveal some of the processes which are specif to a tissue.

3.9 Supplementary information

This section contains the results from the analysis of the additional data set. Table 10 is equvivalent to.Table 7, Figure 16, Figure 17, and Figure 18 are equvivalent to Figure 13, Figure 14, and Figure 15 respectevely.

Table 10 The total number of proteins in the different groups of phyletic age/time of duplication analysed. The numbers in brackets are the numbers of proteins in the group for which expression data were available in the additional dataset.

	Complete dataset	PreMD subset	PostMD subset	Metazoan-specific subset
Number of proteins in the mouse proteome	31535 (13045)	2231 (1317)	3699 (1335)	15394 (5908)
Number of proteins that are either singleton or duplicate proteins	18075 (7127)	811 (489)	2495 (875)	9390 (3281)
Number of singleton proteins	3691 (2054)	667 (396)	0 (0)	1960 (1041)
Number of duplicate proteins	14384 (5073)	144 (93)	2495 (875)	7430 (2240)

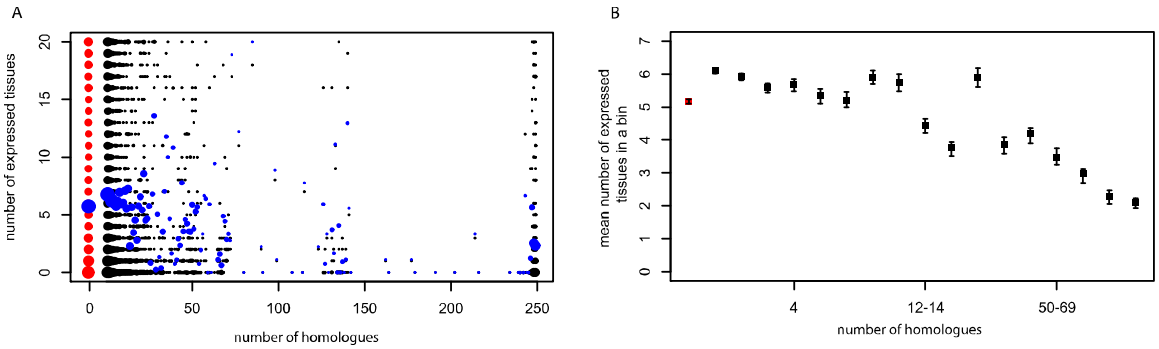

Figure 16 Analysis based on the additional database. Expression breadth versus the number of duplicate-pairs. Red dots – singleton proteins, black dots – duplicate proteins. The tissues tested are the 20 cluster-representing tissues. (A) The size of the dots represents the number of proteins that have the same number of duplicate-pairs and the same expression breadth. The blue dots represent the average expression breadth of proteins with the same number of duplicate-pairs. Sample size: 7127 proteins; kendall's tau = -0.02; *P* value = 8.7e-4; 95% confidence intervals: (-0.04, -0.01). (B) Proteins are ordered according to their number of duplicate-pairs and collected into bins of at least 100 proteins. Each point represents a bin. Error bars indicate the standard deviation from the mean obtained by bootstrapping.

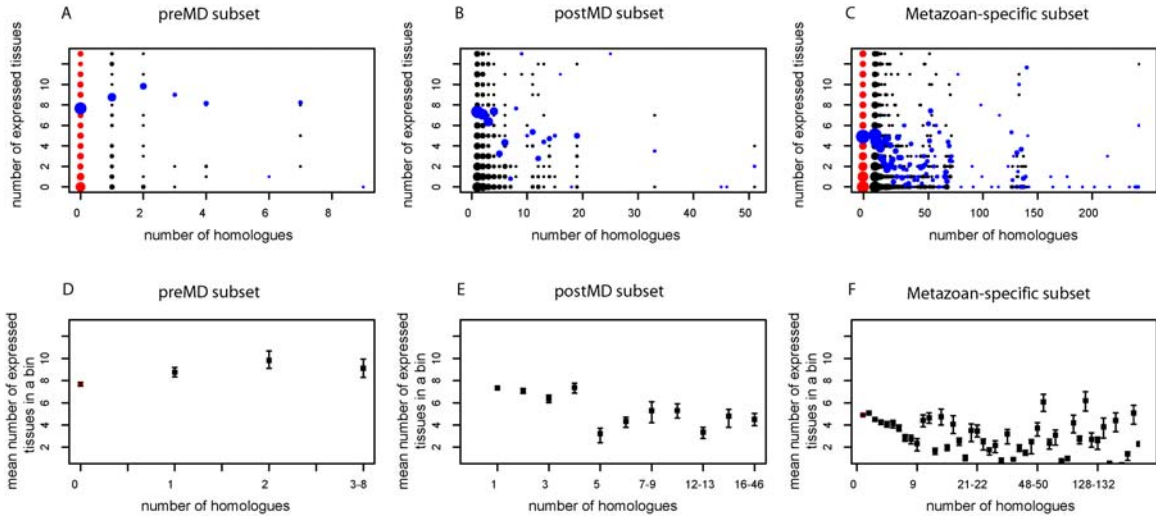

Figure 17 Analysis based on the additional database. Expression breadth versus the number of duplicate-pairs. Red dots – singleton proteins, black dots – duplicate proteins. The tissues tested are the 20 cluster-representing tissues. (A,C,E) The size of the dots represents the number of proteins that have the same number of duplicate-pairs and the same expression breadth. The blue dots represent the average expression breadth of proteins with the same number of duplicate-pairs. preMD subset (A): 489 proteins, kendall's tau = 0.07, *P* value = 0.99, 95% confidence intervals: (-0.01, 0.14); postMD subset (C): 875 proteins, kendall's tau = -0.07, *P* value = 1.5e-3, 95% confidence intervals: (-0.12, -0.02); Metazoan-specific proteins subset (E): 3281 proteins, kendall's tau = -0.11, *P* value = 3.7e-22, 95% confidence intervals: (-0.14, -0.09); (B,D,F) Proteins are ordered according to their number of duplicate-pairs and collected into bins of at least 10 proteins. Each point represents a bin. Error bars indicate the standard deviation from the mean obtained by bootstrapping.

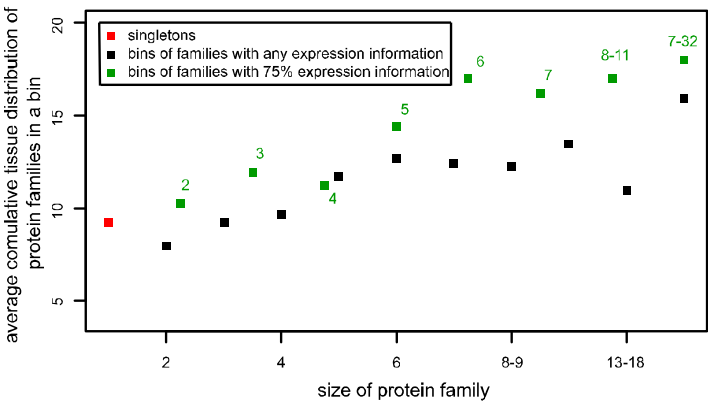

Figure 18 Analysis based on the additional database. Average cumulative expression coverage in bins of protein families, ordered by size of family. Proteins families where expression information is available for at least a single member (black) are grouped into bins of at least 35 proteins (total number of families – 2054). Protein families where expression information is available or at least 75% of the family members (green) are grouped into bins of at least 10 proteins (total number of families – 761). Each point represents a bin. Values in the X-axis describe the size of a family with any expression information (black dots). The size of families with 75% expression information (green dots) is the value on top of each green dot. The tissues tested are the 20 cluster-representing tissues.

Chapter 4: Characterisation of tissue–specific processes

4.1 Overview

The unique physiological role of each mammalian tissue is determined by the unique composition of the genes expressed in the tissue - the tissue's transcriptome. The transcriptome of each tissue comprises genes that are expressed globally and genes whose expression is limited to a subset of tissues. Identification of tissue-specific genes is one approach towards better understanding of the molecular basis behind tissue diversity. Yet, the identification of such proteins cannot always shed light on the nature of pathways unique to a tissue as in many cases tissue-specific proteins have homologues that perform the same function in a larger variety of tissues. Such compensation is less likely for singleton proteins. The singleton tissue-specific proteins are therefore an ideal group for identifying and characterising tissue-specific processes.

In order to characterise such processes we have studied the tissue distribution of specifically expressed singleton and non-singleton proteins. Tissue-specific proteins are defined as those expressed in no more than 3 tissue-clusters (see section 3.3.2). 497 such duplicate proteins and 60 singleton proteins were identified (as detailed in section 4.2). From the singleton proteins, 25 proteins are classified as pre-metazoan proteins and 32 proteins are classified as metazoan-specific proteins. The tissue distribution, SWISS-PROT accessions (Bairoch and Apweiler 2000) and GO annotations (Camon et al. 2003) of all 60 tissue-specific singletons are listed in a table in Appendix C. Firstly, the distribution of tissue-specific singleton and duplicate-proteins in various mammalian tissues was characterised. Secondly, for singleton proteins where annotations are available, a few examples for pre-metazoan and metazoan-specific proteins were looked in detail.

4.2 Data and methods

The data used for the analysis are the same data used in the previous chapter (main data set only).

4.2.1 Assignment of GO annotations to mouse proteins.

The assignment of GO annotations to mouse proteins was done by Eric Blanc. In order to obtain a coverage as extensive as possible, functional annotations have been gathered both from Ensembl release 30.33f (12360 genes assigned to a GO annotation) and directly from MGI (16084 genes assigned to a GO annotation). From MGI (http://www.informatics.jax.org/), we have also obtained a mapping between MGI and Ensembl identifiers, which has allowed us to merge their functional annotations with those obtained from Ensembl. Annotations were then normalized to GO controlled vocabulary (http://www.geneontology.org/) as of September 2004 containing 16805 terms. Those annotations featuring a GO id not present in that version of GO were discarded. In total, 86029 functional annotations describe 14901 genes, or 112901 functional annotations describe 19219 proteins.

4.3 Tissue distribution of tissue-specific proteins

The tissue distribution of specifically expressed singleton and duplicate proteins is listed in Table 11. The highest fraction of tissue-specific duplicate proteins (out of the total number of proteins expressed in the tissue) is observed in the brain, whilst the highest fraction of tissue-specific singleton proteins is observed in the testis. Though, the sample size is too small to enable a conclusive statistical analysis.

The identification of high fraction of tissue-specific duplicate proteins in the brain is compatible with previous studies showing that many of the mammalian proteins that have arisen through recent duplication events are integral for the structure and function of the brain (Fortna et al. 2004). One possible explanation for the identification of a relatively high number of testis-specific singleton proteins might be the rapid divergence rate of proteins mediating sexual reproduction, a phenomenon that has been suggested to play a role in the establishment of fertilization barriers and speciation (Swanson and Vacquier 2002). The size of pre-metazoan and metazoan-specific, tissue specific singleton protein sets are too small to be characterised statistically (Table 11) and therefore we have only looked at a few examples from each set.

Table 11 Tissue distribution of tissue-specific (TS) proteins (expressed in at most 3 tissue-clusters).

Tissue	% tissue specific proteins (4573*)	TS singleton proteins (60)†	TS duplicate proteins (497)†	Phyletic distribution of singleton proteins	
				Pre-metazoan (25)	Metazoan- specific (32)
Antrum	0.021(2958)	2(0.001)	33(0.011)	2	0
Appendix	0.007(2168)	1(0.000)	7(0.003)	0	1
Bladder	0.035(3236)	5(0.002)	54(0.017)	3	2
Brain	0.086(3183)	11(0.003)	147(0.046)	4	7
Cecum	0.033(3219)	7(0.002)	52(0.016)	5	2
Distal colon	0.004(2271)	1(0.000)	5(0.002)	0	1
Proximal colon	0.007(2382)	1(0.000)	10(0.004)	0	1
Duodenum	0.045(3230)	6(0.002)	83(0.026)	2	3
Eye	0.066(3390)	9(0.003)	121(0.036)	5	4
Gall bladder	0.040(3338)	12(0.004)	70(0.021)	10	2
Heart	0.014(2693)	1(0.000)	23(0.009)	0	1
Ileum	0.050(3393)	11(0.003)	82(0.024)	6	5
Jejunum	0.043(3230)	5(0.002)	72(0.022)	2	3
Kidney	0.029(3144)	5(0.002)	58(0.018)	4	1
Liver	0.038(2588)	7(0.003)	66(0.026)	5	2
Lung	0.039(3166)	4(0.001)	72(0.023)	0	3
Muscle	0.013(2241)	2(0.001)	23(0.010)	2	0
Spleen	0.065(3271)	11(0.003)	81(0.025)	6	4
Testis	0.074(3014)	21(0.007)	90(0.030)	6	13
Thymus	0.056(3099)	18(0.006)	58(0.019)	10	8
Vas deferens	0.012(2594)	1(0.000)	16(0.006)	0	1

*Numbers in brackets: total number of expressed proteins
†Numbers in brackets: fraction out of the total number of proteins expressed in the tissue

4.4 Examples of pre-metazoa and metazoan-specific, tissue-specific processes

Metazoan-specific proteins are, in many cases, involved in tissue-specific activities (Freilich et al. 2005; Lehner and Fraser 2004; Subramanian and Kumar 2004) and the recruitment of their encoding genes to the genome, therefore, accompanies the emergence of highly differentiated organs in multicellular species. Sperm protamine P3, Uteroglobin, Neuromedin U-23 and RAG2, all metazoan-specific, tissue-specific proteins (see table in Appendix C), are examples of proteins where function is unique to the tissue where they are expressed. Sperm protamine P3, which in the data set is specifically expressed in the testis, participates in the compaction of chromatin in the spermatid during spermiogenesis (Aoki and Carrell 2003). Uteroglobin, which in the data set is expressed in the lung, is an anti-inflammatory protein specific to the epithelium cells of pulmonary airways, whose decreased

expression is associated with hay fever (Benson et al. 2005). Neuromedin U-23 protein, which in the data is expressed in gastrointestinal track tissues, is thought to stimulate muscle contractions of specific regions in the gastrointestinal tract (http://us.expasy.org/uniprot/NEUU_MOUSE). The RAG2 protein (V(D)J recombination activating protein 2), which in the data set is specifically expressed in the thymus, is essential for the assembly of T-cell-receptor genes in developing lymphocytes (Agrawal et al. 1998).

Tissue-specific expression of pre-metazoan singleton proteins is especially interesting. If we assume such proteins are integral to a biological process (being singletons) in the unicellular ancestor, the process, therefore, has been located in a specialised tissue in multicellular species. It may be that some molecules are only required in one tissue and therefore the enzymes to make them need only to be expressed there. Alternatively, catabolism of some molecules may be restricted to a single tissue, which acts on behalf of the whole organism. Examples for such cases are the enzymes Histidase, Homogentisicase and Inositol-oxygenase (see table in Appendix C). Histidase catalyses the first step of histidine degradation, a process that takes place in the liver and skin of mammals (Taylor et al. 1991). Homogentisicase participates in the catabolism of tyrosine and phenylalanine and its expression in mammals is restricted to liver, kidney, small intestine and prostate (Granadino et al. 1997). Inositol-oxygenase catalyses the first committed step in the only pathway of *myo*-inositol catabolism (http://ca.expasy.org/uniprot/MIOX_MOUSE), which in mammals occurs predominantly in the kidney.

4.5 Summary

Characterisation of the molecular variations behind tissue diversity sheds light on way the evolution of the genome of multicellular species is related to the appearance of cell types. A few examples of pre-metazoan and metazoan-specific proteins that participate in tissue-specific processes are described. These examples demonstrates that 'metabolic organs' such as kidney and liver, perform in mammals functions that took place in the unspecialised unicellular ancestor, while possibly releasing other mammalian cell types from the constraints involved with performing such functions.

The small number of tissue-specific singleton proteins in the data set does not enable the characterisation of tissues as 'conserved' or 'derived, compared to a unicellular ancestor.

The next chapter describes a more robust approach taken in order to try and classify tissues according to the 'novelty' of their transcriptome.

Chapter 5: Relationship between function and a phyletic origin of a protein and its expression pattern

5.1 Introduction and overview

Comparative genomic studies have shown that the evolution of the metazoa lineage involves the expansion of those specific protein families known to participate in cellular communication and transcriptional regulation (Aravind and Subramanian 1999; Chervitz et al. 1998). However at the cellular level, it is not yet clear how processes taking place in specific tissues relate to similar processes, which took place in the ancestral unicellular species. The recent availability of fully sequenced genomes together with expression data from mammalian tissues enables us not only to identify those proteins which are unique to multicellular species but also to examine their contribution to tissue diversity. We can now study the protein content of a mammalian tissues in comparison to the protein content of unicellular organisms. To what extent does the differentiation process involve gaining new functions and to what extent does it involve specialisation of pathways that existed in a unicellular ancestor? Will 'young' proteins (i.e., proteins that are unique to multicellular species) exhibit a different expression pattern than 'ancient' or universal proteins?

Recent studies have related several characteristics of the phyletic age of a protein to its expression profile. Subramanian and Kumar (Subramanian and Kumar 2004) have shown a connection between a protein's phyletic age and the intensity of expression as measured by number of expressed sequence tags. Lehner and Fraser (Lehner and Fraser 2004) show that protein domains differ in their tendency to be specifically or widely expressed and that many of the tissue-specific domains are metazoan-specific.

This chapter describes a study of the relationship between the phyletic age and the function of a protein and its expression profile. Mouse proteins were assigned into functional and phyletic groups and the gene expression patterns of the different protein groupings were examined. As the composition of the phyletic groups highly correlates with that of the functional groups, the data were tested in order to determine which of the two factors -- function or phyletic age -- is dominant in shaping the expression profile of a protein. The observed differences in expression patterns of genes between functional groups

were found to mainly reflect their different phyletic origin. Finally, although metazoan-specific proteins tend to be tissue-specific compared to phyletically conserved proteins present in all domains of life, many such 'universal' proteins are also tissue- specific.

This work has been published in (Freilich et al. 2005).

5.2 General description of the data

Mouse proteins were assigned into functional and phyletic groups in a process described in Figure 19. Firstly, mouse proteins were assigned into a functional category. For simplicity and in order to have a sufficient sample size, the analysis was limited to proteins assigned to one of four main categories from the highest hierarchy level of the GO functional classification: two regulatory categories (signal transduction and transcription regulation) and two metabolic categories (enzymes and transporters).

Next, the proteins were assigned into a phyletic-age category: mammalian-specific proteins, metazoan-specific proteins, eukaryote specific proteins and universal proteins – present in prokaryote species. The universal and eukaryotic specific phyletic groups include proteins that their encoding genes are estimated to be found in the genome of a unicellular ancestor of mouse. The genes encoding metazoan-specific and mammalian-specific proteins are estimated to emerge after the transition to multicellularity.

We compared the expression pattern of the different categories within 14 mouse tissues and studied the tendency of proteins within these groups to be tissue-specific or ubiquitous. For each protein we recorded its expression breadth according to the Absent/Present call in each tissue. The expression data used for the analysis is the same data used in the chapter 3 (main data set only, as described in section 3.3.1). The 14 tissues represent a distinct non-redundant subset of organs that was chosen out of a bigger dataset. The tissues are listed in Figure 20.

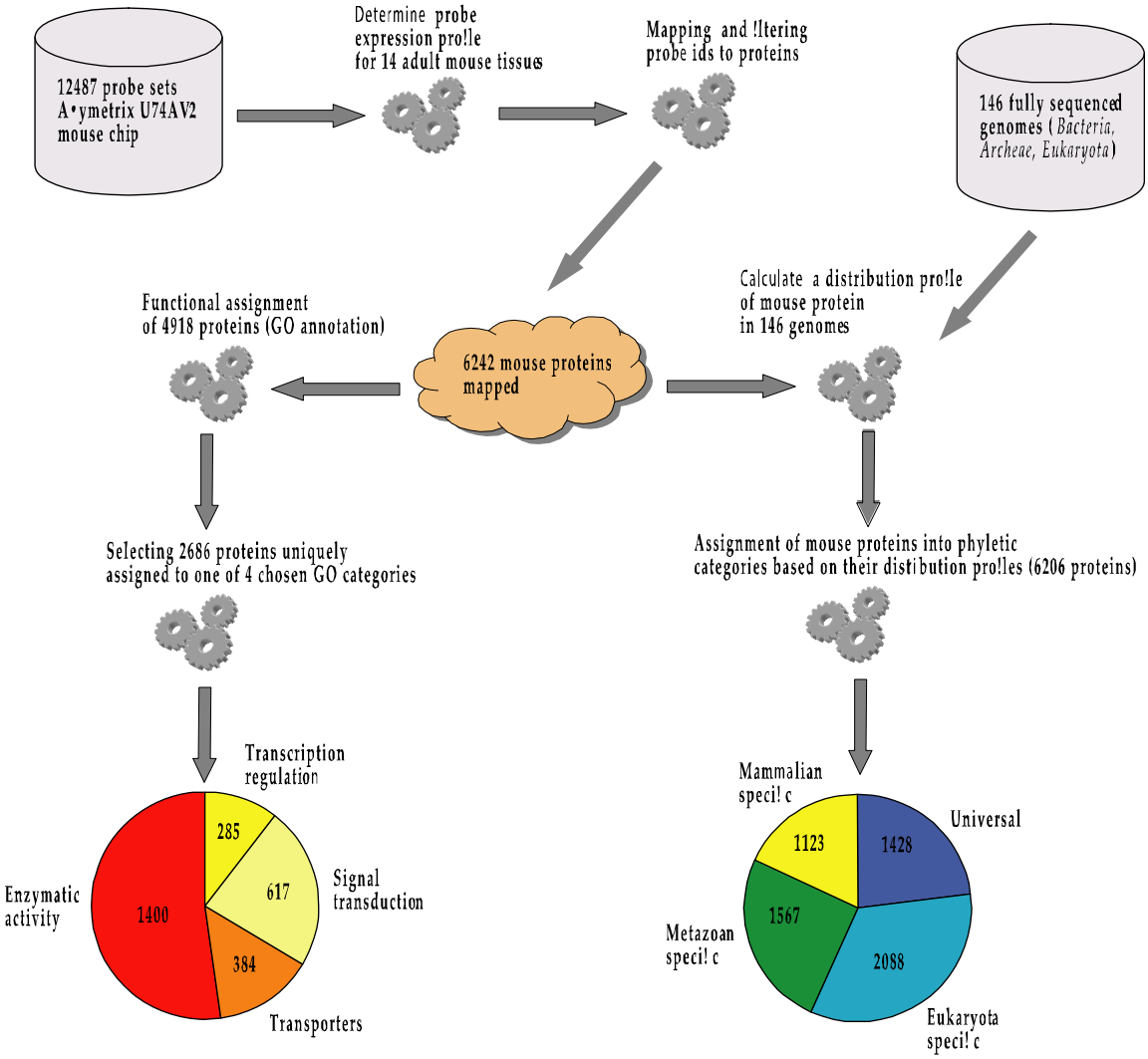

Figure 19 A schematic description of the expression profile determination and protein annotation as described in the methods section. The numbers in the circles indicate the number of proteins assigned to the relevant category.

