Molecular Biology and Evolution

Copy Number Alterations among Mammalian Enzymes Cluster in the Metabolic Network

Michaël Bekaert*,[1] and Gavin C. Conant[1,2]

[1]Division of Animal Sciences, University of Missouri, Columbia

[2]Informatics Institute, University of Missouri, Columbia

***Corresponding author:** E-mail: bekaertm@missouri.edu.

Associate editor: James McInerney

Abstract

Using two high-quality human metabolic networks, we employed comparative genomics techniques to infer metabolic network structures for seven other mammals. We then studied copy number alterations (CNAs) in these networks. Using a graph-theoretic approach, we show that the pattern of CNAs is distinctly different from the random distributions expected under genetic drift. Instead, we find that changes in copy number are most common among transporter genes and that the CNAs differ depending on the mammalian lineage in question. Thus, we find an excess of transporter genes in cattle involved in the milk production, secretion, and regulation. These results suggest a potential role for dosage selection in the evolution of mammalian metabolic networks.

Key words: comparative genomic, gene duplication, gene dosage, mammals, metabolic networks and pathways, milk.

Research article

Introduction

Metabolism's prominent role in facilitating most biological processes and in shaping the availability of ecological niches suggests that strong selective forces have fashioned the metabolic wiring (Raymond and Segre 2006). Likewise, metabolism's central importance to life has made the study of innovation among its systems a topic of particular interest. Morowitz and colleagues (Morowitz et al. 2000; Smith and Morowitz 2004) have argued that life originated through the exploitation of the metabolites of the tricarboxylic acid cycle (although see Bada and Lazcano 2002). Much metabolic innovation appears to have occurred early in evolution: both the general structure and reaction mechanisms of extant enzymes predate the divergence of the major domains of life (Caetano-Anollés et al. 2007). To understand such innovation, other researchers have considered how new catalytic proteins evolve: one model, which is known by several names (the adaptive amplification; adaptive radiation; or innovation, amplification, and divergence model) posits that new enzymes are co-opted from existing enzymes with low levels of the novel activity (Roth and Andersson 2004; Francino 2005; Bergthorsson et al. 2007). New enzymes are shaped by the action of natural selection on large duplicated arrays of these weakly functional enzymes, which are subsequently reduced to single copy once a high activity enzyme has evolved.

Gene duplication itself has long been seen as a major route to evolutionary novelty (Ohno 1970): one topic of recent interest is other mechanisms by which gene duplications promote innovation beyond the classic "neofunctionalization" pathway (reviewed in Conant and Wolfe 2008). One such mechanism is dosage selection, where the new trait is not the acquisition of a novel activity but rather an increased capacity for an existing reaction (Papp et al 2004; Kondrashov and Kondrashov 2006). A noteworthy example is in the amylase gene, responsible for starch digestion. In humans, high copy numbers of this gene are associated with populations having high-starch diets (Perry et al. 2007), suggesting a recent increase in the selective benefit of high amylase activity. Such dosage selection is only part of a larger pattern of requirements for dosage balance that also influence patterns of gene duplication (Papp et al 2003; Freeling and Thomas 2006; Birchler and Veitia 2007; Edger and Pires 2009). A familiar example of this phenomenon is the necessity of X-chromosome inactivation to compensate for dosage imbalances between male and female mammals (Payer and Lee 2008). Any fixed difference in copy number (i.e., duplication) between populations began life as a within-population copy number polymorphism. Such copy number variation contributes significantly to differences in transcript abundance among individuals (Stranger et al. 2007). More significantly, some copy number variations have been shown to be driven to high frequency by positive selection for increased expression of the corresponding gene (Gonzalez et al. 2005; Perry et al. 2007; Nair et al. 2008), highlighting how gene dosage modifications can be targeted by selection. However, the evolutionary constraints that act on gene dosage have yet to be fully elucidated. Discovery and functional assessment of gene dosage alterations between species is therefore an important element of understanding genome evolution.

Using the human metabolic network and orthologous genes from seven other mammals, we explored how differences in enzyme gene copy number in mammals are associated with the structure of the metabolic network. Our work is based on recent advances in cataloging and modeling metabolism. Such models can be used in a variety of ways, but one of the more common is to frame them as metabolic networks (Jeong et al. 2000). In this work, we

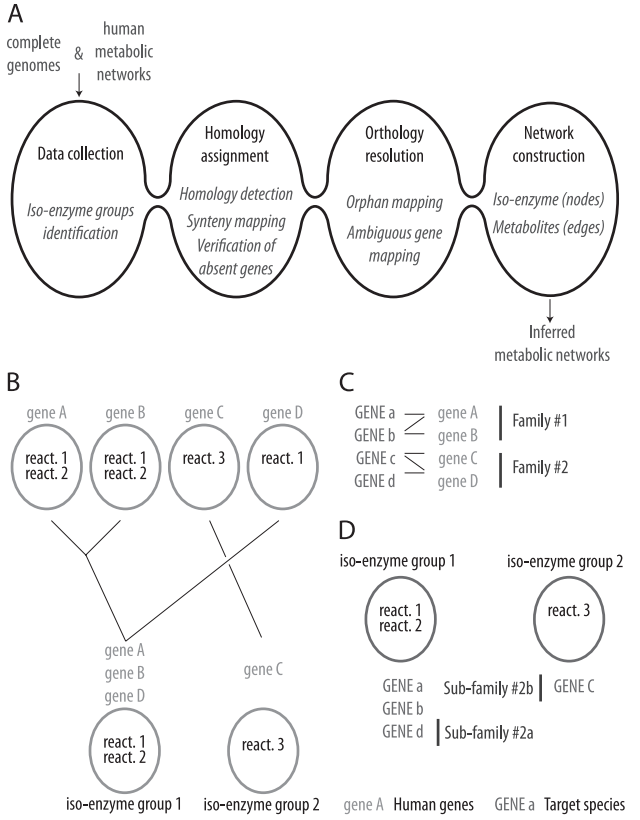

Fig. 1. Metabolic network inference process. (*A*) Overview of the inference process: Four steps produce several lists of genes/groups and eventually generate the full networks. (*B*) Isoenzyme group assignment: We group genes based on their participation in a common set of reactions. Thus, every gene in the group participates in at least a subset of the group's reactions (and in no other reactions). (*C*) Gene family assignment using shared homology: Homology is defined on the basis of a GenomeHistory search of the paired genomes (see Materials and Methods). (*D*) Repartition of the families from (*C*) on the basis of the isoenzyme groups from (*B*). The resulting subfamilies are used to populate the target species metabolic network.

use inferred genome-scale metabolic networks from humans (Duarte et al. 2007; Ma et al. 2007) to study copy number differences: we note that these networks are only two of several available from a variety of organisms (Duarte et al. 2004; Blank et al. 2005; de Oliveira Dal'Molin et al. 2010).

We asked whether the differences in enzyme copy number are distributed nonrandomly in the mammalian metabolic network. In yeast, it is known that enzymes that carry high flux and that lie in highly connected parts of the metabolic network are more likely to undergo duplication (Vitkup et al 2006), and we were curious whether similar forces were at play in multicellular eukaryotes.

Materials and Methods

An overview of our methodology is illustrated in figure 1*A*.

Data Collection and Preprocessing

Complete genome annotations for eight mammals, *Bos taurus* (cattle), *Canis familiaris* (dog), *Equus caballus* (horse), *Homo sapiens* (human), *Macaca mulatta* (macaque), *Mus musculus* (mouse), *Pan troglodytes* (chimpanzee), and *Rattus norvegicus* (rat) were obtained from Ensembl release 50 (Flicek et al. 2010). For the purposes of homology/orthology assignment, we obtained the longest transcript for each protein-coding gene along with its genomic location.

We downloaded two *H. sapiens* metabolic networks, MODEL6399676120 (Duarte et al. 2007) and MODEL2021747594 (Ma et al. 2007), from the BioModels database (Le Novere et al. 2006). Our goal was to use these *H. sapiens* networks to assign metabolic functions to genes in the other seven genomes. In order to do so, we must account for the fact that the only link between the *H. sapiens* metabolic network and the networks to be inferred in the other mammals is the orthology relationships between the genomes. As a result, we need to introduce a level of abstraction to the metabolic networks that we refer to as an "isoenzyme group." These groups attempt to represent sets of enzyme-coding genes all involved in the same reactions. To create them, we agglomerate reactions from the metabolic network in two steps. We first group enzyme-coding genes involved in identical reactions. We then sequentially merge any groups where the reactions of one group are a subset of reactions of second group. The net effect is to create isoenzyme groups such that each gene participates in a subset (possibly complete) of the reactions associated with that node (fig. 1*B*).

Orthology Assignment

Our orthology pipeline has been previously described (Conant 2009). An outline is provided here.

Homology Detection

As a first step, homologous genes within and between genomes are identified by running GenomeHistory (Conant and Wagner 2002) on the combination of two genomes, namely the reference *H. sapiens* genome and a second target genome. GenomeHistory identifies pairs of homologous genes using Blast (Altschul et al. 1997) and estimates their nonsynonymous and synonymous divergences (K_a and K_s, respectively) by maximum likelihood. We configured GenomeHistory to accept only gene pairs meeting the following criteria: *E*-value cutoff of 10^{-9}, protein length $\geq$75 amino acids, pairwise protein alignment length $\geq$70 amino acids, and pairwise protein sequence identity $\geq$45%.

Synteny Mapping

We identify initial orthologs between the two genomes as one-to-one matches in the GenomeHistory analysis (i.e., the two genes have no paralogs in their own genomes) that have synonymous divergence such that $K_s \leq 0.5$ (*P. troglodytes* and *M. mulatta*) or $K_s \leq 0.75$ (all others). Starting with such initial orthologs, any pair of genes that are immediate neighbors of such a pair and are also homologs are now defined to be orthologs themselves. Using these new orthologous pairs, the process is repeated until no further orthologs are located.