5.3 Methods

5.3.1 Mapping probe sets to mouse proteins

8218 out of 12487 probe sets are mapped into Ensembl mouse transcripts using Ensmart (Kasprzyk et al. 2004) (version 13.1, http://www.ensembl.org/Multi/martview (Andrade et al. 1999)). Mapping of Ensembl transcripts to SWISS-PROT (Bairoch and Apweiler 2000) (release 41.25) and TrEmbl proteins (Release 24.13) were obtained from the

International Protein Index (IPI, http://www.ebi.ac.uk/IPI/IPIhelp.html(Tamames et al. 1996)). Proteins that are represented by more than a single probe set were discarded in order to avoid re-counting. Different proteins sharing the same probe sets were also eliminated (with the exception of splice variants). A single and unique probe set therefore represents each of the 6242 remaining proteins.

5.3.2 Functional annotations of mouse proteins

4918 proteins were assigned with a GOA annotation (Camon et al. 2003). Proteins assigned to more than a single category were discarded, leaving 2686 proteins distributed as follows: 1400 enzymes, 384 transporters, 617 proteins involved in signal transduction and 285 proteins that regulate transcription.

5.3.3 Phyletic assignments of mouse proteins

Four categories were used to describe the evolutionary origin of mouse proteins: universal proteins – i.e., ubiquitous in the 3 domains of life (bacteria, archaea and eukaryote), eukaryote specific proteins, metazoan-specific proteins and mammalian-specific proteins.

The 6242 mouse proteins were classified into the phyletic categories according to the results of a BLAST search (Altschul et al. 1997) against 146 fully sequenced species. A protein could only be assigned to a single category. The classification process is hierarchical: first proteins with hits to more than five prokaryote species are classified as universal; the remaining mouse proteins with at least a single hit to non-metazoa eukaryote are classified as eukaryote specific; remaining mouse proteins with at least a single hit to non-mammalian metazoa are classified as metazoan-specific; finally proteins recognising only other mammalian proteins are classified as mammalian-specific. The cut-off used was BLAST E value < 1e-3. Genomes were downloaded from the COGENT database (Janssen et al. 2003) (release 152).

The 6242 proteins are distributed in phyletic categories as follows: 1428 universal proteins, 2088 eukaryote specific proteins, 1567 metazoan proteins, 1123 mammalian-specific proteins; 36 unclassified proteins.

5.3.4 Calculating the evolutionary rate (d_N/d_S values) between mouse proteins and their orthologues in rat

Mouse and rat 1:1 orthologous pairs were obtained from Ensembl (Kasprzyk et al. 2004). In those cases where a mouse protein had more than a single rat orthologue it was discarded from the analysis unless one of the orthologous pairs was annotated as BRH (Best Reciprocal Hit). The ratio of d (the number of non-synonymous substitutions per non-synonymous site) to dS (the number of synonymous substitutions per synonymous site) was calculated using the codeml program from the PAML 3.13d package (Yang 1997). Two sequences had one or fewer nucleotide mutations and so their dN/dS ratio could not be reliably estimated. These sequences were discarded from the analysis. In total the dN/dS ratio was calculated for 4056 mouse proteins from the 5501 proteins classified to a phyletic category and expressed in at least a single tissue.

5.4 Comparing expression patterns within different tissues

For each tissue, the number of expressed probe sets was counted. The fraction of probe sets expressed in each tissue ranges from 0.35 (muscle) to 0.55 (eye). Approximately a constant fraction (~60%) of the probe sets in each tissue is mapped to proteins. Similarly a constant fraction (~45%) of the proteins in each tissue can be assigned a GO annotation (Figure 20A). We compared the tissues for their content of functional and phyletic groups (Figure 20B, Figure 20C). All tissues display a strikingly similar functional and phyletic composition. The functional composition of annotated proteins in a tissue is approximately 60% enzymes, 20% transporters, 15% signal transduction proteins and 5% transcription regulation proteins. The phyletic composition of proteins in a tissue is found to be approximately 25% universal proteins, 40% eukaryotic specific proteins, 20% metazoan-specific proteins and 15% mammalian-specific proteins.

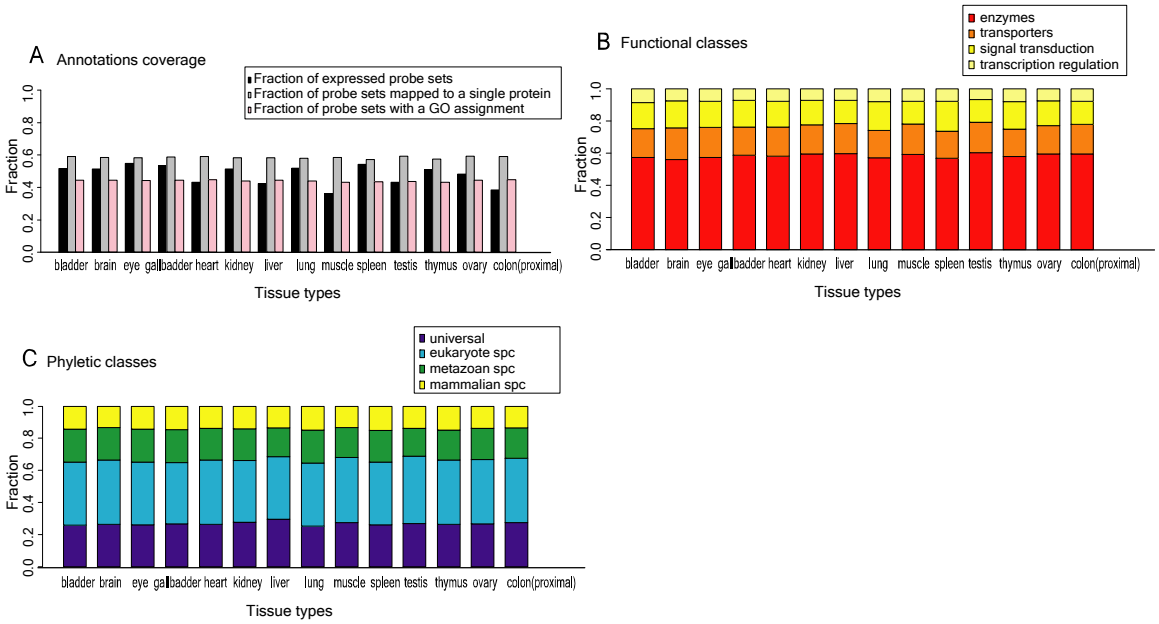

Figure 20 Expression profiles of various tissues. (A) Fraction of expressed probe sets (out of all 12,488 probe sets) in a tissue (black). The grey bars represent the fraction of expressed probe sets in a tissue that can be uniquely mapped to a single protein. The pink bars represent the fraction of expressed probe sets in a tissue that are assigned with a GO annotation. (B)bDistribution of the four functional categories of all annotated proteins expressed in a tissue. (C) Distribution of the four phyletic categories of all the mapped proteins expressed in a tissue.

Since the tissues seem to have almost identical overall composition of functional categories (Figure 20B) tissue diversity must be achieved through differences in the protein composition within each different category. We counted the number of proteins expressed in one tissue, two tissues etc (Figure 21A). About a third of the proteins are expressed in all tissues examined, so variation is seen for 2/3 of the proteins in the sample.

5.5 Comparing expression patterns within functional and phyletic categories

We further studied the contribution of different functional and phyletic groups to variations between tissues. Are some functional categories more tissue-specific than others? We examined the expression profile of proteins from the four functional categories within the different mouse tissues. For each group, the fraction of its protein members expressed in one tissue, two tissues etc was calculated (Figure 21B). Surprisingly less than 1/3 of the enzymes and transporters are ubiquitously expressed in all tissues examined. The fraction is

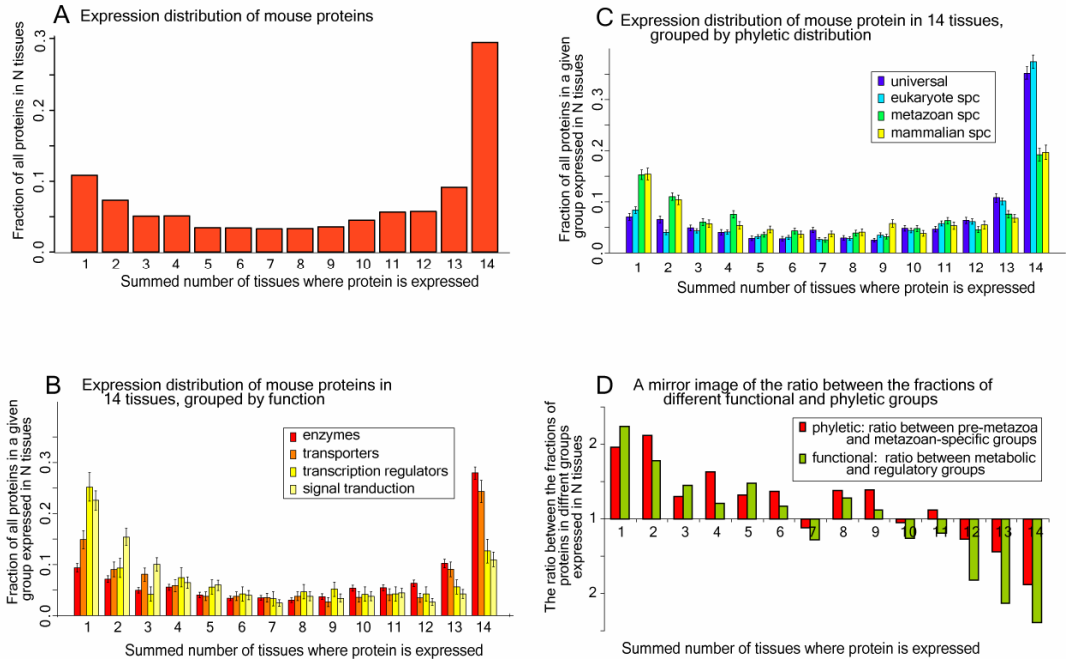

Figure 21 The fraction and the relative fraction of proteins in a group that are expressed in N tissues. (A) Expression pattern of all mapped proteins expressed in at least a single tissue (5528 out of 6242 proteins). (B,C) Expression pattern of proteins in different functional groups (B) and phyletic groups (C). The error bars indicate the standard error estimate using bootstrap resampling. (D) The ratio between the fractions of the different groups shown in B and C. The values for the functional groups are those shown in B where enzymes and transporters are merged (metabolic group) and transcription regulation and signal transduction are merged (regulatory group); The values for the phyletic groups are those shown in C where universal and eukaryotic-specific are merged (pre-metazoa group) and metazoan-specific and mammalian-specific are merged (metazoan-specific group). The values presented are always those retrived when the larger fraction is divided by the smaller fraction. The analysis is restricted to those proteins expressed in at least a single tissue. Sample size (number of assigned proteins): enzyme - 1294; transporters – 343, signal transduction – 450; transcription regulation – 214; universal – 1359; eukaryote specific - 1951; metazoan-specific – 1281; mammalian-specific – 910.

even lower for the other functional groups where only about 1/10 of the transcription factors and signal transduction proteins are expressed in all tissues examined. Two different patterns of expression are observed: the relative abundance of enzymes and transporter proteins is higher among proteins that are ubiquitously expressed. In contrast a larger fraction of transcription factors and signal transduction proteins are tissue-specific (Figure 21D).

Signal transduction proteins and transcription factors are known to be the main functional categories that were expanded in the metazoa lineage while enzymes and

transporter proteins are usually more highly conserved between the different domains of life (Aravind and Subramanian 1999; Chervitz et al. 1998). Therefore, unsurprisingly, the distribution of the functional groups in the data set largely correlates with the phyletic clusters (Table 12).

Reproducing the expression data charts using the phyletic groups naturally reveals the trend predicted from Table 12 - the relative abundance of universal and eukaryote specific proteins is higher among these proteins which are expressed in a wide variety of tissues while mammalian-specific proteins have a higher tendency to be tissue-specific (Figure 21C,D). The observations are compatible with those obtained in a recent study where metazoan-specific protein-domains and protein-domains involved in intercellular communication were shown to be tissue-specific (Lehner and Fraser 2004).

As the expression of the functional (Figure 21B) and phyletic groups ((Figure 21C) represent two sides of the same coin, the remaining question is whether enzymes tend to be ubiquitously expressed due to their phyletic-universal origin or whether phyletic-universal proteins tend to be ubiquitously expressed due to being enzymes.

Table 12 The distribution of function within the phyletic groups.

Phyletic groups\functional groups	Total functionally annotated proteins	%Enzymes	%Transporters	%Transcription factors	%Signal transduction
Universal	833 (31%)	82	15	2	1
Eukaryota specific	823 (31%)	60	18	12	10
Metazoan-specific	656 (24%)	26	7	23	44
Mammalian-specific	372 (14%)	13	17	6	64
All	2684	52	14	11	23

5.6 The inter-relationship between function, 'phyletic age' and expression

In order to identify whether a protein's expression pattern better reflects function or age, the inter-relationship between these three factors was compared statistically using a

contingency table test under the null hypothesis that function and specificity are independent given age. The test was suggested and performed by Tim Massingham.

Conceptually, the genes are divided up according to age and a separate contingency table for specificity and function is formed for each group. The chi-squared test statistic (Howell 1992) for independence between function and specificity is calculated for each table and then pooled, weighted by proportion of genes of each age, to give the test statistic for independence between function and specificity given age. The dependence between specificity and age given function, and age and function given specificity, were calculated similarly. The tables analyzed had cells expected to contain a small number of observations, so it was inappropriate to assess the significance of the test statistic using tables of pre-calculated critical values. Instead 10,000 sets of data, of equal size to that observed, were generated in accordance to the expected contingency table and the test statistics of these were used to form an estimate of their distribution under the null hypothesis, to which the observed test statistic can be compared.

After phyletic age was taken into account, only a weak dependence between function and tissue specificity was detected (test statistic 264.7, P value 0.04) suggesting that most of the relationship observed between function and tissue specificity is accounted for by the age of the gene. The relationship between phyletic age and tissue specificity is not explained by a gene's function (test statistic 339.1, P value < 0.0001), nor is the relationship between phyletic age and function explained by the tissue specificity (test statistic 967.2, P value$<<0.0001$). Therefore the results imply that enzymes and transportrs tend to be ubiquitously expressed mainly due to their phyletic universal origin (rather than due to their functional classification).

In order to show the extent to which ancient metabolic proteins are widely expressed, or young regulatory proteins are tissue specific, we have divided the functional groups according to their phyletic groups (Figure 22). To have a sufficient sample size for the bootstrap error analysis in Figure 22, the four functional categories were merged into two: metabolism (enzymes and transporters) and regulation (transcription factors and signal transduction). The expression pattern of the two functional categories was examined in two phyletic groups: the pre-metazoa group (universal and eukaryote specific groups) and the metazoan-specific group (metazoan and mammalian-specific proteins) (Figure 22).

From the expression distribution of metabolic proteins (enzymes and transporters, Figure 22A) one can observe, as can be inferred from the statistical test reported above, obvious differences between the 'older' pre-metazoa proteins (universal and eukaryotic specific groups) and the more recent metazoan proteins. Differences can be observed for the regulatory proteins as well (Figure 22B): metazoan-specific proteins tend to be more tissue-specific compared to pre-metazoan ones, regardless of their functional class. The tendency of pre-metazoan prteins to be globbaly expressed and of metazoan-specific proteins to be specifically expressed is therefore observed in both functional groups (Figure 22C). in A notable difference between the expression patterns in Figure 22A and Figure 22B occurs for specifically expressed pre-metazoan proteins, where the proportion of regulatory proteins is much higher than metabolic proteins, and it is not significantly different from the fraction of metazoan specific proteins (Figure 22C). This confirms that the function has some influence on expression, independent of age, but suggests that the effect is stronger in specifically expressed pre-metazoan proteins.

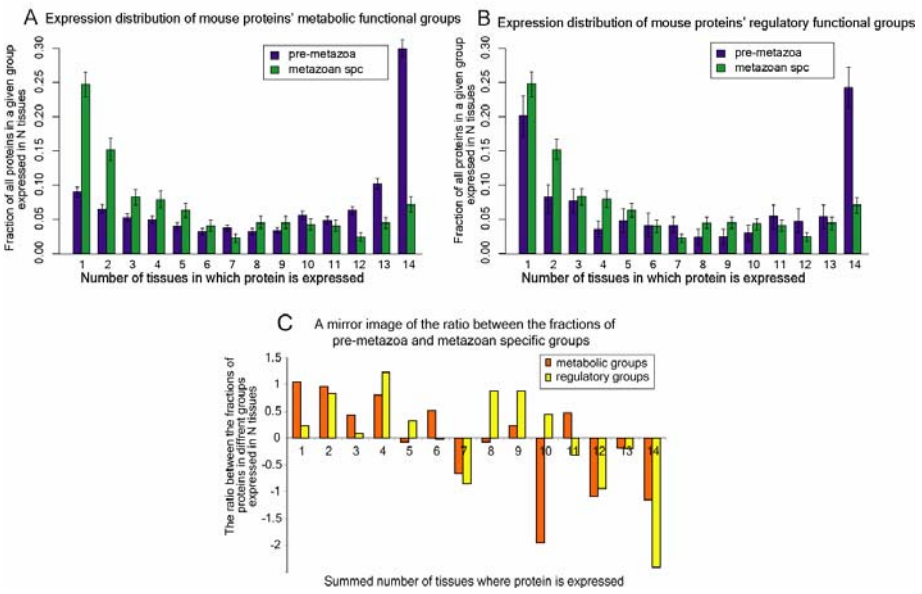

Figure 22 Expression pattern of proteins in different functional groups. The pre-metazoan group includes universal proteins and eukaryote specific proteins. The metazoan-specific group includes metazoan-specific and mammalian-specific proteins. The metabolic functional group includes enzymes and transporter proteins. The regulatory group includes signal transduction and transcription regulation proteins. (A,B) The error bars indicate the standard error estimate using bootstrap resampling. (C) The ratio between the fractions of the different groups shown in A and B. The values for the metabolic groups are those shown in A; The values for the regulatory groups are those shown in C. The values presented are always those retrived when the larger fraction is divided by the smaller fraction. The analysis is restricted to those proteins expressed in at least a single tissue. Sample size (number of assigned proteins): pre-metazoan metabolic proteins – 1370; metazoan metabolic proteins – 267, pre-metazoan regulatory proteins – 169; metazoan regulatory proteins – 493.

5.7 The inter-relationship between 'phyletic age', evolutionary rate and expression

Tissue-specific genes tend to evolve more rapidly than broadly expressed ones (Duret and Mouchiroud 2000; Winter et al. 2004; Zhang and Li 2004). Therefore difficulties might arise in identifying their distant homologues, leading to a correlation between age and rate of evolution. To rule out the possibility that the connection between age and expression is spurious due to the misclassification of rapidly evolving genes and the connection between tissue expression and recent rate of evolution, Subramanian and Kumar (Subramanian and Kumar 2004) showed that the connection still exists in a slowly evolving set of data. However, the assumption that the slow rate of evolution of the genes assumes a correct age classification may not hold if there has been a change in rate during their evolutionary history, for example diversifying selection followed by conservation after a gene duplication (Hughes 1999). Therefore, there is a need for a direct test to show that the connection between phyletic age and tissue expression of a gene cannot be explained by the connection between rate and tissue expression alone, a test which does not assume homotachy and makes use of all the available data.

Such direct test was suggested and performed by Tim Massingham who studied the expression profile of different phyletic groups in a subset of the data where all phyletic groups have evolved at approximately the same rate. By grouping the genes into equal bins of similar recent rate of evolution (d_N/d_S values) and then shuffling the phyletic ages within each group, sample sets of data can be created which have no connection between phyletic age and tissue expression other than through their common connection to recent rate of evolution. The connection between rate and tissue specificity in each of these sample sets is identical to the observed data and, because all genes in each bin have a similar rate of evolution, the connection between age and rate is similar to the observed data. By generating random samples in the manner as described above, the expected contingency table of age/expression dependence and the null distribution of the chi-squared test statistic can be estimated. Using the estimated expected table, the chi-squared statistic for the observed data can be calculated. The significance of the observation was assessed by comparing the observed test statistic to those from 10,000 sets of data, of equal size to the observed data, randomly generated according to the expected contingency table (and so satisfying the null

hypothesis). The chi-squared test statistic for the independence of age and tissue expression given rate in the data was 226.5, whereas the maximum observed statistic in 10,000 random draws, generated as described in the methods section, was 84.3.

Therefore, the connection between phyletic age and tissue specificity that describes here cannot be explained purely in terms of both factors' mutual correlation with the recent rate of evolution.

5.8 Examples for tissue-specific, pre-metazoa enzyme-proteins

Although pre-metazoa proteins tend to be more widely expressed, less than a 1/3 of the pre-metazoa metabolic proteins are expressed in all tissues. Ldhc (testis specific lactate dehydrogenase) is one example of a universal enzyme whose expression is limited to a few cell types in mammals. Ldh participates in anaerobic glycolysis – a nearly universal pathway that converts glucose into pyruvate. The sequence of reactions in the pathway is similar in all organisms and in all cell types. In contrast, the fate of pyruvate is variable. In a variety of microorganisms, lactate is normally formed from pyruvate in a reaction catalysed by Ldh. In higher organisms most cells do not convert pyruvate to lactate and the reaction is limited to a few tissues (Stryer 1995). In germ cells, where lactate is a preferred energy source (Goddard et al. 2003), there is a specific expression of Ldhc (testis specific expression in the expression data used here). The expression of Ldhc is an example of a function occurring in the ancestral unicellular cell that becomes tissue-specific in multicellular species.

The testis specific expression of two other universal enzymes in our data set – Glucose-6-phosphate dehydrogenase 2 (*G6pd-2*) and Phosphoglycerate Kinase 2 (*Pgk-2*) – provides a different example for a specific expression of universal enzymes. *G6pd-2* and *Pgk-2* are believed to arise from their isoenzymes, *G6pd* and *Pgk-1* respectively, by a gene duplication event. *G6pd* and *Pgk-1* are essential, widely expressed, X-chromosome encoded genes. The absence of those two enzymes during the inactivation of the X chromosome in postmeiotic spermatogenic cells is compensated by the expression of their autosomal testis-specific isoenzymes *G6pd-2* and *Pgk-2* (Boer et al. 1987; Hendriksen et al. 1997). Duplication events can therefore explain some of the cases where universal enzymes are specifically expressed.