At the completion of this analysis, the genes in each genome can be divided into four classes: *orthologs*, *orphans*, *ambiguous*, and *absent*. The procedure for identifying *orthologs* has just been described. *Orphans* are genes in one genome that have no hits in the other genome once all the orthologs have been assigned. *Ambiguous* are genes shared between two genomes, but where the synteny and sequence information is insufficient to resolve orthology. *Absent* genes, as their names imply, have no significant homologs in the other genome.

Verification of Absent Genes

Metabolic genes in *H. sapiens* with no identified homologs in the target species were subjected to a second Blast analysis. We searched for these genes in the target genome with an *E*-value cutoff of 10^{-5}. This step allowed us to differentiate weak hits from genes that were truly absent in the target genome.

Metabolic Network Construction

Given the homology data from GenomeHistory, we defined a set of "gene families" that include genes across species. These gene families are defined on the basis of single-linkage clustering using the homology relationships determined by GenomeHistory (fig. 1*C*). We then defined subfamilies within these families such that all *H. sapiens* members of that subfamily with annotations in the metabolic network belong to the same isoenzyme group (fig. 1*D*).

Orphan Genes Mapping

We first attempted to assign the orphan genes (in each species) that fall perfectly into a subfamily. In *H. sapiens*, these orphans are already assigned if they are part of the metabolic network. In the other cases in *H. sapiens* and in all cases in the target species, orphans will not have direct network annotations. However, if such an orphan gene is a member of a gene family where that family is a member of exactly one isoenzyme group, we assign that orphan to that isoenzyme group. In cases where the gene family consists of two subfamilies in different isoenzyme groups, we make no assignment of that orphan to an isoenzyme group because its functional annotation is uncertain.

Ambiguous Genes Mapping

An ambiguous gene between the target genome and *H. sapiens* is one for which orthology cannot be established because the gene is a member of a large gene family in both genomes. In our metabolic analysis, lack of resolved orthology is an issue only if the members of that gene family in *H. sapiens* are split between several isoenzyme groups (i.e., several subfamilies). If all annotated orthologs have the same subfamily, we can reasonably assign all ambiguous genes of that same subfamily to the same isoenzyme group. After this reconciliation, we may also be able to assign functions to further remaining orphans in the same way.

Network Construction

At this point, the assignment of genes to subfamilies is complete. We next collected all such gene families that matched to only a single isoenzyme group. If a newly formed gene family belonged to more than one isoenzyme group, we checked whether this difference could be accounted for as one *H. sapiens* isoenzyme group being a subset of the other. The set of isoenzyme groups for a given gene family is then searched to see if one isoenzyme group can be assigned such that any remaining isoenzyme groups for that gene family are subsets.

Finally, each mapped isoenzyme group is defined as a node in our isoenzyme network. Edges between these nodes are defined by shared metabolites between the included reactions of the two isoenzyme groups (as reported in the *H. sapiens* metabolic network). The network is directed: for irreversible reactions if the product of one reaction is a reactant in the second, we define a directed edge. Reversible reactions are treated similarly, except that both directions of the reaction are allowed and handled independently. Thirteen currency metabolites (H^+, H_2O,

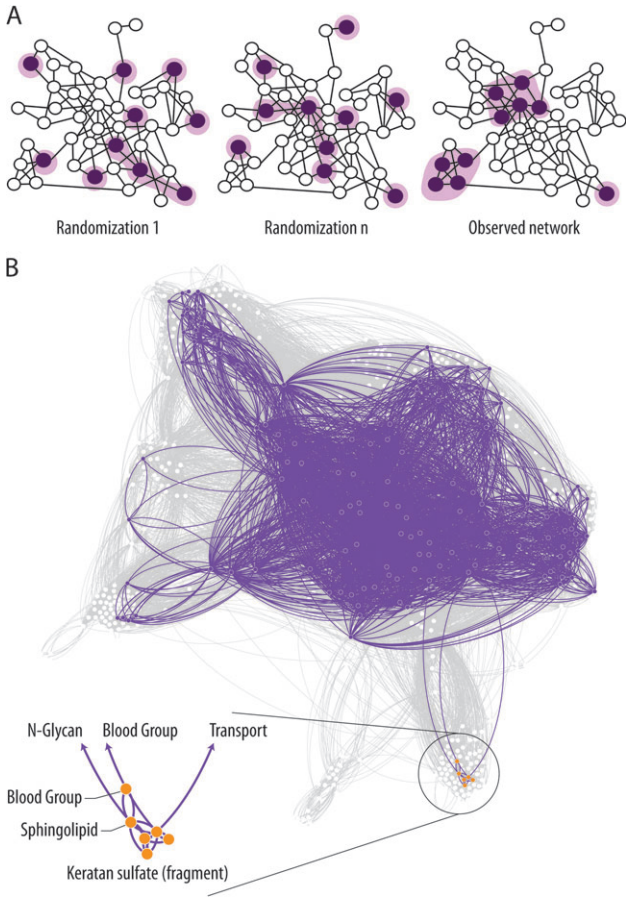

Fig. 2. Detecting duplication-enriched clusters. The 215 CNA nodes from *Bos taurus*. The clusters of isoenzyme group with CNAs of two nodes or more are in lavender. (*A*) Overview of the cluster detection method. The number and size of the connected components (shaded clusters in the figure) for the real network are calculated after non-CNA nodes (white) are removed. These clusters are then compared with those seen in randomized networks with the same number of CNA nodes (see Materials and Methods). (*B*) Clusters observed in *B. taurus*. The detail illustrates a subsection from the Golgi apparatus. The orange nodes are N-acetylglucosaminyl transferases: metabolic pathways associated with each node are indicated.