5.9 Caveats

Several limitations of this analysis must be acknowledged. Firstly, the analysis performed is based only on approximately one-fifth of mouse proteins (proteins that are included in the main expression data) and the expression values are based on a single replicate. In order have a better reassurance that the observations reported in the chapter are not the result of the use of limited, random subset of genes the analysis was repeated while using a different set of tissues (seven components of the gastrointestinal track). The results obtained while using this different subset of genes are compatible with the observations reported here.

Secondly, only Absent/Present calls were used to define expression-specificity. Absent/Present flags are determined by the Microarray Suite 5.0 package (Affymetrix MAS 5.0) and its default setting (*P* value = 0.05). In order to preclude the possibility that the results reflects the choice of cut-off the analysis was repeated under different cut-offs within the range of $0.025 < P \text{ value} < 0.075$ for Absent/Present flags labelling, showing that the choice of cut-off have no effect on the analysis.

Thirdly, the classification of proteins into phyletic categories is based on the results of a BLAST search. Several concerns can be raised about the classification procedure.

- In order to preclude the possibility that the results reflects the choice of cut-offs the analysis was repeated under different cut-offs:
  - ❖ The observation reported here using a cut-off of *E* value < 1e-3 are maintained using different cut-offs within the range of 1e-10 < *E* value < 1e-1. Similar results were also obtained when using the homologous clusters database STRING (von Mering et al. 2003) for a phyletic classification.
  - ❖ The classification of proteins as universal requires the recognition of homologue proteins in at least five prokaryote species (as detailed in section 5.3.3). The observations are maintained when a universal protein is defined as protein with a hit in at least a single prokaryote species or when it is defined as a protein with a hit in at least 10 prokaryote species (examined under cut-off of *E* value < 1e-3).
- The analysis performed here aims to study the expression distribution of proteins, rather than the distribution of protein-domains. The data were tested in order to assure that the trends reported here are truly derived from the expression pattern of full-

sequence proteins, by repeating the analysis while using additional filters for the classification of a protein into a phyletic group. The following filters were added in order to confirm full homology between query and hit (where the cut-off for the BLAST search is *E* value < 1e-3):

- ❖ Pfam domain composition: all query-hit pairs that do not share an identical Pfam (Bateman et al. 2004) domain composition were discarded from the data set.
- ❖ Full coverage: all query-hit pairs where the alignment does not cover almost the full length (80%) of both proteins were discarded from the data set.

When repeating the analysis with the filtered data the results confirm that the trends reported here are maintained.

5.10 Conclusions and discussion

The analyses described in this chapter demonstrate that multicellular specific proteins tend to be more tissue-specific than 'ancient' universal proteins. Most of the 'late' evolutionary proteins are transcription factors and signal transduction proteins, categories that have previously been suggested to play a crucial role in tissue differentiation. However, the analyses suggest that more recent enzymes and transporters also contribute to tissue diversity as many of them are tissue-specific (Figure 22A). The selective expression pattern of recent genes implies that a new protein is often selected to perform a tissue-specific function rather than a global one. A greater evolutionary flexibility of tissue-specific proteins is compatible with previous studies suggesting that tissue-specific proteins evolve more rapidly (Duret and Mouchiroud 2000; Winter et al. 2004; Zhang and Li 2004) due to less strict functional constraints compared to broadly expressed proteins (Duret and Mouchiroud 2000; Hastings 1996).

Despite this trend, many metazoan-specific proteins are ubiquitous and many universal proteins are tissue-specific. The minimal cellular transcriptome of the metazoan cell differs from that of the ancestral unicellular eukaryote: new functions were added (metazoan-specific proteins), whilst other functions became specialised and no longer take place in all cells (tissue-specific pre-metazoa proteins). The extent of the cellular specialisation can be implied from the observation that only 1/3 of the proteins are expressed in all tissues examined. In some of these cases, functions occurring in the

unicellular cell become tissue-specific in multicellular species. In other cases, universal genes that have been duplicated become specific to a tissue whilst a second copy maintains its original expression pattern.

Only about 1/3 of the pre-metazoa metabolic enzymes are expressed in all tissues. Tissue differentiation is at least in part achieved by tissue specialisation of metabolism – either by differentially expressing 2/3 of the ancient metabolic proteins, or by encoding new metabolic proteins. The next chapter concerns the characterisation of the phyletic origin of the mammalian metabolic genes, and the study of the relationship between the evolution of this metabolic set and the development of new, tissue-specific, pathways.

Chapter 6: The phyletic origin of metabolic pathways in mammals

6.1 Overview

Comparative genome studies use the complete sequence of species from different lineages in order to investigate the molecular basis behind the appearance of lineage-specific phenotypes. Several such studies have attempted to elucidate the origin of the eukaryotic cell, the appearance of multicellular life forms, and the evolution of higher animals. Eukaryotic-specific genes are involved in the organisation of DNA within the nucleus, cytoskeleton compartmentation, membrane transport, cell-cycle control, regulatory processes, and RNA splicing (Wood et al. 2002). Metazoan-specific gene families, or gene families whose expansion is specific to the metazoan lineage, include transcription factors (participating in spatial and developmental differentiation), signalling molecules (respond to inter-cellular communication), and adhesion molecules (hold the cells together) (Aravind and Subramanian 1999). Vertebrate specific genes are involved in defence and immunity, and in the nervous system. Gene families whose expansion is specific to the vertebrate lineage include the immunoglobulins, families involved in control of development, and intermediate filament proteins such as keratin and olfactory receptor (Lander et al. 2001).

Comparative studies provide a global view of the innovations in the gene repertoire in a lineage, and they are most useful in identifying novel gene families, or families that have been massively expanded. More subtle changes, such as a modification that enables an enzyme-protein to utilise a new substrate, cannot, in many cases, be detected in such sequence-based analysis. For enzymatic proteins, several schemes link sequence and function (reaction, pathway). Therefore, for enzymes, it is possible to perform a function-based comparative study.

This chapter describes a function-based comparative study of the metabolic reactions and metabolic pathways in mammals. Both metabolic reactions and metabolic pathways were classified as universal, eukaryotic-specific, metazoan-specific and mammalian-specific, according to their distribution in various fully sequenced eukaryote and prokaryote species. Firstly, the phyletic origin of the metabolic reactions repertoire of mammals is analysed. Secondly, the construction of the pathway repertoire in a species performed here is

described. Thirdly, the innovations in the pathway repertoire of mammals are identified and characterised. The main motivation behind this analysis was to investigate how innovations in the metabolic repertoire of mammals are related to the evolution of tissue-specific functions.

The work presented in this chapter is preliminary, and it is mostly based on data gathered in 2002-3. Caveats of this analysis are discussed in the summary section (6.6).

6.2 Data

All data used in this analysis was downloaded from the KEGG database (Kanehisa et al. 2002) in August 2003. KEGG's pathway diagram contained 119 metabolic pathways. The pathways are organised according to the compilations of the Japanese Biochemical Society and the wall charts of Boheringer Mannheim. They are continuously updated according to biochemical evidence. Information was available for 93 fully sequenced species including 7 eukaryotes, 70 bacterial species, and 16 archaeal species. When downloaded, the KEGG catalogue contained the information for only about 2/3 of the human and mouse protein repertoire (17,988 and 13,499 proteins respectively).

For each reaction (four-digit EC number) KEGG provides predictions for its distribution in species. These predictions were used to assign human metabolic reactions with a phyletic age. The procedure is the same as described in chapter 4. Shortly, mammalian specific reactions are reactions found only in mammals; metazoan specific reactions are those reactions that are found only in metazoa (but they are not mammalian specific); eukaryote specific reactions are reactions that are found only in eukaryote species (but they are not metazoan specific); universal reactions are those mammalian reactions that are also found in at least 5 prokaryote species. Out of the 533 human metabolic reactions, 46 reactions were not assigned to any phyletic category.

Networks represantation was done using the Biolayout software (Goldovsky et al. 2005).

6.3 The origin and composition of the mammalian reaction set

487 human metabolic reactions were classified according to their phyletic age. 50% of the reactions are classified as universal, 15% as eukaryotic specific, 11% as metazoan specific, and 24% as mammalian specific (Figure 23). The high fraction of universal reactions in mammals indicates a high level of conservation of the enzymatic sets, which is compatible with previous studies that reported the existence of an extensive conserved core of metabolic enzymes common to archaea, bacteria and eukaryota (Peregrin-Alvarez et al. 2003). The distribution of the remaining reaction groups indicates that a more limited set of reactions is eukaryotic or metazoan specific, suggesting that the phylogenesis of these groups did not involve a massive recruitment of new reactions. Surprisingly, a relatively large group of reactions (24%) is mammalian specific. The results are compatible with these obtained from the analysis of the dataset used in chapter 2.

6.3.1 The reaction class distribution of the different phyletic groups

The EC scheme divides reactions into six classes (which are represented by the first digit in the EC number): oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. The reaction class distribution of each phyletic group is shown in Figure 23. The most conserved reaction class is the ligase class – almost all reactions are universal. Oxidoreductases are the largest reaction class within the metazoan-specific reactions whereas the transeferases are the largest reaction class in all other phyletic groups. Mammalian specific reactions have the highest fraction of transferases. These distributions were obtained using all 4 levels of the EC classification. Similar distributions result when clustering reactions to the third digit of the EC classification.

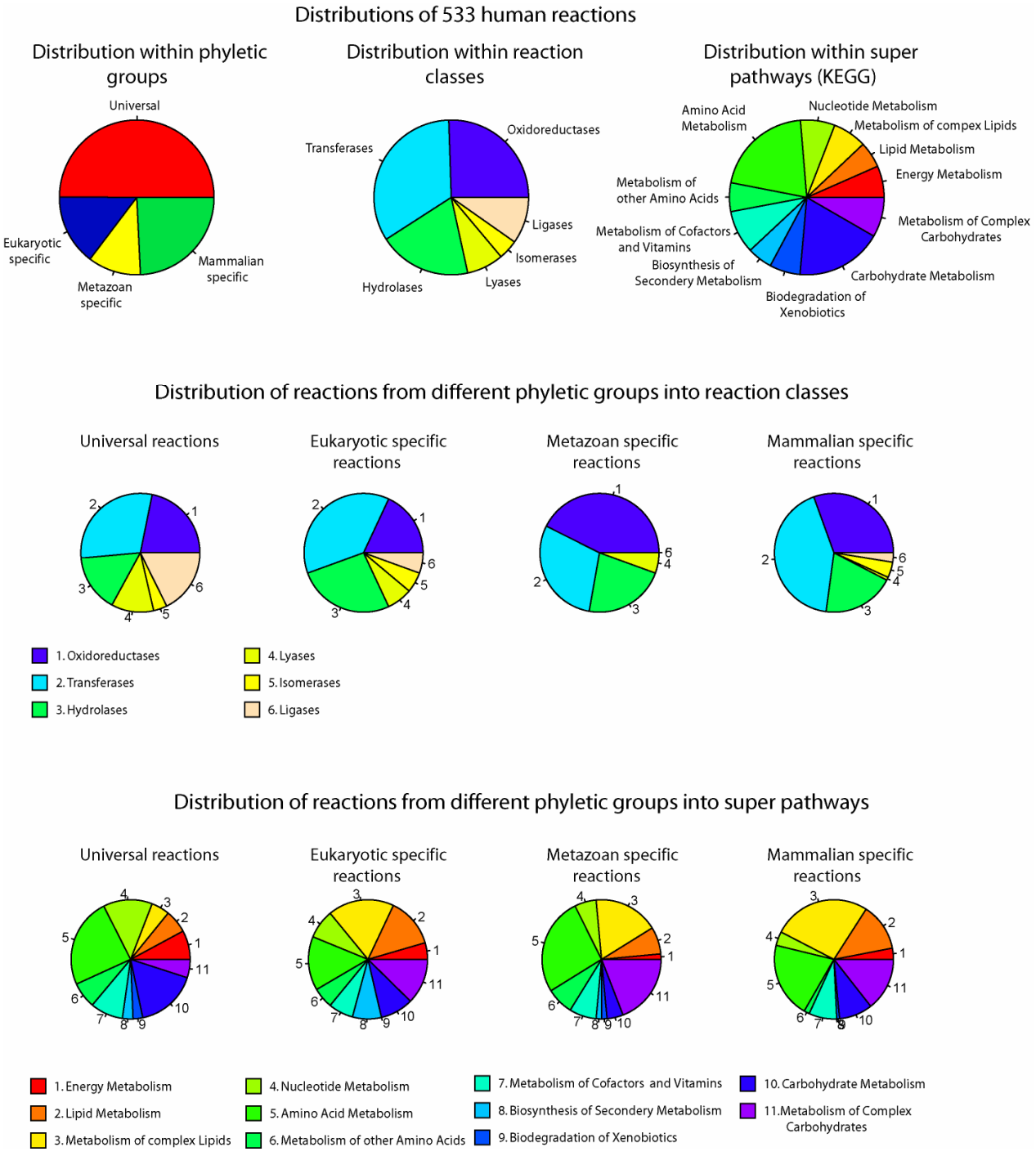

Figure 23 The classification of mammalian reactions into phyletic groups, reaction classes (according to the EC scheme), and super-pathways of metabolic reactions (according to the KEGG database). 46 reactions which are not assigned to any phyletic group were discarded from those charts showing the distribution of phyletic groups (see section 6.2). The number of different reactions is calculated using all four levels of the EC classification.

6.3.2 The metabolic super-pathway distribution of the different phyletic groups

The pathways in KEGG are classified into 11 super-pathways. The super-pathways distribution of each phyletic group is shown in Figure 23. The most ancient pathways are these involved in the metabolism of nucleotides, where the large majority of reactions are universal (26 out of 34). The most recent pathways are these involved in the metabolism of complex lipids, where most of the human reactions have evolved after the emergence of the eukaryotic cell.

We looked in detail into the classification of reactions to pathways (Table 13). Interestingly, most pathways seem to evolve gradually, where each lineage contributes a few reactions. Exceptional cases are the "pentose phosphate pathway", "glutathione metabolism" and "selenoamino acid metabolism" categories where all reactions are universal and possibly conserved since the divergence from a universal ancestor.

6.4 Construction and analysis of the pathway repertoire of a species

6.4.1 Constructing the pathway repertoire of a species

The KEGG database provides the list of reactions assigned to a pathway, and the distribution of each reaction in species. Given the available annotations, only a few pathways are 'complete' in any genome (i.e., proteins assigned to all of the reactions in the KEGG pathway that are present in a species). Therefore, in order to predict the pathway repertoire of a species it is necessary to define a cut-off for the number of reactions in a pathway that are found in a species, which is sufficient to consider a pathway to be present in the species. A cut-off is needed in order to prevent the classification of false positive pathways, which are not truly represented by the reaction set. Yet, assuming that all the metabolic reactions in a species participate in pathways, such a cut-off should enable as large a fraction as possible of the reactions to be included.

Table 13 The distribution of human reactions within metabolic-pathways. Numbers in brackets indicate the number of reactions unique to the pathway. T- total; U – universal; E – eukaryota specific; Z – metazoan-specific; M – mammalian specific; NC – not classified.

Phyletic Class	Super Pathway	Pathway	T	U	E	Z	M	NC
U	Energy metabolism	Nitrogen metabolism	9(1)	6(1)	0(0)	0(0)	1(0)	2(0)
		Oxidative phosphorylation	7(4)	5(3)	1(0)	0(0)	0(0)	1(1)
		Photosynthesis	2(1)	2(1)	0(0)	0(0)	0(0)	0(0)
	Lipid metabolism	Fatty acid biosynthesis	9(7)	5(4)	3(2)	1(1)	0(0)	0(0)
		Fatty acid metabolism	16(5)	9(1)	1(1)	1(0)	2(2)	3(1)
		Sterol biosynthesis	15(9)	8(4)	5(3)	0(0)	1(1)	1(1)
	Complex lipid metabolism	Glycerolipid metabolism	32(14)	14(5)	7(3)	7(3)	2(2)	2(1)
		Phospholipid degradation	4(1)	1(1)	1(0)	2(0)	0(0)	0(0)
	Nucleotide metabolism	Purine metabolism	47(24)	30(15)	7(5)	3(2)	3(0)	4(2)
		Pyrimidine metabolism	33(14)	22(11)	2(0)	2(0)	3(1)	4(2)
	Amino acid metabolism	Arginine and proline metabolism	29(6)	17(3)	2(0)	4(2)	4(0)	2(1)
		Glycine serine and threonine metabolism	27(10)	15(5)	2(1)	3(1)	5(2)	2(1)
		Tryptophan metabolism	23(10)	9(1)	3(1)	4(4)	5(4)	2(0)
		Urea cycle and metabolism of amino groups	18(4)	12(3)	2(1)	0(0)	4(0)	0(0)
		Valine leucine and isoleucine degradation	17(3)	11(3)	0(0)	0(0)	1(0)	5(0)
	Other amino acid metabolism	Glutathione metabolism	9(3)	9(3)	0(0)	0(0)	0(0)	0(0)
	Metabolism of co-factors and vitamins	Biotin metabolism	5(5)	1(1)	3(3)	0(0)	1(1)	0(0)
		Folate biosynthesis	10(6)	7(3)	1(1)	1(1)	0(0)	1(1)
		Nicotinate and nicotinamide metabolism	11(6)	4(2)	0(0)	0(0)	4(2)	3(2)
		One carbon pool by folate	15(4)	10(1)	1(1)	1(1)	1(0)	2(1)
		Porphyrin and chlorophyll metabolism	14(11)	8(7)	1(1)	1(0)	2(2)	2(1)
		Vitamin B6 metabolism	3(1)	2(1)	0(0)	0(0)	1(0)	0(0)
	Carbohydrate metabolism	Citrate cycle (TCA cycle)	15(2)	10(1)	3(1)	0(0)	0(0)	2(0)
		Fructose and mannose metabolism	15(9)	9(5)	3(2)	1(0)	2(2)	0(0)
		Galactose metabolism	14(3)	8(1)	1(0)	1(0)	4(2)	0(0)
		Glycolysis / Gluconeogenesis	25(3)	18(1)	2(0)	0(0)	3(2)	2(0)
		Glyoxylate and dicarboxylate metabolism	9(2)	7(1)	0(0)	0(0)	2(1)	0(0)
		Pentose phosphate pathway	13(5)	12(4)	0(0)	0(0)	0(0)	1(1)
		Pyruvate metabolism	21(3)	16(3)	0(0)	1(0)	2(0)	2(0)
	Metabolism of complex carbohydrates	Aminosugars metabolism	12(7)	4(1)	2(1)	0(0)	3(2)	3(3)
		N-Glycan degradation	8(5)	3(1)	1(1)	1(1)	0(0)	3(2)
		N-Glycans biosynthesis	15(12)	2(2)	6(6)	4(2)	3(2)	0(0)
		Starch and sucrose metabolism	20(10)	11(6)	2(1)	2(1)	3(1)	2(1)
E	Energy metabolism	Sulfur metabolism	7(3)	2(0)	1(1)	1(1)	3(1)	0(0)
	Lipid metabolism	Bile acid biosynthesis	11(4)	3(0)	1(1)	1(0)	6(3)	0(0)
	Complex lipid metabolism	Prostaglandin and leukotriene metabolism	15(13)	1(0)	1(1)	2(1)	11(11)	0(0)
		Sphingoglycolipid metabolism	12(7)	2(0)	2(2)	1(1)	4(3)	3(1)
		Sphingophospholipid biosynthesis	1(1)	0(0)	1(1)	0(0)	0(0)	0(0)
	Amino acid metabolism	Alanine and aspartate metabolism	18(2)	13(0)	2(1)	1(0)	2(1)	0(0)
		Methionine metabolism	8(1)	5(0)	1(1)	0(0)	1(0)	1(0)
	Metabolism of complex carbohydrates	Glycosaminoglycan degradation	9(6)	2(0)	1(1)	0(0)	5(5)	1(0)
Z	Lipid metabolism	Androgen and estrogen metabolism	13(2)	0(0)	2(0)	3(1)	7(1)	1(0)
	Amino acid metabolism	Histidine metabolism	12(2)	2(0)	3(0)	1(1)	4(1)	2(0)
		Lysine degradation	15(6)	7(0)	1(0)	2(2)	4(4)	1(0)
		Phenylalanine, tyrosine and tryptophan biosynthesis	6(1)	4(0)	1(0)	0(0)	0(0)	1(1)
		Tyrosine metabolism	18(5)	5(0)	4(0)	3(2)	4(3)	2(0)
	Other Amino acid metabolism	Taurine and hypotaurine metabolism	5(1)	2(0)	0(0)	2(1)	1(0)	0(0)
	Biodegradation of xenobiotics	Tetrachloroethene degradation	1(1)	0(0)	0(0)	1(1)	0(0)	0(0)
	Metabolism of complex carbohydrates	Chondroitin/heparan sulfate biosynthesis	6(6)	0(0)	0(0)	4(4)	2(2)	0(0)
		O-Glycans biosynthesis	4(3)	0(0)	0(0)	2(2)	2(1)	0(0)
M	Lipid metabolism	C21-Steroid hormone metabolism	9(3)	0(0)	2(0)	0(0)	7(3)	0(0)
	Complex lipid metabolism	Blood group glycolipid biosynthesis-lact series	5(1)	0(0)	1(0)	0(0)	4(1)	0(0)
		Blood group glycolipid biosynthesis-neolact series	8(2)	0(0)	1(0)	1(0)	6(2)	0(0)
		Ganglioside biosynthesis	6(2)	0(0)	0(0)	0(0)	6(2)	0(0)

		Globoside metabolism	8(3)	2(0)	1(0)	0(0)	5(3)	0(0)
		Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	1(1)	0(0)	0(0)	0(0)	1(1)	0(0)
		Inositol phosphate metabolism	11(1)	2(0)	4(0)	2(0)	3(1)	0(0)
	Amino acid metabolism	Cysteine metabolism	7(1)	5(0)	0(0)	1(0)	1(1)	0(0)
		Lysine biosynthesis	3(1)	1(0)	1(0)	0(0)	1(1)	0(0)
	Metabolism of cofactors and vitamins	Retinol metabolism	2(2)	0(0)	0(0)	0(0)	2(2)	0(0)
	Biodegradation of xenobiotics	gamma-Hexachlorocyclohexane degradation	4(1)	3(0)	0(0)	0(0)	1(1)	0(0)
	Metabolism of carbohydrates	Butanoate metabolism	14(1)	8(0)	0(0)	1(0)	1(1)	4(0)
	Metabolism of complex carbohydrates	Keratan sulfate biosynthesis	4(1)	0(0)	0(0)	2(0)	2(1)	0(0)
NC	Other Amino acid metabolism	Selenoamino acid metabolism	9(1)	8(0)	0(0)	0(0)	0(0)	1(1)
	Metabolism of carbohydrates	Pantothenate and CoA biosynthesis	6(1)	3(0)	0(0)	1(0)	1(0)	1(1)

6.4.1.1 Calculating the number of reactions per pathway per species under different cut-offs

The pathways in KEGG are reference pathways – i.e., they include the complement of the relevant reactions found in all species. On average, a reference pathway in KEGG is composed of about 22 reactions and the standard deviation is about 18 reactions. A reaction in KEGG can be classified to several pathways, where more than half of the reactions are unique to a pathway. When considering all pathways that have at least a single detected reaction to be present in the species – almost half of all the pathways recorded in all species have less than three reactions per pathway per species (Figure 24A). About 20% of these pathways contain a single detected reaction. Under a cut-off of a single unique reaction, more than 90% of the pathways examined contain at least three reactions (Figure 24A).

6.4.1.2 Calculating the fraction of reactions per species that are assigned to "present" pathways under different cut-offs

For each species, we calculated the fraction of its reactions (out of all reactions in the species), that are assigned to a "present" pathway when using different cut-offs for the number of unique reactions. When a pathway is considered "present" according to the detection of at least a single, not necessarily unique, reaction (0 unique reactions), than all the reactions (100%) in all species are included in a "present" pathway (Figure 24B). When using a cut-off of a single unique reaction, most of the species examined have 80-100% of their

metabolic reactions assigned to a "present" pathway (Figure 24B). The numbers decrease significantly when using a cut-off of two unique reactions.

Under a cut-off of a single unique reaction, the large majority of reactions in species are assigned to a "present" pathway, where most pathways are assigned with at least three reactions. A lower cut-off results in a significantly lower number of reactions per pathway, where the fraction of participating reactions is not significantly larger. A higher cut-off results in a significantly lower fraction of participating reactions, but not a significant increase in the number of reactions per pathway. Therefore, a cut-off of a single unique reaction was chosen for the analysis done here. The pathway repertoire in the 93 species was constructed and analysed according to this cut-off. The possible caveats of this classification procedure are discussed in the summary section (6.6).

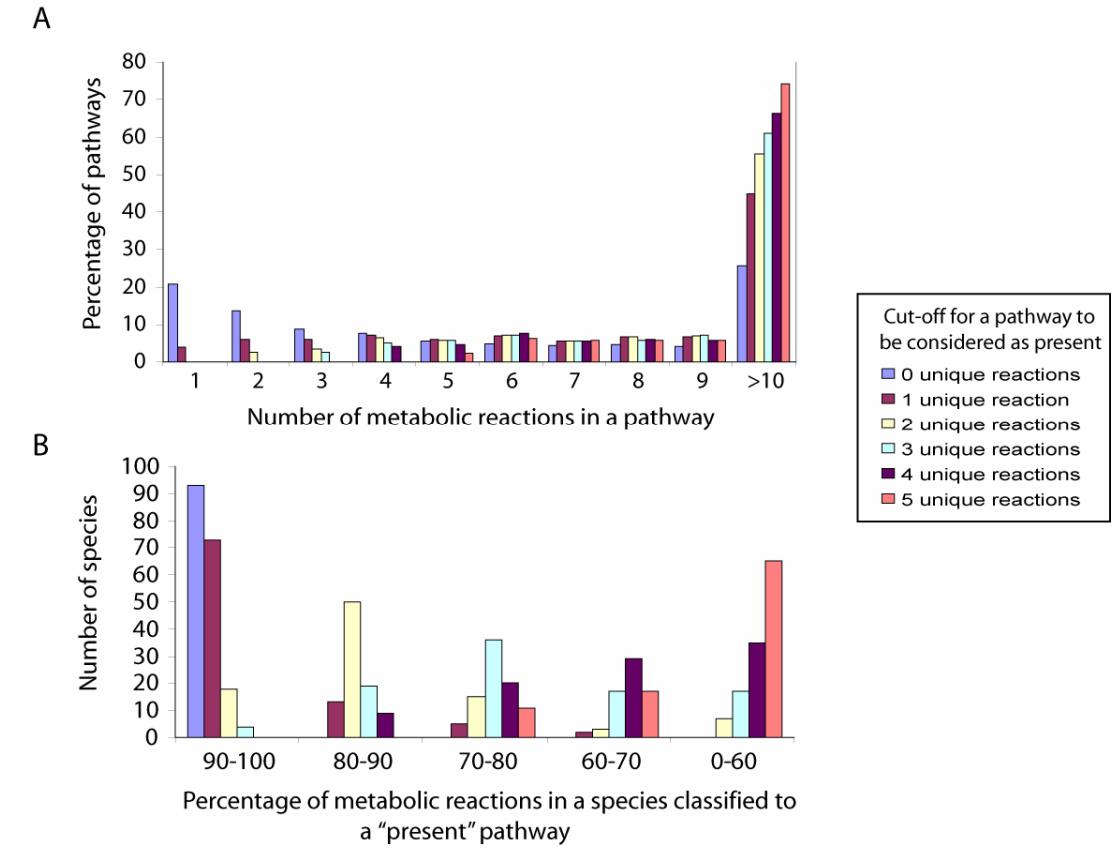

Figure 24 Examining the classification of reaction into pathways under different cut-offs. The cut-offs represnt the number of reactions in a pathway that are required to consider a pathway as present in a species. Unique reactions are reactions which are specific to a single pathway. Under cut-off of 0 unique reactions the presence of a single non-unique reaction in a species which is assigned to a pathway is sufficient to consider the pathway as present. The distribution of reactions in species was retrieved from KEGG. (A) The fraction of reactions in a species classified to a "present" pathway (out of all of the reactions in the species). (B) The distribution of the number of reactions per pathway per species out of all pathways in all species that are considered "present".

6.4.2 Analyses of the pathway repertoire in 93 species

6.4.2.1 The size of species' pathway repertoire

The largest number of pathways is observed in human (65 pathways) followed by five alpha-proteobacteria species (*B.japonicum*, *M.loti*, *S.meliloti*, *B.melitensi*, and *A.tumefaciens*). The smallest number of pathways (11 pathways) is found in *M.genitalium* whose genome is the smallest known genome of any free-living organism, followed by three other Mollicutes Bacteria species (*M.pneumoniae*, *U.urealyticum*, *M.pulmonis*).

Human not only has the highest number of pathways, but also the highest average fraction of reactions assigned per pathway (out of all reactions in the reference pathway) – on average almost half of the reactions on a reference pathway can be detected in human. The lowest average fraction, and also the lowest number of pathways, is detected in a Mollicutes species (U.urealyticum – 0.11). The partial completeness of pathways in all species, including human, is partially explained by the cumulative nature of the reference pathways in KEGG, which are a compilation of metabolic information from many species. It is also possible that some reactions that are present in a species are not detected due to extreme divergence which prevents their identification when using standard methods.

Both the number of pathways in a species and the average fraction of reactions per pathway are correlated with the number of reactions (Pearson's correlation coefficients > 0.9, Figure 25). The increase in the fraction of reactions per pathways accompanying the increase in the total number of reactions possibly reflects the way the pathways in KEGG have been designed, where each pathway can be divided into several modules that are suggested to act as evolutionary conserved functional units (Yamada et al. 2006).

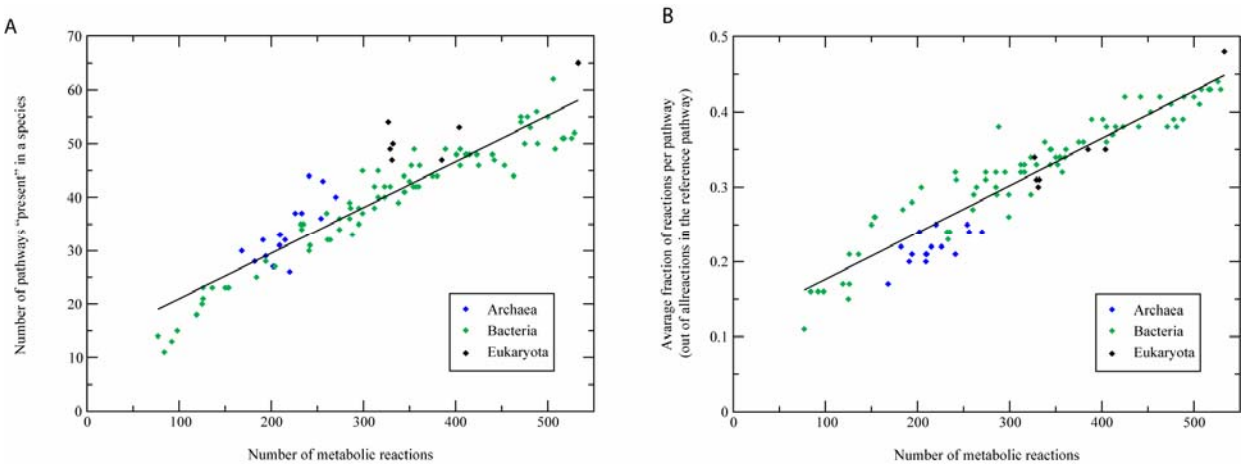

Figure 25 The number of metabolic-pathways and the average fraction of reactions identified per pathway in species plotted against the number of reactions identified for each species.

6.4.2.2 The distribution of pathways within species

The distribution of pathway reveals a vast diversity between species (Figure 27). None of the species have more than 65 "present" pathways (out of 119). Some of the super-pathways tend to be ubiquitous, and they are "present" in most of the species examined, while others are more species specific. The most ubiquitous pathways, i.e., - conserved in all species, are the nucleotide metabolism pathways where all three pathways are found in most species and two of the pathways (purine and pyramidine metabolism) are found in all species examined. Other widely distributed pathways (found in more than 100 species) are carbon hydrate metabolism pathways (glycolysis/gluconeogenesis, pentose phosphate pathway, fructose and mannose metabolism); Oxidative phosphorylation pathway (an energy metabolism pathway); biosynthesis of steroids and glycerolipid metabolism (lipid metabolism pathways); nicotinate and nicotinamide metabolism, folate biosynthsis and porphyrin and chlorophyll metabolism (metabolism of cofactors and vitamins).

The most specific pathways are those involved in biosynthesis of secondary metabolites and biodegradation of xenobiotics where the large majority of pathways (100% and 88%, respectively) are found in less than 20 species. An example of narrowly-distributed pathway is the biphenyl degradation pathway found in a few Proteobacteria and the Actinobacteria species which are capable of degrading biphenyls - a problematic environmental pollutant (Denef et al. 2004). A detailed analysis of species-specific metabolic pathways was performed for human, as detailed below.

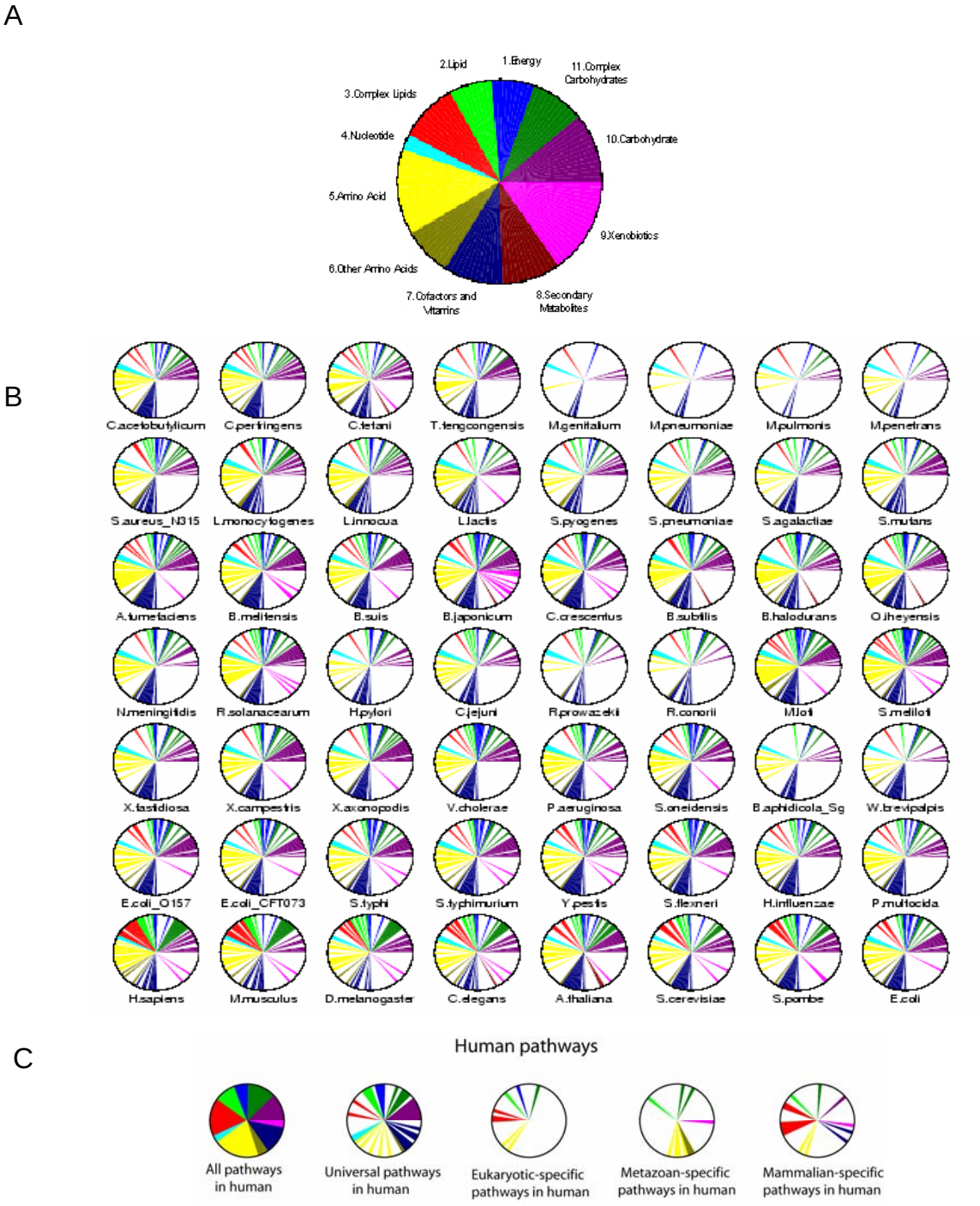

Figure 26 The distribution of metabolic-pathways in species. Each colour represents a super-pathway, each line represents a pathway. (A) All metabolic pathways in KEGG. (B) The distribution of pathways in various prokaryote and eukaryote species. (C) The distribution of metabolic pathways in human, classified according to their estimated time of origin.

6.5 The evolution of the metabolic-network in mammals

6.5.1 The phyletic origin of metabolic-pathways in mammals

65 metabolic pathways are recorded for human. The phyletic origin of these pathways was estimated according to the phyletic age of their reactions. As the pathways are not necessarily a single evolutionary unit (Yamada et al. 2006) different reaction might be classified to more than a single age group. The origin of a pathway was determined according to the age of the most ancient reactions: pathways where at least a single unique reaction is universal are defined as universal pathways; pathways where at least a single unique reaction is eukaryotic-specific and none of the unique reactions are universal were defined as eukaryotic-specific pathways; pathways where at least a single unique reaction is metazoan-specific and none of the unique reactions is of a more ancient origin are defined as metazoan-specific pathways; pathways where all unique reactions are mammalian-specific are defined as mammalian specific pathways. The phyletic classification of pathways is listed in Table 13 and demonstrated in

Figure **26**C. The possible caveats of this classification procedure are discussed in the summary section (6.6).

About half (33 out of 65) of the metabolic pathways in human are universal pathways. The universal pathways include mostly pathways involved in sugar, nucleotide, amino-acid, co-factor and energy metabolism. Unsurprisingly, such pathways were previously defined as a metabolic "skeleton", common to all domains of life (Kyrpides et al. 1999; Peregrin-Alvarez et al. 2003). Not all of the amino acid metabolism pathways are classified as universal and many of them are classified as specific to a more recent lineage, demonstrating the profound differences in amino acid metabolism between mammals and bacteria. Interestingly, only three pathways contain only universal reactions (pentose phosphate pathway, glutathione metabolism and selenoamino acid metabolism), where all other pathways also include reactions from a later lineage. A significant fraction of the eukaryotic, metazoan and human specific reactions (85%, 52% and 45% respectively) are integrated into the skeleton of universal pathways rather than into lineage-specific pathways.

Some of the pathways that are classified as eukaryotic-specific contribute to eukaryotic-specific characteristics (Table 13). Sphingolipids, for example, are a common eukaryotic membrane component (Stryer 1995). The classification of other pathways as eukaryotic-specific is less expected. Pathways like "glycosaminoglycan degradation", "bile acid biosynthesis" and "prostaglandin and leukratine metabolism" are related to characteristic functions of higher animals. Glycosaminoglycans are important components of the extracellular matrix in higher animals where they provide mechanical support to tissues and have an important role in cartilage and other connective tissues. Prostaglandin and leukratine are local hormones that participate in many signalling processes including pain and inflammatory response (Alberts 1994; Stryer 1995). In these pathways, the large majority of enzymes classified to the pathway are from a more recent lineage, where only a few "root" reactions are specific to eukaryotes (Table 13).

Many of the metazoan and human specific pathways are involved in tissue-specific activities. Examples for such activities include neuronal guidance and differentiation (chondrotin/heperan sulphate biosynthesis, ganglioside biosynthesis), hormonal activity in reproduction organs (androgen and estrogen metabolism, C21-steroid hormone metabolism), absorption of ingested lipids in the gastrointestinal track (taurine metabolism), cartilage differentiation (kertan sulphate biosynthesis) and blood cell recognition (blood group glycolipid biosynthesis). Other human specific pathways are involved in inter and intra- cell signalling (retinol metabolism, inositol phosphate metabolism, GPI anchor biosynthesis).

6.5.2 The structure of the metabolic-network in mammals

The KEGG scheme provides information about the links between different pathway maps (e.g., the glycolysis pathway is linked with the pentose phosphate pathway). Firstly, this information was retrieved in order to calculate the avarege connectivity in different groups of pathways. The avarege connectivity of universal, eukaryotic-specific, metazoan-specific, and human-specific pathways is 4.6, 4.5, 2.4, and 2 links per pathway, respectively. Therefore, a tendency exists for more ancient pathways to be more highly connected. Then, we used this information to construct the network of metabolic pathways in human. Each pathway is a node in the network and it is linked to other pathways according to the links in

the KEGG diagram. The constructed network is shown in Figure 27. All the pathways are linked to form a single connected network (Figure 27A).

From the structure of the network one can observe that the core of the networks is mostly formed by universal pathways, where more recent pathways are in many cases added as external nodes. With the exception of two pathways, all universal pathways form a connected component of the network (Figure 27B). The most highly-connected nodes are pathways involved in carbohydrate metabolism: the glycolysis/gluconeogenesis pathway and TCA cycle (14 and 13 edges respectively). Other highly-connected nodes (at least 10 edges) are two amino-acid synthesis pathways (alanine and aspartate metabolism, glycine serine and threonine metabolism). Two of the three pathways where all reactions are universal are only loosely connected to the network: selenoamino acid metabolism pathway is connected to the network via a single edge; the glutathione metabolism pathway is connected via two edges.

When constructing a network only from the lineage-specific pathways, one can observe that unlike the universal pathways, lineage-specific pathways do not link to form a single component-network (Figure 27C). The lineage-specific pathways are clustered into six single-nodes components and five networks where the number of nodes ranges between two and ten. The biggest lineage-specific network contains mostly pathways involved in amino-acid metabolism (eight out of the ten pathways) and all members are directly linked to a universal-pathway. Of special interest are a few examples where lineage-specific pathways are added to the network as external extensions to the core network of universal pathways. Three such examples (marked in Figure 27A) are further discussed as examples to lineage specific adaptations.

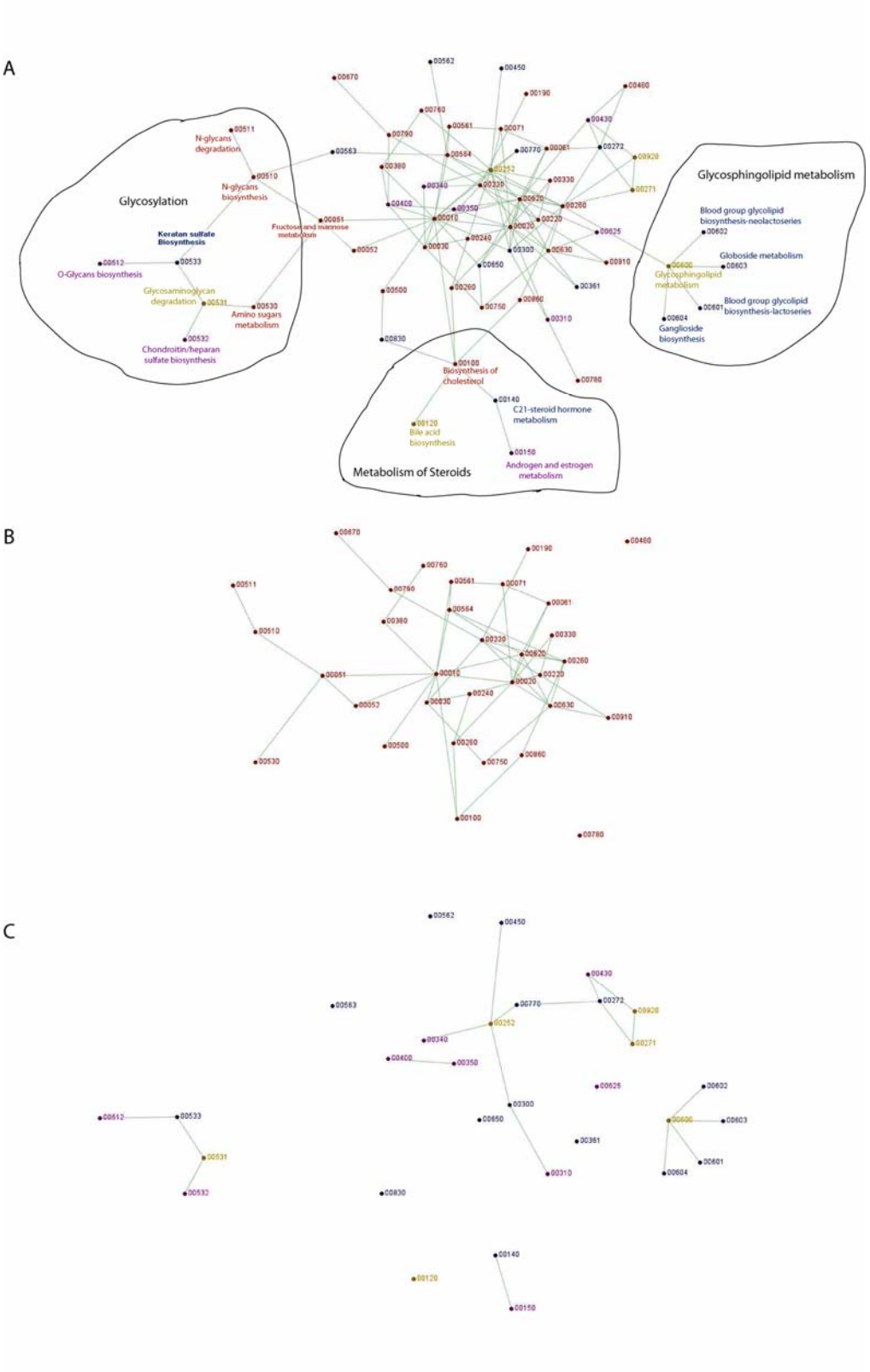

Figure 27 A network representation of the metabolic pathways. (A) The complete network. (B) The universal pathways only. (C) The eukaryotic-, metazoan- and mammalian-specific pathways, without the universal pathways.