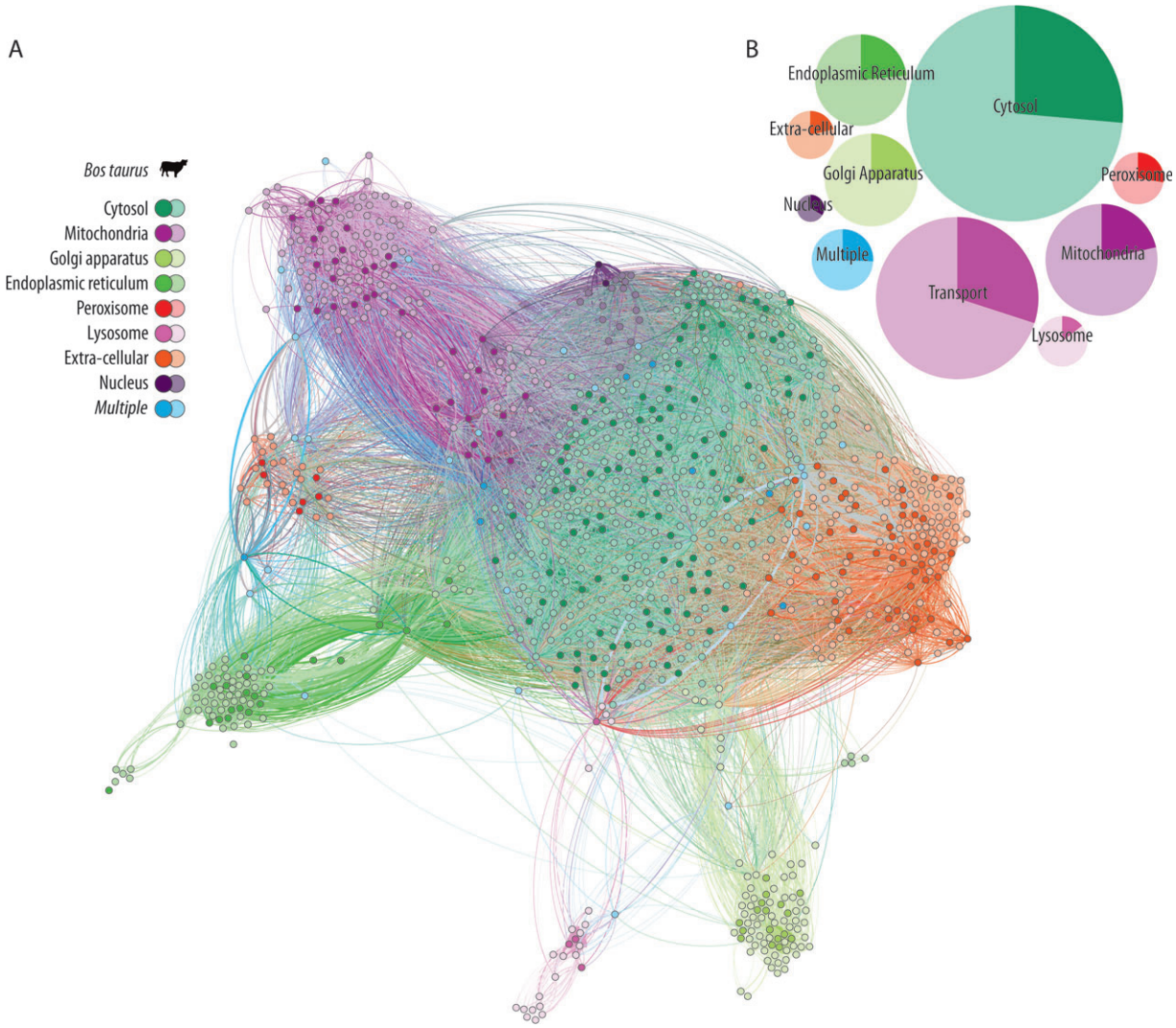

Fig. 3. Inferred metabolic network for *Bos taurus*. (*A*) The complete metabolic network, the cellular compartment, and the location of the nodes with CNAs are shown. Node and edge colors indicate the cellular compartment. Darkly shaded nodes are the isoenzyme groups with CNAs; pale nodes are those nodes without variation. (*B*) The distribution of the number of nodes with CNAs (the equivalent of the "node count" shown in fig. 5) for each compartment: the pale fraction again represents nodes without CNAs.