6.5.2.1 Glycosylation

The glycosylation of proteins is the most frequent post-translational modification in eukaryotic species (Helenius and Aebi 2004). Glycosylation is the attachment of a carbohydrate residue to a protein either via its serine or threonine residues in O-linked glycosylation or via asparagine residues in N-linked glycosylation (Wildt and Gerngross 2005). N-linked protein glycosylation is the most ubiquitous glycosylation mode. In mammals, this process begins with the transfer of pre-assembled oligosaccharide ($Glc_3Man_9GlcNAc_2$) from the lipid carrier dolichyl-pyrophosphate to asparagine residues of nascent polypeptide chains (Figure 28), in a process that takes place in the endoplasmic reticulum (ER). $Glc_3Man_9GlcNAc_2$ then undergoes trimming of the glucose and of some of the mannose residues, first in the ER and then in the Golgi, followed by the addition of branching N-acetylglucosamine and additional sugars, such as galactose, fucose and sialic-acid to form complex N-glycans (Herscovics 1999).

For a long time it was unclear why cells have evolved such a complicated and apparently wasteful biosynthetic strategy. Why synthesise a large oligosaccharide in the ER, and then – after transferring it to a polypeptide – subject it to trimming in order to build it up again with different sugars? Why should this process be so important as to be conserved in all eukaryotes? The explanation proposed is that different configurations of the N-linked glycans play a role in at least three different stages during the existence of the glycoprotein. For each of the stages the N-linked glycan has to look different. The first phase occurs in the ER where partially trimmed versions of the core oligosaccharide are needed for proper protein folding and quality control. Here the glycans help to secure the fidelity of protein production. The second phase involves a role in intracellular transport and targeting. This phase occurs in the ER, in the Golgi complex, and in trans-Golgi network. The third phase takes place after extensive modification in the Golgi. It occurs when the mature protein has reached the extracellular space, the lysosome, the plasma membrane, or wherever the protein is targeted. The functions of the glycans in the mature proteins are as varied as the structures themselves (Helenius and Aebi 2004). In the immune system, for example, glycosylation plays a critical role in protein interactions and recognition. Specific glycoforms are involved in the folding, quality control and assembly of antigens in the major histocompatibility

complex (MHC) and T cell receptor complex, where the pattern in which oligosaccharides are presented provides a mechanism for distinguishing self from non-self (Rudd et al. 2001).

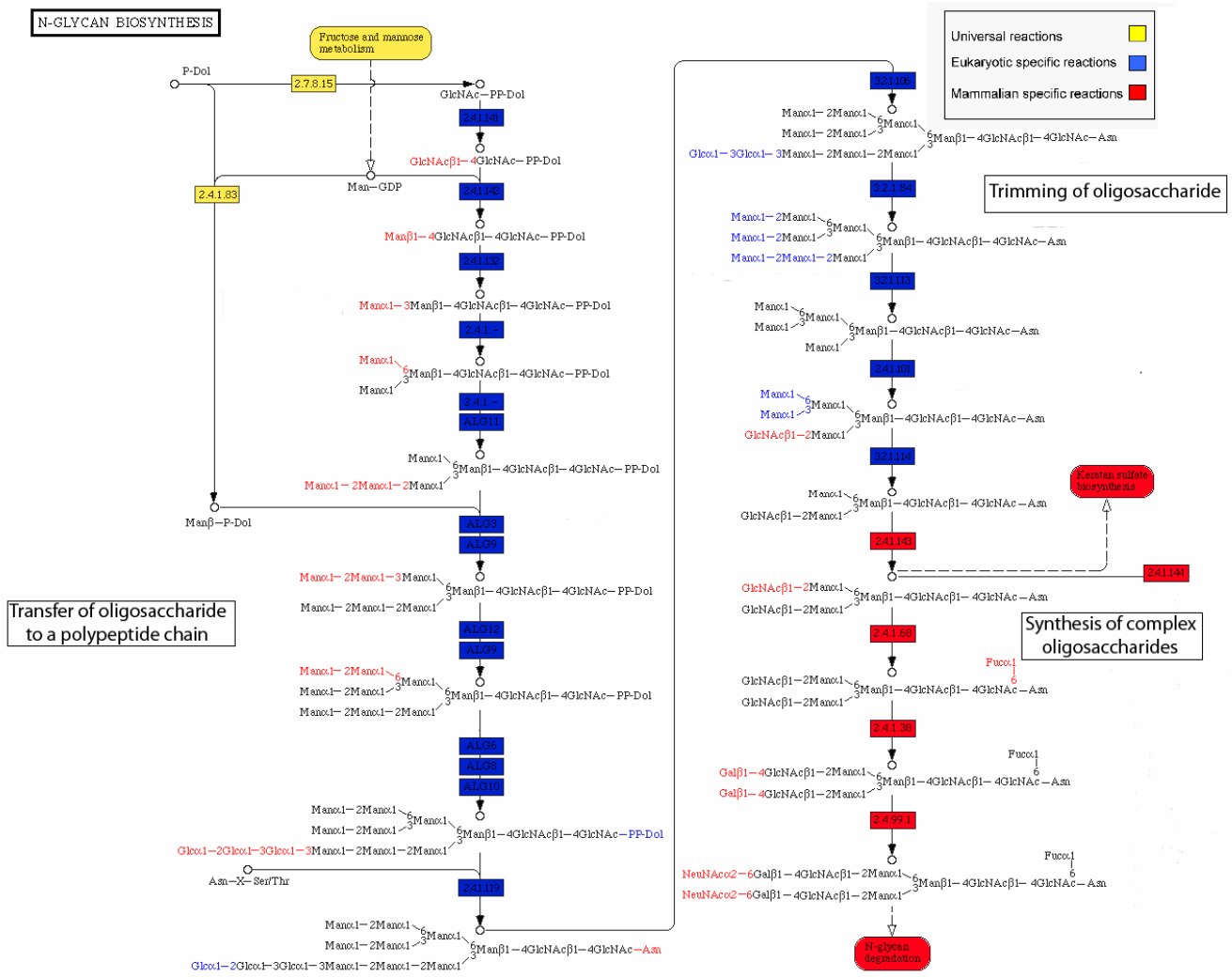

Figure 28 Diagram describing the N-glycan biosynthesis pathway. The diagram was constructed according to information in the KEGG database. Each box represents a reaction and each oval represents a pathwayThe colours represent the estimated phyletic age of the reactions and pathways. The estimated phyletoc age oof pathways are those sshown in Figure 27.

The functions of N-glycans can be further illuminated by considering the evolutionary origin of ER glycosylation. As in most functions of the ER, the synthesis of N-linked glycans stems from homologous processes in the plasma membrane of archaea or bacteria. Proteins with N-linked glycan are present in the outermost layer of the archaeal cell wall and in the cell wall of certain gram negative bacteria. The early stages of glycosylation are homologous between prokaryote and eukaryote where both use similar mechanism to

transfer the oligosaccharide residue. In contrast, the identity of the oligosaccharide residue is not conserved between eukaryotes and prokaryotes. Whereas prokaryote species exhibit a great diversity with regard to the oligosaccharide used, almost all eukaryotes transfer the same structure - $Glc_3Man_9GlcNAc_2$. A notable exception is within the clad of the trypanosomatids, primitive eukaryotes in which a different oligosaccharide is transferred to proteins ($Man_{6-9}GlcNAc_2$). This oligosaccharide, assembled in the cytoplasm, probably corresponds to the glycan transferred to proteins in the common ancestor of all eukaryotic cells, whereas the additional saccharides residues, added in the lumen of the ER, are extensions that have occurred during eukaryotic evolution. The addition of further saccharides to the ancestral oligosaccharide during eukaryotic evolution is probably driven by the internalization of glycoprotein biosynthesis from the plasma membrane to the ER and by the concomitant need to export newly synthesised proteins to the cell surface (Helenius and Aebi 2004). In most eukaryote species the early steps in N-glycan processing in the ER are conserved. Later modifications of the carbohydrate groups in the Golgi exhibit considerable diversity between species and cell types (Herscovics 1999), whereas the switch from use of oligomannose (as in yeast) to complex N-glycans (as in mammals) correlates with the appearance of multicellular organisms (Schachter 2000).

The importance of complex N-glycans to the evolution of multicellular species might be related to their role in the formation of the extracellular matrix. Glycosaminoglycans are essential to the integrity of connective tissues. As polysaccharide chains are too stiff to fold up into compact globular structure (like polypeptide chains) they tend to adopt a highly extended conformation that occupies a huge volume relative to their mass. Therefore, the proteoglycans molecules in connective tissues form a gel-like "ground substance" even at very low concentrations and fill most of the extracellular space. Fibrous proteins such as collagen and elastin are embedded in this proteoglycan gel. The sulfate groups on the saccharide residue make the glycosylaminoglycan highly negatively charged. The high density of negative charges attracts a cloud of cations that are osmotically active causing large amounts of water to be sucked into the matrix. This creates a swelling pressure that enables the matrix to withstand compressive forces (in contrast to collagen fibrils, which resist stretching forces). The cartilage matrix that lines the knee joint, for example, can support pressures of hundreds of atmospheres in this way. Keratan sulfate, chondroitin sulfate, heperan sulfate and dermatan sulfate are among the main groups of glycosaminoglycans in

vertebrates. In invertebrates and plants, other types of polysaccharide often dominate the extracellular matrix, mainly cellulose and chitin, respectively. As can be predicted from their great abundance and structural diversity, the functions of proteoglycans in vertebrates are not limited to shaping the extracellular matrix. Functions of proteoglycans vary greatly and include chemical signalling between cells, binding and regulation of secreted proteins and regulation of the traffic of molecules and cells (Alberts 1994)

The classification approach taken here has assigned the N-glycosylation pathway as universal. This pathway leads to more recent, lineage-specific pathways (Figure 27A), which are involved in the biosynthesis of glycoproteins typical of mammalian tissues. The specific reactions are detailed in Figure 28.

6.5.2.2 Glycosphingolipid (GSL) metabolism

Glycolipids are sugar-containing lipids found in the membrane of cells (Stryer 1995). In eukaryotic cells, glycolipids are derived from sphingosine and form a large and heterogeneous family of GSLs (Sandhoff and Kolter 2003). The first step in the biosynthesis of GSLs, which is common to the large majority of eukaryotic species, is the formation of ceramide from the condensation of L-serine with fatty-acyl coenzyme A (Figure 29). Ceramide consists of a long chain alcohol (D-erythro-sphingiosine) which is N-acylated with fatty acids. The variation in type, number, linkage and further modification of the sugar residues give rise to a combinatorial variety of GSLs, generated by only a few enzymes (Sandhoff and Kolter 2003). Different GSL series are characteristic of different animals, where lactosylceramide is the common precursor for the GSL series found in vertebrates.

Different series of lactosylceramide-derived GSLs have their unique expression patterns in specific cell types (Hakomori 2003). This pattern defines the specific role of the GSLs as antigens, mediators of cell adhesion, and modulators of signal transduction. Blood cells and nerve cells, in particular, have their characteristic composition of GSL. Nerve cells have the highest concentration of ganglosides, the most complex GSL (Stryer 1995), and different nerve cell types express different types of ganglosides (Wiegandt 1995). Ganglosides are essential to the function of the nerve system and disorders in their metabolism have severe clinical consequences such as the Tay-Sachs disease (Stryer 1995).

They are assumed to function as neuro-modulators in connection with calcium, and thus to contribute to the transmission and storage of information (Rahmann 1995).

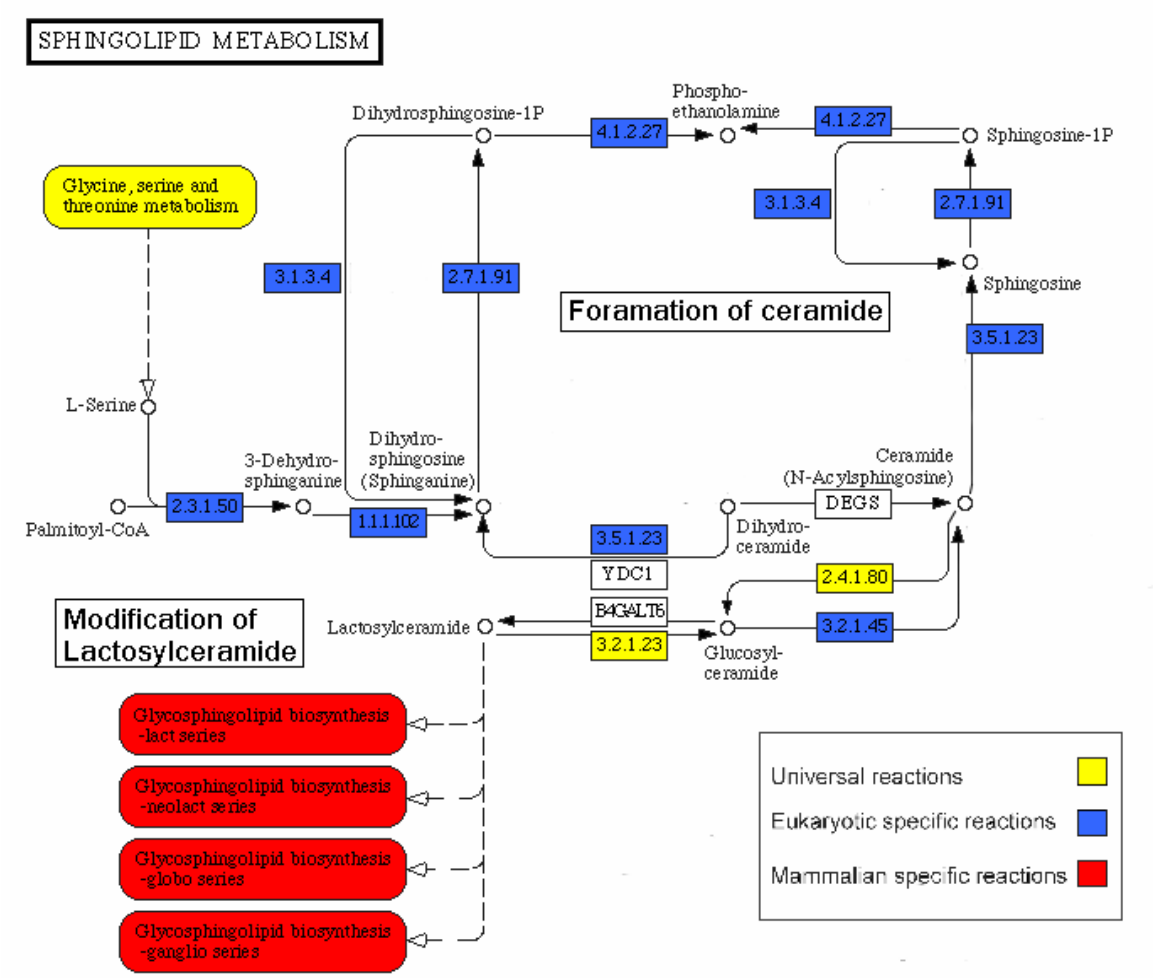

Figure 29 Diagram describing the sphingolipid metabolism pathway. The diagram was constructed according to information in the KEGG database. Each box represents a reaction and each oval represents a pathwayThe colours represent the estimated phyletic age of the reactions and pathways. The estimated phyletoc age oof pathways are those sshown in Figure 27

The structure of the network in Figure 27A demonstrates that the eukaryotic pathway of GSL metabolism is linked to the core network via the universal serine metabolism pathway. The eukaryotic-specific GSL pathway leads to mammalian-specific pathways, which are involved in the metabolism of cell-type specific GSLs – the metabolism of ganglosides which are characteristics of nerve cells, and the metabolism of the lactosylceramide series which are characteristic of groups of blood cells. The specific reactions are detailed in Figure 29.

6.5.2.3 Biosynthesis of cholesterol

Cholesterol is a lipid present in the membrane of eukaryotes but not in most prokaryotes. It is found in varying degrees in virtually all animal membranes. The plasma membrane is usually rich in cholesterol, whereas it is typically much less abundant in the membrane of the organelles of the cell. In animals, cholesterol is also a precursor of many signalling molecules including steroid hormones and bile acids.

Cholesterol is built from four-linked hydrocarbon ring with a hydrocarbon tail linked at one end, and a hydroxyl group attached at the other end. The synthesis of cholesterol is illustrated in Figure 30. The first stage in its synthesis is the formation of isopentenyl pyrophosphate from acetyl-CoA. The second stage is the condensation of six molecules of isopentenyl pyrophosphate (C_5) to form squalene (C_{30}). In the final stage squalene cycles and the tetra-cyclic product is subsequently converted to cholesterol. Cholesterol is then a precursor for other important steroid molecules: the bile salts, steroid hormones, and vitamin D (Stryer 1995).

Bloch, who had studied the evolution of the sterol synthesis pathway, suggested that in prokaryotes the synthesis of sterol had stopped at squalene (Rohmer et al. 1979). In this ancient pathway, which is assumed to have evolved before the appearance of oxygen in the atmosphere (and before the appearance of eukaryote species), squalene is hydrated to form hopanoid – a sterol-like molecule found in the membrane of prokaryote species. Once aerobic conditions had developed, the oxidation of squalene by O_2 gave rise to the formation of genuine precursors of sterols – lanosterol in vertebrates and fungi, and cyclortenol in plants. The degradation of those precursors to form sterols may have paved the way towards the eukaryotic membrane, with its efficient combination of n-acyl chains and sterol (Rohmer et al. 1979).

Despite an intensive research, the role of cholesterol in the plasma membrane remains unclear. As cholesterol is not found in the membrane of prokaryotes, and in the eukaryotic ER and membrane of cellular organelles it can only be found in small amounts, it is clearly not required for the integrity of the membrane. Yet, the absence of cholesterol is lethal to all animal cells. Usually cholesterol is thought to affect the rigidity or fluidity of the

membrane. It is also suggested that cholesterol might inhibit leakage of cations through the plasma membrane and therefore to save energy (ATP) for the cell (Haines 2001).

Derivatives of cholesterol, such as steroid hormones and bile acid, have roles, which are specific to mammals. Bile acids are polar derivates of cholesterol, and because they contain polar and non-polar regions, they are highly effective detergents. Bile salts are synthesized in the liver, stored and concentrated in the gall bladder, and released into the small intestine where they facilitate the absorption of lipids (Stryer 1995).

Cholesterol is also the precursor of the five major classes of steroid hormones: progestagenes, glucocorticoids, mineralocorticoids, androgens, and estrogens. These hormones are powerful signalling molecules that regulate a host of organismal functions. Progesterone, a progestagen, prepares the lining of the uterus for implantation of an ovum, and is also essential for the maintenance of pregnancy. Androgens (such as testosterone) are responsible for the development of the male secondary sex characteristics, whereas estrogens (such as estrone) are required for the development of female sex characteristics. Estrogens, along with progesterone, also participate in the ovarian cycle. Glucocorticoids (such as cortisol) promote gluconeogenesis and the formation of glycogen, enhance the degradation of fat and protein, and inhibit the inflammatory response. They enable animals to response to stress and the absence of glucocorticoids can be fatal. Mineralocorticoids act on the distal tubules of the kidney to increase the reabsorption of Na^+ and the excretion of K^+ and H^+, which leads to an increase in blood volume and blood pressure. The major sites of synthesis of these classes of hormones are the corpus luteum, for progestagens; the ovaries, for estrogens; the testes, for androgens; and the adrenal cortex, for glucocorticoids and mineralocorticoids (Stryer 1995).

The structure of the network in Figure 27A demonstrates that the universal pathway of sterol biosynthesis leads to mammalian-specific pathways. The specific reactions are detailed in Figure 30.

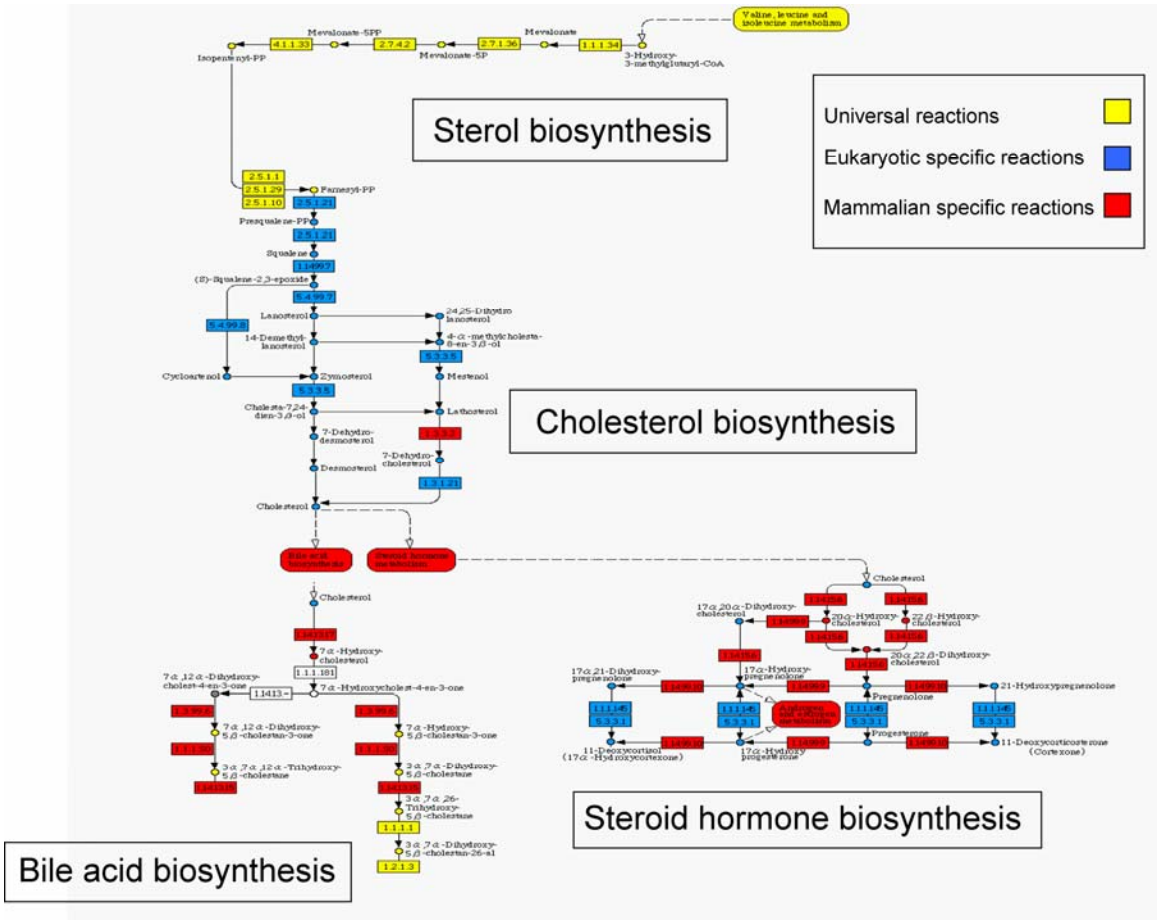

Figure 30 Diagram describing the cholesterol biosynthesis pathway. The diagram was constructed according to information in the KEGG database. Each box represents a reaction. The colours represent the estimated phyletic age of the reaction.