ATP, ADP, P_i, PP_i, Na^+, coenzyme A, O_2, NAD^+, NADH, $NADP^+$, and NADPH) were removed from all analyses in whichever compartments they occurred (Huss and Holme 2007). We then used the *H. sapiens* reference network of Duarte et al. (2007) to locate each metabolite in a cellular compartment. We then assigned each isoenzyme group to the compartment where the product of that reaction is located. A few reactions, coded by the same set of genes, are located in multiple compartments; hence they were assigned to a virtual compartment termed "Multiple." Our approach necessarily assigns the transporters to the destination compartment. In the second set of analyses, a specific compartment was created for the transporters (which we defined as reactions having metabolites in two compartments).

Using these fully constructed networks, we analyzed gene copy number alterations (CNAs) between each species and *H. sapiens* for each isoenzyme group node. For our purposes, we defined a CNA as any node possessing a different number of included genes in the target species as compared with that number in *H. sapiens*. The inferred metabolic networks were deposited at the EBI BioModels database using the references: MODEL1008120000–MODEL1008120006.

Visualization

The networks were visualized with Gephi v0.7 (Bastian et al. 2009) using the Force-based algorithm ForceAtlas. ForceAtlas works similarly to the Früchterman–Rheingold algorithm (Fruchterman and Reingold 1991), with the difference that the repulsion between two nodes is proportional to degree (n_1) × degree (n_2). Thus, the former will tend to bring nodes of degree 1 closer to their neighbors than will the latter.

Pathway Enrichment Analysis

Each gene from the reference *H. sapiens* networks is associated with one or more Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa et al. 2010). We imported the KEGG pathways (209 pathways, 5,281 genes).

Table 1. Summary of the Seven Inferred Metabolic Networks.

	Pan troglodytes	*Macaca mulatta*	*Mus musculus*	*Rattus norvegicus*	*Bos taurus*	*Equus caballus*	*Canis familiaris*
Genomes[a]							
Total protein-coding genes	19,829	21,905	23,493	22,503	21,036	20,322	19,305
Orthology stage[a]							
Orthologous pairs[b]	19,620	17,168	16,683	15,350	15,892	16,515	15,904
Orphan genes	59	1,457	1,963	2,280	1,656	1,824	1,002
Ambiguous genes	648	2,479	2,934	4,056	3,450	2,175	2,117
Orthology stage							
Gene families	1,509	1,500	1,382	1,340	1,357	1,347	1,362
Genes (Duarte et al. 2007)							
Assigned orthologs	1,412	1,363	1,389	1,256	1,295	1,342	1,296
Orphan genes	9	234	331	332	163	80	165
Ambiguous genes	36	79	334	317	379	105	163
Nodes (Duarte et al. 2007)							
Variable nodes	36	154	143	208	215	135	155
Invariable nodes	886	755	772	677	675	770	742
Unassigned groups in target	22	35	29	59	54	39	47
Total isoenzyme groups	944	944	944	944	944	944	944
Genes (Ma et al. 2007)							
Assigned orthologs	2,180	2,104	2,101	1,919	1,991	2,057	2,015
Orphan genes	7	283	136	353	187	93	183
Ambiguous genes	45	137	149	278	271	102	173
Nodes (Ma et al. 2007)							
Variable nodes	43	171	123	180	193	135	148
Invariable nodes	791	656	705	617	621	682	667
Unassigned groups in target	13	20	19	50	33	30	32
Total isoenzyme groups	847	847	847	847	847	847	847
Visualization[c]	Supplementary figure S1A	Supplementary figure S1B	Supplementary figure S1C	Supplementary figure S1D	Figure 3	Supplementary figure S1E	Supplementary figure S1F

NOTE.—Supplementary Material online.

[a] All protein-coding genes (including gene from the metabolic network).

[b] Includes inter- and intraspecies pairs.

[c] The Visualization field refers to the Duarte et al., inferred networks only.

We then collected for each pathway the number of isoenzyme groups with CNAs (from the total of 944 isoenzymes, 1,437 genes, and 171 pathways). We plotted the number of isoenzyme groups with CNAs in a pathway as a function of the total number of isoenzyme groups involved in that pathway. Simultaneously, we inferred the best linear fit of these two variables; doing so allowed us to calculate the normalized residual for each value. Any value with a residual significantly different from the expectation was defined as an outlier, that is, it has fewer or more CNAs than would be expected. These outliers represent a pool of reactions with the potential to be under copy number selection.

Milk Production Particulars

For each isoenzyme group including extracellular transporters and exhibiting CNAs, we collected the associated KEGG pathways (supplementary table S3, Supplementary Material online) in *B. taurus*.

Network Metric Calculations

We used Gephi to calculate network statistics, including in- and out-degrees, betweenness centrality, closeness centrality, network diameter, average clustering coefficient, the average shortest path, eccentricity, and network modularity (Newman 2006; Opsahl. et al. 2010). Statistical evaluations were performed with R (http://www.r-project.org/) using nonparametric tests (Kolmogorov–Smirnov test).

Clustering Tests

We were interested in to what extent CNAs tended to cluster in the metabolic network. To assess this, we first removed from the network all nodes without CNAs. We then calculated the number of connected components among the remaining nodes having CNAs (blue regions; fig. 2*A*). To assess whether these components were bigger than would be expected, we used network randomization. We began by copying the original network and reassigning the duplication status at random. The result was randomized networks with the same number of nodes with CNAs but for which the location of those nodes was random (fig. 2*A*). We again removed the unaltered nodes and computed connected components for the random networks. We performed 10,000 permutations and used the distribution of component sizes to determine whether the clusters in the real network were larger than expected. The procedure was implemented in C++ using the Boost Libraries (http://www.boost.org/). The code is available upon request.