6.6 Summary

This chapter describes a function-based analysis of the evolution of human metabolic properties. The limitations of this analysis must be acknowledged. Firstly, the observations reported here are highly dependent on the scheme of the database used. The classification of reactions into pathways is ultimately subjective and could have been done in more than one way. Therefore, the construction of the pathway repertoire of a species and the signalling of a pathway as present or absent in a species or in a lineage are not definitive. Secondly, the classification of reactions into phyletic groups is limited in its resolution as a result of the small number of fully sequenced multicellular species. Therefore, the estimated phyletic origin of a pathway is obviously influenced by the scheme of the database, the

estimated age of the reactions, and the choice of the criteria for the phyletic assignment of the pathway (a single unique enzyme). These limitations affect the strength of such an analysis for detecting the contribution of each lineage to the development of the metabolic repertoire along species evolution. Yet, despite these limitations, when studying the classifications retrieved from this robust analysis in light of the experimental literature, it seems that the general picture drawn here is usually accurate. Moreover, as will be discussed below, the structural organisation of the metabolic network provides a valuable insight on the gradual evolution of phenotypes which are characteristics of a lineage.

The function-based analysis performed here enables one to use the genomic data in order to identify lineage specific functional-innovations, understand their contribution in the wider context of the metabolic network, and outline the gradual process of species evolution. The main novelty in this function-based anlysis is in providing a global view on the evolution of the mammalin network. Where previous studies have characterized the metabolic "skeleton", common to all domains of life (Kyrpides et al. 1999; Peregrin-Alvarez et al. 2003), a global view of the metabolic innovations in mammals was not previously reported. Therefore, although the phyletic classification of pathways are in most cases in agreement with the biochemical literature, this study provides for the first time a comprehensive analysis of the origin of the complement of the mammalian metabolic pathways. Furthermore, this study not only concerns the characterization of 'anicient' and 'novel' pathways, but also describes for the first time the way they are integrated to form a metabolic network. The preliminary results presented here imply that a universal metabolic core vertically inherited from a pre-eukaryotic ancestor remained highly conserved along the mammalian phylogenesis. Compatible with previous studies which have charactarised a universal metabolic 'skeleton', the universal pathways in mammals are mainly involved in the metabolism of the basic building blocks of every living cell: carbohydrates, lipids, and amino-acids. Only a limited number of reactions and pathways seem to be specific to eukaryotes, metazoan and mammals. The eukaryotic specific pathways are in many cases involved in the metabolism of the complex structures found in the membrane of the eukaryotic cells, or the membrane of the cellular organelles. In multicellular species, these eukaryotic-specific reactions are the first steps in the synthesis of extracellular molecules that participate in inter-cellular communication and cell adhesion. Finally, in reactions specific to mammals, these

extracellular molecules provide precursors for the biosynthesis of more complex molecules, whose function is in many cases characteristic of a specific cell type.

The structure on the network, which is described here for the first time, emphasises the gradual evolution of processes which are specific to a unique cell type, where the development of a phenotype in an ancient lineage provides a platform for the evolution of more recent characteristics. As the phyletic representation of the metabolic network put gene innovations in an evolutionary context, it demonstrates how essential was the development of the eukaryotic membrane, cellular organelles and the system of transportation between them to the onward emergence of multicellular life. Whereas the basic mechanisms of inter-cellular communication are common between invertebrates and vertebrates, many of the mammalian specific pathways are involved in tissue-specific functions. The evolution of these pathways must have coincided with the emergence of cell types and organs. A more extensive collection of metazoan species, including prebilaterian species, can provide a better resolution for studying the morphological context in which new functions (as viewed by the gene repertoire) had appeared. Better understanding of the evolutionary context in which new reactions had evolved can shed light on of the events leading to formation of the complex body organisation in mammals.

Finally, the network approach highlights evolutionary junctions when an ancestor species gained the ability to catalyze a new reaction type or to use a new substrate. Examples for such junctions from the analysis performed here are cholesterol and sphingiosine, molecules which their biosynthesis is lineage specific. The study of such junctions can contribute to our understanding of the co-evolution between enzyme and substrate, and more generally on the mechanism behind the evolutionary process.

Chapter 7: Discussion

7.1 Overview

The evolution of phenotype is a reflection of the selective forces shaping the genome. The construction of the gene repertoire of a species, followed by its comparison to the gene repertoire of other species, either similar or distant, illuminate the way changes in the gene repertoire are related to phenotypic innovations. The interpretation of genomic data becomes harder when the complexity of species grows. Complex species are characterized by a growing variety of differentiation modes, and therefore the presence of a gene in the genome cannot indicate where and when a gene is required. In mammals, the embryonic developmental process, the differentiated body structure, and the complex regulatory system makes the availability of expression data essential for understanding the evolution of the mammalian genome. Combining expression data together with sequence data can provide an evolutionary perspective for understanding development and differentiation since it enables to detect the evolutionary origin of globally and specifically expressed genes.

In recent years expression data from various differentiation modes were accumulated. Several databases contain expression information from tissues of adult mammals. These databases, together with sequence data, can be used in order to characterize a relationship between the evolution of the gene repertoire and the specialization process of tissues in multicellular metazoa. The specialization of tissues requires not only the development of enhanced regulatory mechanism but also the emergence and specialization of genes that are required in order to perform a tissue specific function. The unique combination of genes that are expressed in a tissue – the tissue's transcriptome - determines its unique physiological role. One way to better understand the specialization of the transcriptome in tissues is to view the transition from unicellularity to multicellularity as a transition from a studio flat to a 'room-differentiated' house. Some of the essential functions from the studio flat became specific to one room type (shower) while other functions had mild adaptations in a few rooms (desk, dining table). Possibly, some rooms acquired house-specific functions that cannot be found in the studio flat (conservatory).

In this thesis I have aimed to study how to 'put the house in order' in order to

illuminate the genomic processes accompanying tissue specialization. Therefore, I have related the evolutionary origin of both genes and functions to their expression pattern in mammalian tissues. The following section ties together the main observation from this thesis and points out few possible implications. The novelty of each finding is discussed in the relevant chapter. It is important to note that the formation of spatially differentiated body structure is only one aspect in the evolution of mammals. Other expansions in the variety of differentiation modes had occurred during the evolution of the lineage leading to mammals. For example, the emergence of eukaryotic life was accompanied by a dramatic increase in the variety of temporal differentiation modes (Aravind and Subramanian 1999). The study of types of expression data other than these used here (adult mammalian tissues) is possibly useful for exploring the origin of non-spatial differentiation modes, but is beyond the scope of this thesis.

7.2 Observations, implications and suggestions for future work

7.2.1 'New' characteristics are added on top of an existing ancient core

Understanding the contribution of genomic innovations to the specialization of tissues requires the appreciation of both the differences between tissues, and the differences between each tissue to the last unicellular ancestor of metazoa (which represents a pre-specialized form of all tissues). A first step in comparing the way different tissues relate to the last unicellular ancestor of metazoa is to study the fraction of 'old' and 'young' genes in each tissue. 'Old' genes are those that have descendant from a unicellular ancestor of mammals (pre-metazoan genes), and 'young' genes are innovations from a multicellular lineage (metazoan-specific genes). The analysis of the distribution of the phyletic age of genes expressed in adult mouse tissues indicates that all tissues have a similar fraction of 'old' and 'young' genes (section 5.4). If tissues are similar in their fraction of 'old' and 'young' genes, where do the differences between tissues lie? The demonstrated tendency of 'old' genes to be globally expressed and 'young' genes to be specifically expressed (section 5.5) indicates that, to a large extent, variations between tissues are achieved by specific expression of 'young' genes, whereas 'old' genes are conserved between all tissues. Since none of the tissues is entirely 'old', and none of the tissues is entirely 'young', in the transcriptomes of all

tissues 'young' genes were added to an existing 'ancient' core which is composed of 'old' genes.

How do proteins encoded by the 'young' genes operate together with the proteins encoded by the 'old' genes? The cellular networks of proteins, which are uniquely formed according the transcriptome of each tissue, were not constructed as part of this thesis. Construction of the cellular networks of different tissues is useful for exploring the way the networks have evolved. For example, it is possible to identify these junctions where 'young' proteins are added to the network, and to characterize the molecular functions of the 'old' and 'young' proteins in the junction. The structure of the cellular network is possibly similar to the structure of the metabolic-network (as reported in section 6.5.2) – 'old' functions form a core network and 'young' functions are in many cases added as external additions. Currently, the full construction of the cellular networks is not yet possible, at least for mammalian cells. In the lack of such information it is possible to construct partial segments of the networks, or to compare the connectivity of 'old' and 'young' proteins.

7.2.2 Innovations in the metabolic repertoire were not a main factor in the formation of a complex body structure

A possible way to reveal the cellular role of 'old' and 'young' proteins is to assign them into functional categories. We analyzed not only the expression pattern of proteins from different phyletic groups, but also the expression pattern of proteins from different functional groups. Regulatory proteins were shown to have a tendency to be specifically expressed whereas metabolic proteins tend to be globally expressed (section 5.6). How do adult mammalian tissues refer to the unicellular ancestor they have descended from? The phyletic assignment of proteins from different functional groups shows that most of the regulatory proteins are 'young' proteins and most of the metabolic proteins are 'old' proteins (section 5.6). The scarcity of tissue-specific, 'young' metabolic proteins indicates that innovations in the metabolic repertoire were not a main factor in the formation of a complex body structure (although a disclaimer to this statement will be discussed in the following section).

A support to this view is provided by the analyses of the enzyme functional group that was performed as part of this thesis. Two lines of evidence indicate that the repertoire

of enzymatic reactions in mammals is very similar to the repertoire in the unicellular ancestor. Firstly, the number of metabolic reactions in multicellular species is not significantly higher than the number of reactions in unicellular eukaryotes although the total number of genes in mammals is significantly higher (section 2.5). Secondly, almost 2/3 of the mammalian enzymatic reactions have descendend from a unicellular ancestor (section 6.3).

To a great extent, it is therefore possible to view the mammalian body as a collection of cells of the unicellular ancestor, where each cell type is under different differentiation mode, and the role of the novel proteins is to lead each cell type to its unique differentiation mode, or to communicate between the cell types and hold them together. It is therefore interesting to explore whether a relation exists between temporal differentiation modes in unicellular eukaryotes and spatial differentiation modes in metazoa, where a temporal differentiation mode in a unicellular ancestor precede the emergence of a specific cell type.

7.2.3 The 'conserved ancient metabolic core' is not truly conserved

If the majority of the tissue-specific proteins are regulatory proteins, what are the variations between tissues that these proteins regulate? Variations between tissues in their metabolic repertoire seem essential in order to perform a unique physiological role. Does such variation exist? Several findings indicate that despite (and along with) the conservation in the metabolic repertoire (as discussed above), there are still significant variations between tissues in their metabolic capabilities. Mainly, less than a third of the metabolic proteins are expressed in all tissues whereas the remaining metabolic proteins are differentially expressed (section 5.5).

The comparison of the enzymatic repertoire in mammals to the enzymatic repertoire in a unicellular ancestor illuminates the genomic innovations followed by specialization of expression of the enzymatic proteins. Two major changes exist between these enzymatic sets. Firstly, the mammalian enzymatic repertoire contains additional reactions, where about a third of the mammalian enzymatic reactions are specific to metazoa (section 6.3). Many of the metazoan-specific reactions are involved in tissue-specific activities such as the synthesis of steroid hormones or ganglioside residues (section 6.5). Secondly, the mammalian enzymatic repertoire is significantly more functionally redundant (i.e. more enzyme proteins

are assigned to an EC reaction) in comparison to a unicellular ancestor (section 2.5.4). Whereas in yeast only 37% of the enzymatic reactions are redundant, in mammals 58% of the reactions are redundant. Since most of the functional redundancy is the result of gene duplication events (section 2.6), in many cases the isoenzyme duplicates become specifically expressed where each duplicate is expressed in a different set of tissues (section 3.4). The specialization of expression of the isoenzyme duplicates is in many cases accompanied by the specialization of their functions (section 3.7).

The findings from this thesis therefore indicate that along a conserved metabolic core, variations exist between tissues in their metabolic capabilities. The increased functional redundancy facilitates the emergence of tissue specific adaptations. Novel reactions often participate in tissue-specific pathways, and some examples exist for 'old' pathways that have become specialized to a few cell types (section 4.4).

7.2.4 In a multicellular system, it is possible that a need for coincidence between the specialization of enzymes and the specialization of transporters is a limiting factor in the appearance of metabolic innovations

In some of the metabolic pathways which are specific to a lineage a new compound, which was not used by a preceding ancestor, is utilized. Examples for such compounds are cholesterol and sphingiosine, molecules that have a crucial role in the emergence of eukaryotic and multicellular life forms (section 6.5.2). In a multicellular species, it is possible that the usage of a new compound (or its derivates) involves its export from the cell where it was synthesized, to other tissues. Such export requires that the synthesizing tissue will also express transporters which recognize the novel metabolites. It is therefore possible that a mutual dependence between the specialization of enzymes and the specialization of transporters contributed to the slow rate of appearance of new metabolic reactions in multicellular species.

Although enzymes are the functional group which was analyzed in the greatest detail as part of this thesis, evidence exists that a relative stagnation in the reaction repertoire occurs also in the transporter functional group. The phyletic distribution of transporter proteins indicates that more than 2/3 of them are 'old', pre-metazoan proteins (section 5.5), implying that the repertoire of metabolites remained relatively conserved between mammals

and their unicellular ancestor. A detailed study of the size and composition of the transporters functional group, as well as a study of their expression pattern in mammalian tissues can illuminate the way transports and enzymes have co-evolved in multicellular species.

7.2.5 The chicken and the egg dilemma – how does a group of cells become committed to perform a unique physiological role?

The sequence of events that initiate a process where a group of cells had specialized to perform a unique physiological role remains vague. One possible scenario is that a change in regulation had limited the expression of a protein, or a group of proteins, to a defined spatial space. A mutation in the protein had lead to the emergence of a new role for the set of cells where it is expressed, and encourage positive selection for the specialization of other proteins. When does the emergence of a new function define a unique role of a tissue? Few reported examples link specification of function to the emergence of new tissue types. One such example is the duplication of an ancestral *opsin* gene into two paralogues, *c-opsin* and *r-opsin*. It has been suggested that this duplication event, that took place in an early metazoan ancestor, had allowed the diversification of the cilliary and rhabdomeric photoreceptor cell types from a precursor photoreceptor ancestor cell (Arendt et al. 2004). In mammals, ciliary photoreceptor cells (expressing *c-opsin*) have became the main visual photoreceptor cells (rods and cones), whereas rhabdomeric photoreceptor cells (expressing *r-opsin*) are thought to give rise to cells involved in photoperiodicity regulation.

The number of fully sequenced metazoan species constantly grows. The additional genomes will provide a better resolution for identifying the genomic events that lead to the emergence of new organs. Analysis of such data together with expression data can mark which tissues are conserved between different species or different families, and which tissues evolve faster during speciation. Relating genotypic changes to the phenotype of complex body structure together with relating tissue differentiation to speciation will contribute to our understanding of the way complex species evolve.

Bibliography

Affymetrix
[http://www.affymetrix.com/Auth/support/downloads/manuals/data_analysis_fundamentals_manual.pdf].

http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/I/Invertebrates.html.

Adams, K.L., R. Cronn, R. Percifield, and J.F. Wendel. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc Natl Acad Sci U S A* **100:** 4649-4654.

Agrawal, A., Q.M. Eastman, and D.G. Schatz. 1998. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* **394:** 744-751.

Aguilar, D., F.X. Aviles, E. Querol, and M.J. Sternberg. 2004. Analysis of phenetic trees based on metabolic capabilites across the three domains of life. *J Mol Biol* **340:** 491-512.

Alberts, B. 1994. *Molecular biology of the cell*. Garland Pub., New York.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Andrade, M.A., C. Ouzounis, C. Sander, J. Tamames, and A. Valencia. 1999. Functional classes in the three domains of life. *J Mol Evol* **49:** 551-557.

Andrade, M.A. and C. Sander. 1997. Bioinformatics: from genome data to biological knowledge. *Curr Opin Biotechnol* **8:** 675-683.

Aoki, V.W. and D.T. Carrell. 2003. Human protamines and the developing spermatid: their structure, function, expression and relationship with male infertility. *Asian J Androl* **5:** 315-324.

Apic, G., J. Gough, and S.A. Teichmann. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* **310:** 311-325.

Apweiler, R., A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.S. Yeh. 2004. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32 Database issue:** D115-119.

Aravind, L. and G. Subramanian. 1999. Origin of multicellular eukaryotes - insights from proteome comparisons. *Curr Opin Genet Dev* **9:** 688-694.

Arendt, D., K. Tessmar-Raible, H. Snyman, A.W. Dorresteijn, and J. Wittbrodt. 2004. Ciliary photoreceptors with a vertebrate-type opsin in an invertebrate brain. *Science* **306:** 869-871.

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25-29.

Bairoch, A. and R. Apweiler. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28:** 45-48.

Barrett, A.J. 1994. Classification of peptidases. *Methods Enzymol* **244:** 1-15.

Bateman, A., L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L. Sonnhammer, D.J. Studholme, C. Yeats, and S.R. Eddy. 2004. The Pfam protein families database. *Nucleic Acids Res* **32:** D138-141.

Benson, M., L. Jansson, M. Adner, A. Luts, R. Uddman, and L.O. Cardell. 2005. Gene profiling reveals decreased expression of uteroglobin and other anti-inflammatory genes in nasal fluid cells from patients with intermittent allergic rhinitis. *Clin Exp Allergy* **35:** 473-478.

Birney, E., T.D. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, G. Coates, J. Cuff, V. Curwen, T. Cutts, T. Down, E. Eyras, X.M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H.R. Hotz, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, K.C. Woodwark, G. Cameron, R. Durbin, A. Cox, T. Hubbard, and M. Clamp. 2004. An overview of Ensembl. *Genome Res* **14:** 925-928.

Blanc, G. and K.H. Wolfe. 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* **16:** 1679-1691.

Boeckmann, B., A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31:** 365-370.

Boer, P.H., C.N. Adra, Y.F. Lau, and M.W. McBurney. 1987. The testis-specific phosphoglycerate kinase gene pgk-2 is a recruited retroposon. *Mol Cell Biol* **7:** 3107-3112.

Bolger, G., T. Michaeli, T. Martins, T. St John, B. Steiner, L. Rodgers, M. Riggs, M. Wigler, and K. Ferguson. 1993. A family of human phosphodiesterases homologous to the dunce learning and memory gene product of Drosophila melanogaster are potential targets for antidepressant drugs. *Mol Cell Biol* **13:** 6558-6571.

Bonner, J.T. 2003. Evolution of development in the cellular slime molds. *Evol Dev* **5:** 305-313.

Bonner, J.T. 2003. On the origin of differentiation. *J Biosci* **28:** 523-528.

Camon, E., M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox, and R. Apweiler. 2003. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res* **13:** 662-672.

Cases, I., V. de Lorenzo, and C.A. Ouzounis. 2003. Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol* **11:** 248-253.

Chervitz, S.A., L. Aravind, G. Sherlock, C.A. Ball, E.V. Koonin, S.S. Dwight, M.A. Harris, K. Dolinski, S. Mohr, T. Smith, S. Weng, J.M. Cherry, and D. Botstein. 1998. Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* **282:** 2022-2028.

Clamp, M., D. Andrews, D. Barker, P. Bevan, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, T. Hubbard, A. Kasprzyk, D. Keefe, H. Lehvaslaiho, V. Iyer, C. Melsopp, E. Mongin, R. Pettett, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and E. Birney. 2003. Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res* **31:** 38-42.

Copley, R.R., J. Schultz, C.P. Ponting, and P. Bork. 1999. Protein families in multicellular organisms. *Curr Opin Struct Biol* **9:** 408-415.

Denef, V.J., J. Park, T.V. Tsoi, J.M. Rouillard, H. Zhang, J.A. Wibbenmeyer, W. Verstraete, E. Gulari, S.A. Hashsham, and J.M. Tiedje. 2004. Biphenyl and benzoate metabolism in a genomic context: outlining genome-wide metabolic networks in Burkholderia xenovorans LB400. *Appl Environ Microbiol* **70:** 4961-4970.

Devos, D. and A. Valencia. 2000. Practical limits of function prediction. *Proteins* **41:** 98-107.

Doolittle, R.F., D.F. Feng, S. Tsang, G. Cho, and E. Little. 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* **271:** 470-477.

Duret, L. and D. Mouchiroud. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* **17:** 68-74.

Enright, A.J., V. Kunin, and C.A. Ouzounis. 2003. Protein families and TRIBES in genome sequence space. *Nucleic Acids Res* **31:** 4632-4638.

Force, A., M. Lynch, F.B. Pickett, A. Amores, Y.L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531-1545.

Fortna, A., K. Young, E. MacLaren, K. Marshall, G. Hahn, L. Meltesen, M. Bernton, R. Hink, S. Burgers, T. Hernandenz-Boussard, A. Karimpour-Fard, G. D, L. McGavarn, R. Berry, J. Pollack, and J. Sikela. 2004. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biology* **2:** 937-954.

Freilich, S., T. Massingham, S. Bhattacharyya, H. Ponsting, P.A. Lyons, T.C. Freeman, and J.M. Thornton. 2005. Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. *Genome Biol* **6:** R56.

Freilich, S., T. Massingham, E. Blanc, L. Goldovsky, and J.M. Thornton. 2006. Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. *Genome Biol* **7:** R89.

Freilich, S., R.V. Spriggs, R.A. George, B. Al-Lazikani, M. Swindells, and J.M. Thornton. 2005. The complement of enzymatic sets in different species. *J Mol Biol* **349:** 745-763.

Goddard, I., A. Florin, C. Mauduit, E. Tabone, P. Contard, R. Bars, F. Chuzel, and M. Benahmed. 2003. Alteration of lactate production and transport in the adult rat testis exposed in utero to flutamide. *Mol Cell Endocrinol* **206:** 137-146.

Goldberg, A.L. 1995. Functions of the proteasome: the lysis at the end of the tunnel. *Science* **268:** 522-523.

Goldovsky, L., I. Cases, A.J. Enright, and C.A. Ouzounis. 2005. BioLayout(Java): versatile network visualisation of structural and functional relationships. *Appl Bioinformatics* **4:** 71-74.

Goodstadt, L. and C.P. Ponting. 2006. Phylogenetic Reconstruction of Orthology, Paralogy, and Conserved Synteny for Dog and Human. *PLoS Comput Biol* **2**.

Granadino, B., D. Beltran-Valero de Bernabe, J.M. Fernandez-Canon, M.A. Penalva, and S. Rodriguez de Cordoba. 1997. The human homogentisate 1,2-dioxygenase (HGO) gene. *Genomics* **43:** 115-122.

Gu, Z., D. Nicolae, H.H. Lu, and W.H. Li. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* **18:** 609-613.

Gu, Z., S.A. Rifkin, K.P. White, and W.H. Li. 2004. Duplicate genes increase gene expression diversity within and between species. *Nat Genet* **36:** 577-579.

Haines, T.H. 2001. Do sterols reduce proton and sodium leaks through lipid bilayers? *Prog Lipid Res* **40:** 299-324.
Hakomori, S. 2003. Structure, organization, and function of glycosphingolipids in membrane. *Curr Opin Hematol* **10:** 16-24.
Hastings, K.E. 1996. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. *J Mol Evol* **42:** 631-640.
Hedges, S.B., J.E. Blair, M.L. Venturi, and J.L. Shoe. 2004. A molecular timescale of eukaryote evolution and the rise of complex molecular life. *BMC Evolutionary Biology* **4**.
Hegyi, H. and M. Gerstein. 2001. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res* **11:** 1632-1640.
Helenius, A. and M. Aebi. 2004. Roles of N-linked glycans in the endoplasmic reticulum. *Annu Rev Biochem* **73:** 1019-1049.
Hellman, A. and A. Chess. 2002. Olfactory axons: a remarkable convergence. *Curr Biol* **12:** R849-851.
Hendriksen, P.J., J.W. Hoogerbrugge, W.M. Baarends, P. de Boer, J.T. Vreeburg, E.A. Vos, T. van der Lende, and J.A. Grootegoed. 1997. Testis-specific expression of a functional retroposon encoding glucose-6-phosphate dehydrogenase in the mouse. *Genomics* **41:** 350-359.
Herscovics, A. 1999. Importance of glycosidases in mammalian glycoprotein biosynthesis. *Biochim Biophys Acta* **1473:** 96-107.
Howell, D.C. 1992. Statistical methods for psychology, pp. 132-140. Duxbury Press, Belmont, CA.
Hsiao, L.L., F. Dangond, T. Yoshida, R. Hong, R.V. Jensen, J. Misra, W. Dillon, K.F. Lee, K.E. Clark, P. Haverty, Z. Weng, G.L. Mutter, M.P. Frosch, M.E. Macdonald, E.L. Milford, C.P. Crum, R. Bueno, R.E. Pratt, M. Mahadevappa, J.A. Warrington, G. Stephanopoulos, G. Stephanopoulos, and S.R. Gullans. 2001. A compendium of gene expression in normal human tissues. *Physiol Genomics* **7:** 97-104.
Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. 2002. The Ensembl genome database project. *Nucleic Acids Res* **30:** 38-41.
Hughes, A.L.P. 1999. *Adaptive evolution of genes and genomes.* Oxford University Press, New York.
Hulsen, T., M.A. Huynen, J. de Vlieg, and P.M. Groenen. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* **7:** R31.
Huminiecki, L. and K.H. Wolfe. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res* **14:** 1870-1879.
Huynen, M., B. Snel, W. Lathe, 3rd, and P. Bork. 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* **10:** 1204-1210.
Janssen, P., A.J. Enright, B. Audit, I. Cases, L. Goldovsky, N. Harte, V. Kunin, and C.A. Ouzounis. 2003. COmplete GENome Tracking (COGENT): a flexible data environment for computational genomics. *Bioinformatics* **19:** 1451-1452.

Kanehisa, M. and S. Goto. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27-30.

Kanehisa, M., S. Goto, S. Kawashima, and A. Nakaya. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Res* **30:** 42-46.

Karp, P.D., M. Riley, M. Saier, I.T. Paulsen, S.M. Paley, and A. Pellegrini-Toole. 2000. The EcoCyc and MetaCyc databases. *Nucleic Acids Res* **28:** 56-59.

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

Kerszberg, M. and L. Wolpert. 1998. The Origin of Metazoa and the Egg: a Role for Cell Death. *J Theor Biol* **193:** 535-537.

Keseler, I.M., J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I.T. Paulsen, M. Peralta-Gil, and P.D. Karp. 2005. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res* **33:** D334-337.

Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc Natl Acad Sci U S A* **78:** 454-458.

Kondrashov, F.A., I.B. Rogozin, Y.I. Wolf, and E.V. Koonin. 2002. Selection in the evolution of gene duplications. *Genome Biol* **3:** RESEARCH0008.

Krakauer, D.C. and J.B. Plotkin. 2002. Redundancy, antiredundancy, and the robustness of genomes. *Proc Natl Acad Sci U S A* **99:** 1405-1409.

Kyrpides, N., R. Overbeek, and C. Ouzounis. 1999. Universal protein families and the functional content of the last universal common ancestor. *J Mol Evol* **49:** 413-423.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh R. Funke D. Gage K. Harris A. Heaford J. Howland L. Kann J. Lehoczky R. LeVine P. McEwan K. McKernan J. Meldrim J.P. Mesirov C. Miranda W. Morris J. Naylor C. Raymond M. Rosetti R. Santos A. Sheridan C. Sougnez N. Stange-Thomann N. Stojanovic A. Subramanian D. Wyman J. Rogers J. Sulston R. Ainscough S. Beck D. Bentley J. Burton C. Clee N. Carter A. Coulson R. Deadman P. Deloukas A. Dunham I. Dunham R. Durbin L. French D. Grafham S. Gregory T. Hubbard S. Humphray A. Hunt M. Jones C. Lloyd A. McMurray L. Matthews S. Mercer S. Milne J.C. Mullikin A. Mungall R. Plumb M. Ross R. Shownkeen S. Sims R.H. Waterston R.K. Wilson L.W. Hillier J.D. McPherson M.A. Marra E.R. Mardis L.A. Fulton A.T. Chinwalla K.H. Pepin W.R. Gish S.L. Chissoe M.C. Wendl K.D. Delehaunty T.L. Miner A. Delehaunty J.B. Kramer L.L. Cook R.S. Fulton D.L. Johnson P.J. Minx S.W. Clifton T. Hawkins E. Branscomb P. Predki P. Richardson S. Wenning T. Slezak N. Doggett J.F. Cheng A. Olsen S. Lucas C. Elkin E. Uberbacher M. Frazier R.A. Gibbs D.M. Muzny S.E. Scherer J.B. Bouck E.J. Sodergren K.C. Worley C.M. Rives J.H. Gorrell M.L. Metzker S.L. Naylor R.S. Kucherlapati D.L. Nelson G.M. Weinstock Y. Sakaki A. Fujiyama M. Hattori T. Yada A. Toyoda T. Itoh C. Kawagoe H. Watanabe Y. Totoki T. Taylor J. Weissenbach R. Heilig W. Saurin F. Artiguenave P. Brottier T. Bruls E. Pelletier C. Robert P. Wincker D.R. Smith L. Doucette-Stamm M. Rubenfield K. Weinstock H.M. Lee J. Dubois A. Rosenthal M. Platzer G. Nyakatura S. Taudien A. Rump H. Yang J. Yu J. Wang G. Huang J. Gu L. Hood L. Rowen A. Madan S. Qin R.W. Davis N.A. Federspiel A.P. Abola M.J. Proctor R.M. Myers J. Schmutz M. Dickson J. Grimwood D.R. Cox M.V. Olson R. Kaul C. Raymond N. Shimizu K. Kawasaki S. Minoshima G.A. Evans M. Athanasiou R. Schultz B.A. Roe F. Chen H. Pan J. Ramser H. Lehrach R. Reinhardt W.R. McCombie M. de la Bastide N. Dedhia H.

Blocker K. Hornischer G. Nordsiek R. Agarwala L. Aravind J.A. Bailey A. Bateman S. Batzoglou E. Birney P. Bork D.G. Brown C.B. Burge L. Cerutti H.C. Chen D. Church M. Clamp R.R. Copley T. Doerks S.R. Eddy E.E. Eichler T.S. Furey J. Galagan J.G. Gilbert C. Harmon Y. Hayashizaki D. Haussler H. Hermjakob K. Hokamp W. Jang L.S. Johnson T.A. Jones S. Kasif A. Kaspryzk S. Kennedy W.J. Kent P. Kitts E.V. Koonin I. Korf D. Kulp D. Lancet T.M. Lowe A. McLysaght T. Mikkelsen J.V. Moran N. Mulder V.J. Pollara C.P. Ponting G. Schuler J. Schultz G. Slater A.F. Smit E. Stupka J. Szustakowski D. Thierry-Mieg J. Thierry-Mieg L. Wagner J. Wallis R. Wheeler A. Williams Y.I. Wolf K.H. Wolfe S.P. Yang R.F. Yeh F. Collins M.S. Guyer J. Peterson A. Felsenfeld K.A. Wetterstrand A. Patrinos M.J. Morgan P. de Jong J.J. Catanese K. Osoegawa H. Shizuya S. Choi and Y.J. Chen. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lang, B.F., C. O'kelly, T. Nerad, M.W. Gray, and G. Burger. 2002. The closest unicellular relatives of animals. *Curr Biol* **12:** 1773-1778.

Lehner, B. and A.G. Fraser. 2004. Protein domains enriched in mammalian tissue-specific or widely expressed genes. *Trends Genet* **20:** 468-472.

Lipshutz, R.J., D. Morris, M. Chee, E. Hubbell, M.J. Kozal, N. Shah, N. Shen, R. Yang, and S.P. Fodor. 1995. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* **19:** 442-447.

Liu, W.M., R. Mei, X. Di, T.B. Ryder, E. Hubbell, S. Dee, T.A. Webster, C.A. Harrington, M.H. Ho, J. Baid, and S.P. Smeekens. 2002. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* **18:** 1593-1599.

Lux, R., K. Jahreis, K. Bettenbrock, J.S. Parkinson, and J.W. Lengeler. 1995. Coupling the phosphotransferase system and the methyl-accepting chemotaxis protein-dependent chemotaxis signaling pathways of Escherichia coli. *Proc Natl Acad Sci U S A* **92:** 11583-11587.

Lynch, M. and J.S. Conery. 2003. The origins of genome complexity. *Science* **302:** 1401-1404.

Lynch, M. and A. Force. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154:** 459-473.

Marcotte, E.M. 2000. Computational genetics: finding protein function by nonhomology methods. *Curr Opin Struct Biol* **10:** 359-365.

Marcotte, E.M., M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285:** 751-753.

Martindale, M.Q., J.R. Finnerty, and J.Q. Henry. 2002. The Radiata and the evolutionary origins of the bilaterian body plan. *Mol Phylogenet Evol* **24:** 358-365.

Miki, R., K. Kadota, H. Bono, Y. Mizuno, Y. Tomaru, P. Carninci, M. Itoh, K. Shibata, J. Kawai, H. Konno, S. Watanabe, K. Sato, Y. Tokusumi, N. Kikuchi, Y. Ishii, Y. Hamaguchi, I. Nishizuka, H. Goto, H. Nitanda, S. Satomi, A. Yoshiki, M. Kusakabe, J.L. DeRisi, M.B. Eisen, V.R. Iyer, P.O. Brown, M. Muramatsu, H. Shimada, Y. Okazaki, and Y. Hayashizaki. 2001. Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc Natl Acad Sci U S A* **98:** 2199-2204.

Moran, N.A. 2002. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108:** 583-586.

Moran, N.A. 2003. Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr Opin Microbiol* **6:** 512-518.

Mulder, N.J., R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti, R. Copley, E. Courcelle, U. Das, R. Durbin, W. Fleischmann, J. Gough, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, J. McDowall, A. Mitchell, A.N. Nikolskaya, S. Orchard, M. Pagni, C.P. Ponting, E. Quevillon, J. Selengut, C.J. Sigrist, V. Silventoinen, D.J. Studholme, R. Vaughan, and C.H. Wu. 2005. InterPro, progress and status in 2005. *Nucleic Acids Res* **33:** D201-205.

Muller, A., R.M. MacCallum, and M.J. Sternberg. 2002. Structural characterization of the human proteome. *Genome Res* **12:** 1625-1641.

Muller, W.E. 2001. Review: How was metazoan threshold crossed? The hypothetical Urmetazoa. *Comp Biochem Physiol A Mol Integr Physiol* **129:** 433-460.

Nelson, D.R. 1999. Cytochrome P450 and the individuality of species. *Arch Biochem Biophys* **369:** 1-10.

O'Brien, K.P., M. Remm, and E.L. Sonnhammer. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33:** D476-480.

O'Halloran, D.M., D.A. Fitzpatrick, G.P. McCormack, J.O. McInerney, and A.M. Burnell. 2006. The molecular phylogeny of a nematode-specific clade of heterotrimeric G-protein alpha-subunit genes. *J Mol Evol* **63:** 87-94.

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin; New York.

Overbeek, R., N. Larsen, G.D. Pusch, M. D'Souza, E. Selkov, Jr., N. Kyrpides, M. Fonstein, N. Maltsev, and E. Selkov. 2000. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res* **28:** 123-125.

Peregrin-Alvarez, J.M., S. Tsoka, and C.A. Ouzounis. 2003. The phylogenetic extent of metabolic enzymes and pathways. *Genome Res* **13:** 422-427.

Pfeiffer, T. and S. Bonhoeffer. 2003. An evolutionary scenario for the transition to undifferentiated multicellularity. *Proc Natl Acad Sci U S A* **100:** 1095-1098.

Prince, V.E. and F.B. Pickett. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet* **3:** 827-837.

Raff, R.A. 1996. *The Shape of Life*. The University of Chicago Press.

Rahmann, H. 1995. Brain gangliosides and memory formation. *Behav Brain Res* **66:** 105-116.

Rajagopalan, D. 2003. A comparison of statistical methods for analysis of high density oligonucleotide array data. *Bioinformatics* **19:** 1469-1476.

Ranea, J.A., D.W. Buchan, J.M. Thornton, and C.A. Orengo. 2004. Evolution of protein superfamilies and bacterial genome size. *J Mol Biol* **336:** 871-887.

Ranson, H., C. Claudianos, F. Ortelli, C. Abgrall, J. Hemingway, M.V. Sharakhova, M.F. Unger, F.H. Collins, and R. Feyereisen. 2002. Evolution of supergene families associated with insecticide resistance. *Science* **298:** 179-181.

Remm, M., C.E. Storm, and E.L. Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314:** 1041-1052.

Rohmer, M., P. Bouvier, and G. Ourisson. 1979. Molecular evolution of biomembranes: structural equivalents and phylogenetic precursors of sterols. *Proc Natl Acad Sci U S A* **76:** 847-851.

Rubin, G.M., M.D. Yandell, J.R. Wortman, G.L. Gabor Miklos, C.R. Nelson, I.K. Hariharan, M.E. Fortini, P.W. Li, R. Apweiler, W. Fleischmann, J.M. Cherry, S. Henikoff, M.P. Skupski, S. Misra, M. Ashburner, E. Birney, M.S. Boguski, T. Brody, P. Brokstein, S.E. Celniker, S.A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R.F. Galle, W.M.

Gelbart, R.A. George, L.S. Goldstein, F. Gong, P. Guan, N.L. Harris, B.A. Hay, R.A. Hoskins, J. Li, Z. Li, R.O. Hynes, S.J. Jones, P.M. Kuehl, B. Lemaitre, J.T. Littleton, D.K. Morrison, C. Mungall, P.H. O'Farrell, O.K. Pickeral, C. Shue, L.B. Vosshall, J. Zhang, Q. Zhao, X.H. Zheng, and S. Lewis. 2000. Comparative genomics of the eukaryotes. *Science* **287:** 2204-2215.

Rudd, P.M., T. Elliott, P. Cresswell, I.A. Wilson, and R.A. Dwek. 2001. Glycosylation and the immune system. *Science* **291:** 2370-2376.

Sandhoff, K. and T. Kolter. 2003. Biosynthesis and degradation of mammalian glycosphingolipids. *Philos Trans R Soc Lond B Biol Sci* **358:** 847-861.

Schachter, H. 2000. The joys of HexNAc. The synthesis and function of N- and O-glycan branches. *Glycoconj J* **17:** 465-483.

Shimeld, S.M. and P.W. Holland. 2000. Vertebrate innovations. *Proc Natl Acad Sci U S A* **97:** 4449-4452.

Soding, J. and A.N. Lupas. 2003. More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* **25:** 837-846.

Steenkamp, E.T., J. Wright, and S.L. Baldauf. 2006. The protistan origin of animals and fungi. *Mol Biol Evol* **23:** 93-106.

Stryer, L. 1995. *Biochemistry*. Freeman, New York, N.Y.

Su, A.I., M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, A. Patapoutian, G.M. Hampton, P.G. Schultz, and J.B. Hogenesch. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99:** 4465-4470.

Su, A.I., T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, and J.B. Hogenesch. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101:** 6062-6067.

Subramanian, S. and S. Kumar. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* **168:** 373-381.

Swanson, W.J. and V.D. Vacquier. 2002. The rapid evolution of reproductive proteins. *Nat Rev Genet* **3:** 137-144.

Swindells, M., M. Rae, M. Pearce, S. Moodie, R. Miller, and P. Leach. 2002. Application of high-throughput computing in bioinformatics. *Philos Transact Ser A Math Phys Eng Sci* **360:** 1179-1189.

Tamames, J., C. Ouzounis, C. Sander, and A. Valencia. 1996. Genomes with distinct function composition. *FEBS Lett* **389:** 96-101.

Tatusov, R.L., N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, and D.A. Natale. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4:** 41.

Tatusov, R.L., E.V. Koonin, and D.J. Lipman. 1997. A genomic perspective on protein families. *Science* **278:** 631-637.

Taylor, R.G., H.L. Levy, and R.R. McInnes. 1991. Histidase and histidinemia. Clinical and molecular considerations. *Mol Biol Med* **8:** 101-116.

Tipton, K. and S. Boyce. 2000. History of the enzyme nomenclature system. *Bioinformatics* **16:** 34-40.

Todd, A.E., C.A. Orengo, and J.M. Thornton. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* **307:** 1113-1143.

Valencia, A. 2005. Automatic annotation of protein function. *Curr Opin Struct Biol* **15:** 267-274.

van Nimwegen, E. 2003. Scaling laws in the functional content of genomes. *Trends Genet* **19:** 479-484.

van Noort, V., B. Snel, and M.A. Huynen. 2003. Predicting gene function by conserved co-expression. *Trends Genet* **19:** 238-242.

von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* **31:** 258-261.

Warrington, J.A., A. Nair, M. Mahadevappa, and M. Tsyganskaya. 2000. Comparioson of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics* **2:** 143-147.

Waterston, R.H. K. Lindblad-Toh E. Birney J. Rogers J.F. Abril P. Agarwal R. Agarwala R. Ainscough M. Alexandersson P. An S.E. Antonarakis J. Attwood R. Baertsch J. Bailey K. Barlow S. Beck E. Berry B. Birren T. Bloom P. Bork M. Botcherby N. Bray M.R. Brent D.G. Brown S.D. Brown C. Bult J. Burton J. Butler R.D. Campbell P. Carninci S. Cawley F. Chiaromonte A.T. Chinwalla D.M. Church M. Clamp C. Clee F.S. Collins L.L. Cook R.R. Copley A. Coulson O. Couronne J. Cuff V. Curwen T. Cutts M. Daly R. David J. Davies K.D. Delehaunty J. Deri E.T. Dermitzakis C. Dewey N.J. Dickens M. Diekhans S. Dodge I. Dubchak D.M. Dunn S.R. Eddy L. Elnitski R.D. Emes P. Eswara E. Eyras A. Felsenfeld G.A. Fewell P. Flicek K. Foley W.N. Frankel L.A. Fulton R.S. Fulton T.S. Furey D. Gage R.A. Gibbs G. Glusman S. Gnerre N. Goldman L. Goodstadt D. Grafham T.A. Graves E.D. Green S. Gregory R. Guigo M. Guyer R.C. Hardison D. Haussler Y. Hayashizaki L.W. Hillier A. Hinrichs W. Hlavina T. Holzer F. Hsu A. Hua T. Hubbard A. Hunt I. Jackson D.B. Jaffe L.S. Johnson M. Jones T.A. Jones A. Joy M. Kamal E.K. Karlsson D. Karolchik A. Kasprzyk J. Kawai E. Keibler C. Kells W.J. Kent A. Kirby D.L. Kolbe I. Korf R.S. Kucherlapati E.J. Kulbokas D. Kulp T. Landers J.P. Leger S. Leonard I. Letunic R. Levine J. Li M. Li C. Lloyd S. Lucas B. Ma D.R. Maglott E.R. Mardis L. Matthews E. Mauceli J.H. Mayer M. McCarthy W.R. McCombie S. McLaren K. McLay J.D. McPherson J. Meldrim B. Meredith J.P. Mesirov W. Miller T.L. Miner E. Mongin K.T. Montgomery M. Morgan R. Mott J.C. Mullikin D.M. Muzny W.E. Nash J.O. Nelson M.N. Nhan R. Nicol Z. Ning C. Nusbaum M.J. O'Connor Y. Okazaki K. Oliver E. Overton-Larty L. Pachter G. Parra K.H. Pepin J. Peterson P. Pevzner R. Plumb C.S. Pohl A. Poliakov T.C. Ponce C.P. Ponting S. Potter M. Quail A. Reymond B.A. Roe K.M. Roskin E.M. Rubin A.G. Rust R. Santos V. Sapojnikov B. Schultz J. Schultz M.S. Schwartz S. Schwartz C. Scott S. Seaman S. Searle T. Sharpe A. Sheridan R. Shownkeen S. Sims J.B. Singer G. Slater A. Smit D.R. Smith B. Spencer A. Stabenau N. Stange-Thomann C. Sugnet M. Suyama G. Tesler J. Thompson D. Torrents E. Trevaskis J. Tromp C. Ucla A. Ureta-Vidal J.P. Vinson A.C. Von Niederhausern C.M. Wade M. Wall R.J. Weber R.B. Weiss M.C. Wendl A.P. West K. Wetterstrand R. Wheeler S. Whelan J. Wierzbowski D. Willey S. Williams R.K. Wilson E. Winter K.C. Worley D. Wyman S. Yang S.P. Yang E.M. Zdobnov M.C. Zody and E.S. Lander. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Watson, S.J. and H. Akil. 1999. Gene chips and arrays revealed: a primer on their power and their uses. *Biol Psychiatry* **45:** 533-543.

Wiegandt, H. 1995. The chemical constitution of gangliosides of the vertebrate nervous system. *Behav Brain Res* **66:** 85-97.