Results

Reference Networks

The *H. sapiens* metabolic network of Ma et al. (2007) (Biomodel MODEL2021747594; 2007) consists of 2,716 metabolites, 2,566 reactions (1,052 with unique Enzyme Commission number), and 2,322 genes. There are 889 reactions associated

with at least one gene, and 2,339 metabolites are used by these reactions. Using this network, we established 847 isoenzyme groups. The overall network includes 2 isolated nodes not connected to others nodes: these nodes occur because some reactions do not have associated genes and hence cannot be part of isoenzyme groups. The *H. sapiens* isoenzyme network has 189,247 edges and the following network statistics: diameter: 4, average shortest path: 1.81, density: 0.267 (Sabidussi 1966; Coleman and More 1983).

The metabolic network of Duarte et al. (2007) (Biomodel MODEL6399676120; 2007) includes cellular compartments for all metabolites and includes 3,188 metabolites, 3,742 reactions, and 1,496 genes. Of the reactions, 2,307 reactions are associated with at least one gene, and 2,331 metabolites are used by these reactions. We established 944 isoenzyme groups. The overall network included 4 isolated nodes and 81,759 edges. Network statistics: diameter: 6, average shortest path: 2.31, density: 0.092.

Inferred Networks

Our goal was to study differences in enzyme copy number among eight mammalian genomes. We made an initial orthology assignment (see Materials and Methods) that produced a list of assigned orthologs and absent genes, as well as orphan and ambiguous genes. For each of the seven other mammals (fig. 3 and supplementary fig. S1, Supplementary Material online), we then mapped the *H. sapiens* network onto that target genome in four steps, with the aim of assigning target genes to isoenzyme groups so as to evaluate the CNAs. Supplementary figure S2 (Supplementary Material online) shows the assignment results in *B. taurus* at each step of the process. The first step assigns only orthologs, resulting in many unassigned nodes, whereas the full process significantly reduces this number. The remaining cases of unassigned isoenzyme groups may either represent true missing functions in the target genome or nonsequenced/annotated genes in that genome. Table 1 summarizes the results of the full process for the seven target species. Because some species have more metabolic genes than *H. sapiens*, the number of genes we can identify in the target genome was between 92% and 130% of the number of reference human genes. The number of assigned isoenzyme groups (groups we can identify in the target genome relative to the complete *H. sapiens* metabolic network) was between 94% and 98% of the total set of isoenzyme groups for both metabolic networks.

Pathways Enrichment Analysis

We investigated whether particular metabolic pathways seemed to be over or underrepresented among the gene CNAs. We extracted for each pathway the number of isoenzyme groups with CNAs (944 isoenzymes, 1,437 genes, 171 pathways) and estimated the overall relationship between the number of genes in each pathway and the number of CNAs using linear regression (fig. 4*A*). Adjusted R^2 were 0.659, 0.634, 0.389, 0.591, 0.448, 0.081, and 0.656 for *B. taurus*, *C. familiaris*, *E. caballus*, *M. mulatta*, *M. musculus*, *P. troglodytes*, and *R. norvegicus*, respectively. The extremely low R^2 value for the human–chimpanzee comparison is due to the very small number of CNAs found between these very recently diverged taxa. The normalized residuals (fig. 4*B*) were calculated and outliers ($>1.96\ \sigma$ or $<-1.96\ \sigma$) collected (table 2). Out of the 171 pathways present, 20 are significantly overrepresented with CNAs, and 9 are underrepresented. Examples of pathways found to have an excess of CNAs include the steroid hormone and retinol pathways as well as pathways involved in cytochrome P450 metabolism and associated with Huntington's disease.