Wildt, S. and T.U. Gerngross. 2005. The humanization of N-glycosylation pathways in yeast. *Nat Rev Microbiol* **3:** 119-128.

Wilson, C.A., J. Kreychman, and M. Gerstein. 2000. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* **297:** 233-249.

Winter, E.E., L. Goodstadt, and C.P. Ponting. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res* **14:** 54-61.

Wood, V. R. Gwilliam M.A. Rajandream M. Lyne R. Lyne A. Stewart J. Sgouros N. Peat J. Hayles S. Baker D. Basham S. Bowman K. Brooks D. Brown S. Brown T. Chillingworth C. Churcher M. Collins R. Connor A. Cronin P. Davis T. Feltwell A. Fraser S. Gentles A. Goble N. Hamlin D. Harris J. Hidalgo G. Hodgson S. Holroyd T. Hornsby S. Howarth E.J. Huckle S. Hunt K. Jagels K. James L. Jones M. Jones S. Leather S. McDonald J. McLean P. Mooney S. Moule K. Mungall L. Murphy D. Niblett C. Odell K. Oliver S. O'Neil D. Pearson M.A. Quail E. Rabbinowitsch K. Rutherford S. Rutter D. Saunders K. Seeger S. Sharp J. Skelton M. Simmonds R. Squares S. Squares K. Stevens K. Taylor R.G. Taylor A. Tivey S. Walsh T. Warren S. Whitehead J. Woodward G. Volckaert R. Aert J. Robben B. Grymonprez I. Weltjens E. Vanstreels M. Rieger M. Schafer S. Muller-Auer C. Gabel M. Fuchs A. Dusterhoft C. Fritzc E. Holzer D. Moestl H. Hilbert K. Borzym I. Langer A. Beck H. Lehrach R. Reinhardt T.M. Pohl P. Eger W. Zimmermann H. Wedler R. Wambutt B. Purnelle A. Goffeau E. Cadieu S. Dreano S. Gloux V. Lelaure S. Mottier F. Galibert S.J. Aves Z. Xiang C. Hunt K. Moore S.M. Hurst M. Lucas M. Rochet C. Gaillardin V.A. Tallada A. Garzon G. Thode R.R. Daga L. Cruzado J. Jimenez M. Sanchez F. del Rey J. Benito A. Dominguez J.L. Revuelta S. Moreno J. Armstrong S.L. Forsburg L. Cerutti T. Lowe W.R. McCombie I. Paulsen J. Potashkin G.V. Shpakovski D. Ussery B.G. Barrell and P. Nurse. 2002. The genome sequence of Schizosaccharomyces pombe. *Nature* **415:** 871-880.

Yamada, T., M. Kanehisa, and S. Goto. 2006. Extraction of phylogenetic network modules from the metabolic network. *BMC Bioinformatics* **7:** 130.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13:** 555-556.

Yang, Z. and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17:** 32-43.

Young, J.M. and B.J. Trask. 2002. The sense of smell: genomics of vertebrate odorant receptors. *Hum Mol Genet* **11:** 1153-1160.

Zhang, L. and W.H. Li. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol* **21:** 236-239.

Appendix A: Papers published

Freilich, S., Blanc, E., Massingham, T., Goldovsky, L. & Thornton, J. M (2006). Relating tissue specialisation to the differentiation of expression of singleton and duplicate mouse genes. *Genome Biol* **7**, R89.

Freilich, S., Massingham, T., Bhattacharyya, S., Ponsting, H., Lyons, P. A., Freeman, T. C. & Thornton, J. M. (2005). Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. *Genome Biol* **6**, R56.

Freilich, S., Spriggs, R. V., George, R. A., Al-Lazikani, B., Swindells, M. & Thornton, J. M. (2005). The complement of enzymatic sets in different species. *J Mol Biol* **349**, 745-63.

Appendix B: Estimating the reliability of annotations describing the participation of a protein in a complex, as retrieved from the KEGG database

This analysis was performed in support of the analysis described in section 2.6.1.

Using KEGG annotations for the identification of hetero-oligomeric complexes.

In order to verify that we do not miss out hetero-oligomeric complexes when text-searching the annotations retrieved from KEGG, we also looked for such complexes in EcoCyc. The coverage in EcoCyc has expanded from its original focus on metabolic pathways and it now includes annotations of all protein functions and is not limited to proteins involved in 'small molecule metabolism' (Keseler et al. 2005).

We looked for all protein complexes that are assigned any EC number and that are hetero-oligomeric in EcoCyc (version 8.5). 46 such EC numbers were identified. As the KEGG-based analysis is limited to enzymes with four digit EC numbers, the EcoCyc set was filtered and was finally left with 36 EC numbers. In KEGG, 39 hetero-oligomeric EC reactions were identified. 25 of these reactions are annotated as hetero-oligomeric complexes in both KEGG and EcoCyc. In most cases (7/11) where an EC reaction is identified only in EocCyc, the KEGG database does annotate the reaction as hetero-oligomeric but does not assign genes to the reaction (for E.coli). In most cases (11/14) where an EC reaction is identified as a complex by KEGG only, the EcoCyc database does not contain the EC number. When using HumanCyc (version 9.0) 14 hetero-oligomeric complexes are identified, compared with 21 such complexes in KEGG. For B.subtilis we find 26 complexes in MetaCyc (version 9.0) compared to 30 in KEGG. The coverage of the two data sources was compared and found to be similar: 626 four-digit EC reactions are found in EcoCyc compared with 656 E.coli reactions in KEGG; 860 four-digit EC reactions are found in HumanCyc compared with 816 human reactions in KEGG.

The number of complexes is similar between the two data sources used. Therefore, the estimated number of complexes retrieved from KEGG provides a reliable reflection of our current state of knowledge.

Homologous hetero-oligomeric complexes.

In section 2.6.1 it is reported that, for the nine model organisms, between 42% and 76% of the multi-protein reactions have all of their proteins homologous to each other (i.e., all classified to a single orthologue cluster). Yet, the analysis fails to look for cases where those proteins work in concert as an homologous-hetero-oligomeric complex. When using the text search used to identify protein complexes in multi-orthologue-cluster reactions (chapter 2.6.1), no such cases were found in KEGG for the nine model species.

Next, we wanted to verify that KEGG's assignments of proteins into orthologue clusters are sensitive enough to enable us to distinguish between proteins that are known to have different functions (i.e., distinct subunits in a protein complex), although homologous. We looked for reactions where only EcoCyc (and not KEGG) describes hetero-oligomeric complexes and find only three such cases (EC 1.6.6.9, 1.6.81, 2.3.1.16). In all cases, the gene assignments for the EC reaction differ between EcoCyc and KEGG and that might explain the lack of a complex annotation in KEGG. There is clearly a discrepancy between EcoCyc and KEGG. However, these results indicate that, at least for prokaryote species, we do not miss hetero-oligomeric complexes where the subunits are homologous and that the classification of proteins to orthologous clusters, as extracted from KEGG, is sensitive enough to enable us to distinguish between proteins that are known to have different functions (although homologous).

In order to examine whether KEGG is missing out references for hetero-oligomeric complexes of homologous proteins in higher species, we looked in detail at those nine reactions where a hetero-oligomeric complex is reported by HumanCyc and not by KEGG (for human). In four of these cases (6.3.2.2, 3.4.24.18, 3.2.1.52, 1.1.1.41) the proteins classified to these reactions are all homologous to each other (classified to the same orthologous cluster in KEGG), though, unlike EcoCyc, KEGG does not provide any indication that the proteins are working in concert. Yet, although a more accurate classification would have classified the proteins in those reactions to more than a single orthologue-cluster, these cases are still rare. The number of reactions with a single orthologue cluster should be reduced to 202 (73%) instead of 206 (74%). Therefore, it seems that for higher species there are cases where KEGG's annotations are not sensitive enough to enable us to distinguish between proteins that are known to have distinct functions in a

protein complex, although homologous. Nevertheless, these cases are relatively rare and in most cases the use of KEGG does provide a reliable reflection of our current state of knowledge, for higher species as well.

Finally, we looked at the specific example of the proteosome complex. The eukaryotic proteosome consists of seven α and seven β subunits that are distinct, but related proteins (Goldberg 1995). In KEGG there are 17 human proteins that are assigned to EC 3.4.25.1 (proteosome, 26S protease). Each protein is assigned to a unique orthologue-cluster, seven representing distinct proteosome α subunits and ten representing distinct β subunits. The assignment of proteins into unique orthologue-clusters provides additional reassurance as to the overall accuracy of the KEGG classification and annotation procedure.

References:

Goldberg, A.L. 1995. Functions of the proteasome: the lysis at the end of the tunnel. *Science* **268:** 522-523.

Keseler, I.M., J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I.T. Paulsen, M. Peralta-Gil, and P.D. Karp. 2005. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res* **33:** D334-337.

Appendix C: Table from section 4.1

Table Appendix c Annotations of specifically expressed singleton proteins (expressed in at most 3 tissue-clusters, as described in section 3.3.2). Possibly, under different conditions, in a different set of tissues (for example only tissues found in male mice were analysed here) or during embryonic development, these proteins will be expressed in additional tissues.

	ENSEMBL id (phyletic group*)	SWISSPROT†/ MGI‡ accession and name	GO Molecular Function	GO Biological Process	GO Cellular Component
antrum	ENSMUSP00000023510 (PM)	PYR5_MOUSE† Uridine 5'-monophosphate synthase	nucleobase, nucleoside, nucleotide and nucleic acid metabolism,'de novo' pyrimidine base biosynthesis,aromatic compound metabolism,biosynthesis,nucleobase metabolism,heterocycle metabolism,	transferase activity,lyase activity,	
antrum, bladder, gall bladder, spleen, thymus	ENSMUSP00000047720 (PM)	Mett11d1‡ methyltransferase 11 domain containing 1			
appendix	ENSMUSP00000043369 (MS)	IL5_MOUSE† Interleukin-5 [Precursor]	phosphorus metabolism,phosphate metabolism,immune response,response to biotic stimulus,positive regulation of metabolism,peptidyl-tyrosine phosphorylation,peptidyl-tyrosine modification,regulation of phosphate metabolism,regulation of metabolism,macromolecule metabolism,positive regulation of phosphate metabolism,positive regulation of peptidyl-tyrosine phosphorylation,	receptor binding,	
bladder, cecum, ileum, kidney	ENSMUSP00000006692 (PM)	ER19_MOUSE† Diphosphomevalonate decarboxylase	alcohol metabolism,lipid metabolism,cholesterol biosynthesis,phosphorus metabolism,phosphate	nucleotide binding,kinase activity,lyase activity,	

			metabolism,biosynthesis,macromolecule metabolism,		
bladder, thymus	ENSMUSP00000021130 (MS)				
bladder, cecum, gall bladder, ileum, spleen, thymus	ENSMUSP00000025773 (PM)	DPD4_MOUSE† DNA polymerase delta subunit 4	nucleobase, nucleoside, nucleotide and nucleic acid metabolism,cell cycle,cell growth and/or maintenance,	transferase activity,	nucleus,
bladder, cecum, gall bladder, lung, spleen	ENSMUSP00000071235 (MS)	PHX4_MOUSE† Putative per-hexamer repeat protein 4			integral to membrane,
brain, eye	ENSMUSP00000018595 (MS)	6330403K07Rik ‡ RIKEN cDNA 6330403K07 gene			
brain, eye	ENSMUSP00000026466 (MS)	TKNK_MOUSE† Neurokinin B [Precursor]	signal transduction,cell surface receptor linked signal transduction,tachykinin signaling pathway,neuropeptide signaling pathway,cell-cell signaling,transmission of nerve impulse,neurophysiological process,organismal movement,neuromuscular physiological process,	receptor binding,	
brain, cecum, duodenum, eye, ileum, jejunum, testis	ENSMUSP00000031268 (PM)	2310057D15Rik ‡ RIKEN cDNA 2310057D15 gene	metabolism,	hydrolase activity,	
brain, eye	ENSMUSP00000033198 (PM)	CRYM_MOUSE† Mu-crystallin homolog	morphogenesis,organogenesis,	structural constituent of eye lens,	cytoplasm,
brain	ENSMUSP00000034023 (MS)	SRG1_MOUSE † Scrapie-responsive protein 1 [Precursor]			
brain, duodenum, testis	ENSMUSP00000039418 (MS)	VHL_MOUSE† Von Hippel-Lindau disease tumor suppressor	cell cycle,cell growth and/or maintenance,protein ubiquitination,macromolecule metabolism,	protein binding,	nucleus,
brain, eye	ENSMUSP00000040342 (MS)	Pcsk1n ‡ proprotein convertase subtilisin/kexin type 1 inhibitor	peptide hormone processing,hormone metabolism,macromolecule metabolism,	enzyme inhibitor activity,	cytoplasm,

brain, duodenum, ileum, jejunum	ENSMUSP0000004 6012 (MS)	APEL_MOUSE† Apelin [Precursor]	signal transduction,cell surface receptor linked signal transduction,	receptor binding,	
brain, muscle, testis	ENSMUSP0000006 5423 (PM)	4932416N17Rik‡ RIKEN cDNA 4932416N17 gene		carbohydrate binding,	
brain, eye, testis	ENSMUSP0000006 7770 (MS)				
brain, kidney	ENSMUSP0000007 5395 (PM)		nucleobase, nucleoside, nucleotide and nucleic acid metabolism,regulation of transcription, DNA-dependent,regulation of transcription from Pol II promoter,transcription from Pol II promoter,positive regulation of metabolism,regulation of nucleobase,regulation of metabolism,positive regulation of transcription,positive regulation of nucleobase,positive regulation of transcription from Pol II promoter,	transcription cofactor activity,protein binding,	nucleus,
cecum, distal colon, proximal colon	ENSMUSP0000003 1146 (MS)	NEUU_MOUSE† Neuromedin U-23 [Precursor]	cell motility,signal transduction,cell surface receptor linked signal transduction,neuropeptide signaling pathway,		
cecum, duodenum, ileum, jejunum, testis, thymus	ENSMUSP0000004 5737 (PM)	CND2_MOUSE† Condensin complex subunit 2	M phase of mitotic cell cycle,nuclear division,cytokinesis,cell cycle,cell growth and/or maintenance,		nucleus,
cecum, ileum	ENSMUSP0000007 1936 (PM)	YV03_MOUSE† Hypothetical protein PP2447 homolog			
duodenum, ileum, jejunum, testis, thymus	ENSMUSP0000002 5704 (MS)	Cdca5‡ cell division cycle associated 5	cytokinesis,cell growth and/or maintenance,	protein binding,	nucleus,
duodenum	ENSMUSP0000005 6752				
eye	ENSMUSP0000003 6541 (PM)	Trappc5 ‡ trafficking protein	ER to Golgi vesicle-mediated transport,		endoplasmic reticulum,

		particle complex 5			
eye, spleen, testis, thymus	ENSMUSP00000063590 (PM)	N107_MOUSE† Nuclear pore complex protein Nup107	cell growth and/or maintenance,		nucleus,endomembrane system,integral to membrane,
eye, testis	ENSMUSP00000065836 (PM)	Nup188‡ nucleoporin 188		porin activity,	outer membrane,
gall bladder, liver	ENSMUSP00000016031 (PM)	HUTH_MOUSE† Histidine ammonia-lyase	organic acid metabolism,amino acid and derivative metabolism,amino acid metabolism,histidine metabolism,catabolism,biosynthesis,histidine family amino acid catabolism,amine metabolism,amine catabolism,	lyase activity,ligase activity,	cytoplasm,
gall bladder	ENSMUSP00000022722 (PM)	IRG1_MOUSE† Immune-responsive protein 1 [Fragment]			
gall bladder, kidney, liver	ENSMUSP00000023519 (PM)	HGD_MOUSE† Homogentisate 1,2-dioxygenase	organic acid metabolism,amino acid and derivative metabolism,amino acid metabolism,aromatic compound metabolism,catabolism,aromatic amino acid family catabolism,amine metabolism,amine catabolism,aromatic compound catabolism,	oxidoreductase activity,	
gall bladder, liver	ENSMUSP00000023832 (PM)	SM30_MOUSE† Regucalcin		ion binding,	nucleus,cytoplasm,
gall bladder, kidney, liver	ENSMUSP00000025249 (MS)	APOM_MOUSE† Apolipoprotein M	cell growth and/or maintenance,	lipid transporter activity,	
gall bladder, liver	ENSMUSP00000029645 (PM)	T23O_MOUSE† Tryptophan 2,3-dioxygenase	organic acid metabolism,amino acid and derivative metabolism,amino acid metabolism,tryptophan metabolism,biogenic amine metabolism,aromatic compound metabolism,amine metabolism,indole and derivative metabolism,heterocycle metabolism,	oxidoreductase activity,	

gall bladder, spleen	ENSMUSP0000003 2141 (PM)				
gall bladder, ileum, spleen, thymus	ENSMUSP0000003 4539 (PM)	Dcps‡ decapping enzyme, scavenger	nucleobase, nucleoside, nucleotide and nucleic acid metabolism,catabolism,macro molecule catabolism,macromolecule metabolism,	hydrolase activity,	nucleus,
gall bladder, liver, testis	ENSMUSP0000007 7192 (PM)	Pemt ‡ phosphatidylethanol amine N-methyltransferase	lipid metabolism,phospholipid metabolism,phosphatidylchol ine biosynthesis,phospholipid biosynthesis,biosynthesis,mac romolecule metabolism,membrane lipid biosynthesis,phosphatidylcho line metabolism,glycerophospholi pid biosynthesis,	transferase activity,	cytoplasm,m itochondrial membrane,e ndomembra ne system,integ ral to membrane,
heart, lung	ENSMUSP0000002 5554 (MS)	UTER_MOUSE† Uteroglobin [Precursor]		enzyme inhibitor activity,steroid binding,	
ileum	ENSMUSP0000000 1667 (MS)	CASK_MOUSE† Kappa-casein [Precursor]			
ileum, jujunum	ENSMUSP0000003 4359 (MS)				
ileum, testis	ENSMUSP0000003 7484 (MS)	1110039B18Rik‡ RIKEN cDNA 1110039B18 gene	regulation of transcription, DNA-dependent,	ion binding,	extracellular region
kidney	ENSMUSP0000002 3282 (PM)	MIOX_MOUSE† Inositol oxygenase	monosaccharide metabolism,alcohol metabolism,regulation of cell volume,signal transduction,intracellular signaling cascade,cell growth and/or maintenance,catabolism,cellul ar osmoregulation,membrane organization and biogenesis,carbohydrate catabolism,osmoregulation,m yo-inositol catabolism,hexose catabolism,cell homeostasis,macromolecule metabolism,alcohol catabolism,monosaccharide	oxidoreductase activity,	cytoplasm,in clusion body,

			catabolism,inositol phosphate (MS)ediated signaling,		
liver	ENSMUSP00000060495 (MS)	LCT2_MOUSE† Leukocyte cell-derived chemotaxin 2 [Precursor]	sensory perception,response to external stimulus,neurophysiological process,		
lung, spleen, thymus	ENSMUSP00000044493 (MS)	Tmem71 ‡ transmembrane protein 71			
lung, testis	ENSMUSP00000044862				
muscle, spleen, thymus	ENSMUSP00000030705 (PM)	RFA2_MOUSE† Replication protein A 32 kDa subunit	nucleobase, nucleoside, nucleotide and nucleic acid metabolism,cell cycle,cell growth and/or maintenance,	nucleic acid binding,	nucleus,
spleen, thymus	ENSMUSP00000027162 (MS)	Icos ‡ inducible T-cell co-stimulator	response to biotic stimulus,	antigen binding,receptor binding,peptide binding,	plasma membrane,integral to membrane,
spleen, thymus	ENSMUSP00000030180 (MS)	Sit1 ‡ suppression inducing transmembrane adaptor 1	cell activation,immune response,signal transduction,response to biotic stimulus,T-cell activation,immune cell activation,lymphocyte activation,regulation of T-cell activation,regulation of cell activation,	receptor signaling protein activity,protein binding,	plasma membrane,integral to membrane,
spleen, testis	ENSMUSP00000078413	FBS1_MOUSE† Fibrosin-1 [Fragment]		receptor binding,	
testis	ENSMUSP00000020990 (MS)	COLI_MOUSE† Corticotropin-lipotropin [Precursor]	signal transduction,cell surface receptor linked signal transduction,neuropeptide signaling pathway,	receptor binding,	
testis	ENSMUSP00000022019 (MS)	IL9_MOUSE† Interleukin-9 [Precursor]	immune response,response to biotic stimulus,	receptor binding,	
testis	ENSMUSP00000031876 (MS)	Stra8 ‡ stimulated by retinoic acid gene 8			
testis	ENSMUSP00000032481 (MS)	Acrbp‡ proacrosin binding protein	reproduction,gametogenesis,spermatid development,sexual reproduction,cell differentiation,	nucleic acid binding,protein binding,enzyme activator activity,	nucleus,cytoplasm,
testis	ENSMUSP00000003	TSX_MOUSE†			

	3691 (MS)	Testis-specific protein TSX			
testis	ENSMUSP0000004 5925 (MS)D	Fbxo15 ‡ F-box protein 15	macromolecule metabolism,	protein binding,	ubiquitin ligase complex,cytoplasm,
testis	ENSMUSP0000004 7588 (MS)	Coil‡ coilin		protein binding,	nucleus,
testis	ENSMUSP0000005 9448 (MS)	TUR8_MOUSE† Tumor rejection antigen P815A			nucleus,integral to membrane,
testis	ENSMUSP0000005 9630 (MS)	HSP3_MOUSE† Sperm protamine P3	reproduction,M phase of mitotic cell cycle,nuclear division,cell cycle,gametogenesis,cell growth and/or maintenance,sexual reproduction,	nucleic acid binding,	nucleus,chromosome,
thymus	ENSMUSP0000000 0028 (PM)	CC45_MOUSE† CDC45-related protein	nucleobase, nucleoside, nucleotide and nucleic acid metabolism,cell cycle,cell growth and/or maintenance,		nucleus,
thymus	ENSMUSP0000002 0765 (PM)	DPD2_MOUSE† DNA polymerase delta subunit 2	nucleobase, nucleoside, nucleotide and nucleic acid metabolism,cell cycle,cell growth and/or maintenance,	transferase activity,	nucleus,
thymus	ENSMUSP0000002 1164 (MS)	6720460F02Rik‡ RIKEN cDNA 6720460F02 gene			
thymus	ENSMUSP0000003 8204 (MS)	RAG2_MOUSE† V(D)J recombination-activating protein 2	cell activation,nucleobase, nucleoside, nucleotide and nucleic acid metabolism,immune response,response to biotic stimulus,morphogenesis,organogenesis,lymphocyte differentiation,cell differentiation,immune cell activation,lymphocyte activation,	nucleic acid binding,peroxidase activity,oxidoreductase activity,hydrolase activity,	nucleus,

*PM – pre-metazoan protein; MS – metazoan-specific protein