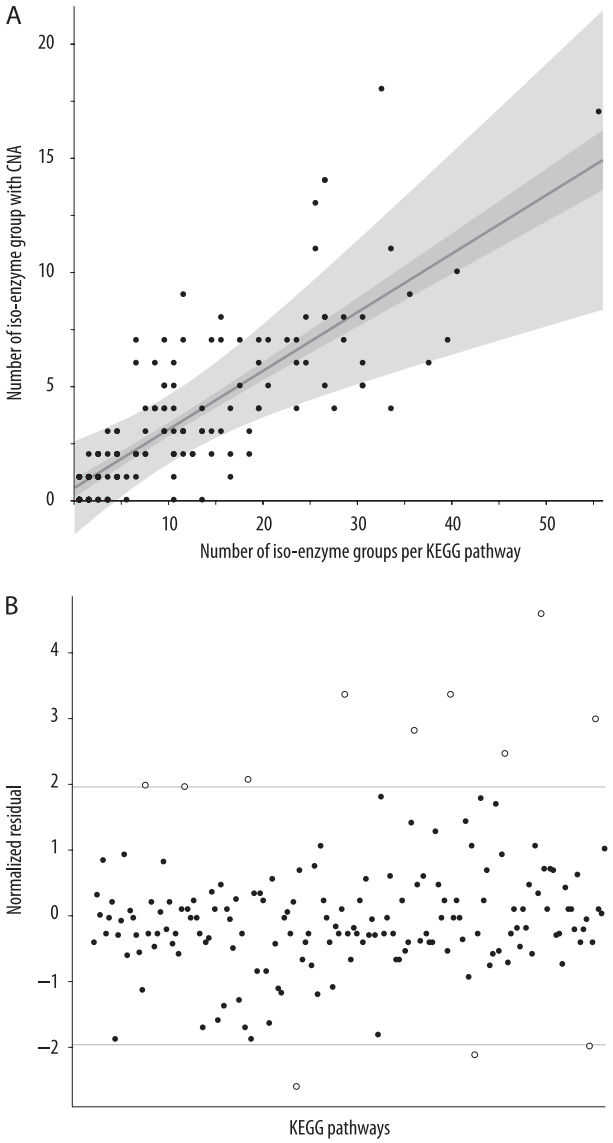

FIG. 4. Pathways enrichment analysis for *Bos taurus*. (*A*) Number of isoenzyme groups with CNAs in a pathway (*y* axis) versus the total number of isoenzyme groups involved in that pathway (*x* axis). The black line illustrates the linear regression line (adjusted R^2: 0.659); the darker shaded area represents one standard deviation from the expected trend, the lighter area, two standard deviations. (*B*) Normalized residuals from the linear regression. Gray lines are the significance thresholds (± 1.96 ; $\alpha = 0.05$). "●," Values not significantly different from the regression model prediction; "○," Significantly divergent values (outliers).

Table 2. List of the Pathways Over- or Undertargeted by CNAs.

KEGG Pathway[a]	*Pan troglodytes*	*Macaca mulatta*	*Mus musculus*	*Rattus norvegicus*	*Bos taurus*	*Equus caballus*	*Canis familiaris*
Glycolysis/gluconeogenesis		●	●	●			●
Steroid biosynthesis					○		
Steroid hormone biosynthesis			●	●	●	●	
Oxidative phosphorylation	●	●					●
Purine metabolism							●
Pyrimidine metabolism		○					○
Alanine, aspartate, and glutamate metabolism			○				
Glycine, serine, and threonine metabolism							○
Cysteine and methionine metabolism		●					
Taurine and hypotaurine metabolism						●	●
Starch and sucrose metabolism							●
Amino sugar and nucleotide sugar metabolism		○					
Glycosylphosphatidylinositol-anchor biosynthesis				○			
Glycerophospholipid metabolism			○	○	○		○
Arachidonic acid metabolism	●					●	
Linoleic acid metabolism	●					●	
Glycosphingolipid biosynthesis							●
Butanoate metabolism		●					
Retinol metabolism			●	●	●	●	
Metabolism of xenobiotics by cytochrome P450			●	●	●	●	
Drug metabolism—cytochrome P450			●	●	●	●	●
Drug metabolism—other enzymes			●		●		
ABC transporters					●		
PPAR[b] signaling pathway		○					
Peroxisome		○	○	○	○		
Vascular smooth muscle contraction						●	
Alzheimer's disease	●	●					●
Parkinson's disease	●				●		
Huntington's disease	●	●			●		●

NOTE.—"○," Pathway less variable in copy number than expected (9); "●" Pathway more variable than expected (20). PPAR, Peroxisome proliferator-activated receptor.

[a] For the 17 pathways under selection compared with the average for each species, the distribution of CNAs across the seven species is detailed.

[b] Peroxisome proliferator-activated receptor.

To explore the discrepancy in observed CNA frequencies, we evaluated the distribution of the isoenzyme groups showing the CNAs across the seven species, examining how often a given isoenzyme group exhibited a CNA. Supplementary figure S3 (Supplementary Material online) shows the proportion of isoenzyme groups with CNAs in a given number of species (using the network of Duarte et al. 2007) as compared with the expected distribution whether the seven networks were independent and CNAs were randomly occurring. The two distributions are statistically distinguishable, but we cannot rule out the influence of the phylogenic relationships among the species.

Metrics

We next assessed if there was an association between several network statistics (node degree, betweenness, and closeness centrality; Sabidussi 1966; Brandes 2001) and the propensity of a node to possess a CNA. These measures all evaluate, to one degree or another, the "importance" of a particular enzyme (node) in the metabolic network. In other words, nodes of high betweenness or degree represent parts of the network that affect many other nodes, meaning that damage to them is likely to have large effects on metabolism. Because the network of Ma et al. (2007) lacked compartment information, we excluded the latter from this analysis. As described in the Materials and Methods, we introduced a virtual cellular "Transport" compartment: we performed our network statistics analysis both with and without this compartment. For many cellular compartments, we found significant associations between network statistics and CNAs when transport reactions were included in those compartments, but the association was no longer significant when the transport reactions were removed (supplementary tables S1 and S2, Supplementary Material online). The distribution of copy number changes is nonrandom, as judged both by the structure of the network itself and by the distribution of network statistics for nodes with changes in copy number (fig. 5). Specifically, the order Rodentia (fig. 5, node 1) shows copy number changes among Golgi apparatus transporters, whereas in the superorder Laurasiatheria (fig. 5, node 2), we find an excess of duplication/loss among the extracellular transporters. It is especially intriguing that although these patterns are lineage specific, there is an overall trend toward apparent duplication among the transporters.

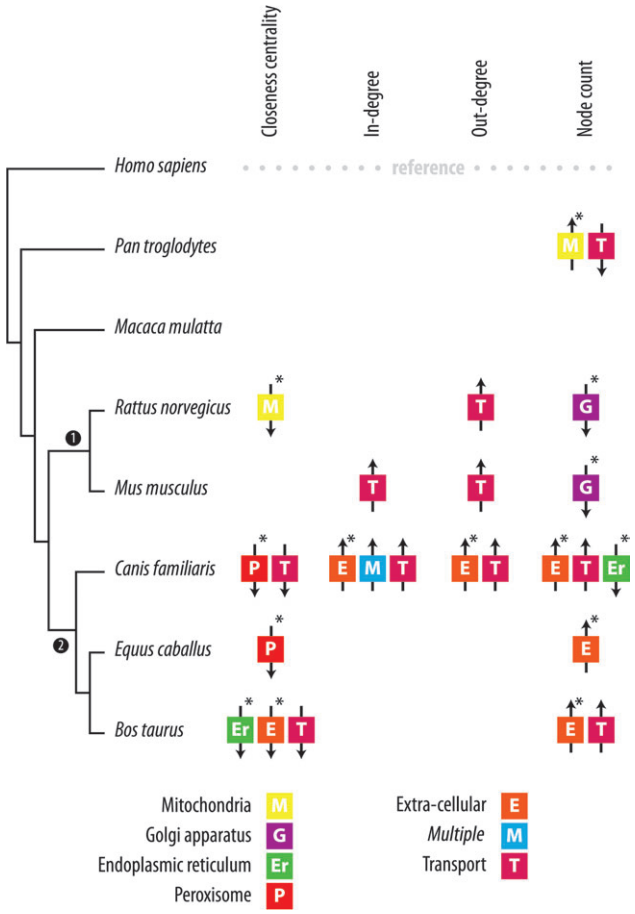

Fig 5. Association of CNAs and network statistics. Several metrics have been used to describe the networks and the distribution of CNAs between species. Using the species tree (Murphy et al. 2007; Prasad et al 2008) at left, we show how the number of CNAs increases with greater evolutionary distance from the reference human network ("node count," far right). For each lineage, we show the cellular compartments for which the metric in question significantly differs between nodes with CNAs and those without (supplementary tables S1 and S2, Supplementary Material online). In-/out-degrees describe the number of reactants or products for each isoenzyme group, respectively. Closeness centrality evaluates the proximity of a node to every other isoenzyme node. The node count is the number of isoenzyme groups with CNAs. The squares indicate the compartment name. The "*" denotes the transporters from that compartment rather than the compartment itself. The arrows indicate whether the nodes with CNAs have an increased or decreased mean value compared with the invariant nodes. Labeled branch points in the phylogeny: 1, Rodencia; 2, Laurasiatheria. For example, the *Rattus norvegicus* network shows a *P* value significant for the mitochondria "Closeness centrality" metric (supplementary table S1, Supplementary Material online): 0.0270. It is reported as a yellow square. By adding a "transporter" compartment (supplementary table S2, Supplementary Material online) and subtracting the mitochondrial transporters from the mitochondrial compartment, this value became nonsignificant: 0.6245. This illustrates that the mitochondria transporters are carrying the signal; an "*" indicates cases where this is true.

Clustering

The nonrandom distribution of CNAs among pathways and cellular compartments led us to ask whether the CNAs might be clustered in the overall network. We thus searched the network for clusters enriched in CNAs. To do so, we first removed from the network all nodes without CNAs and then calculated the number of connected components among the subset of nodes with CNAs. Note that this removal implicitly removes any edges that end at nodes lacking CNAs, drastically reducing the number of edges in the network. The result is to reduce the network from one large component (of the form illustrated in fig. 3) to numerous isolated ones (blue regions in fig. 2*A*). We assessed the statistical significance of these induced clusters by network randomization (fig. 2*A*; see Materials and Methods). For all seven genomes surveyed, and using either network (Ma et al. 2007 or Duarte et al. 2007, when disregarding the compartmental information), we found that there were significantly fewer and larger connected components (e.g., clusters) in the real network than would be expected based on the distribution of component sizes and number seen in the randomized networks ($P < 0.001$). When we examine the compartmentalized networks, we find fewer cases of significant clustering (table 3), likely because our test for significant clustering relies on interconnected metabolic pathways, pathways that can be hidden in the compartmentalized analysis when shared metabolites are present in separate cellular compartments. The real networks also showed higher than expected in- and out-degrees within these clusters (table 3). For example, the largest cluster found in the *B. taurus* metabolic network includes 104 isoenzyme groups linking numerous metabolic pathways. The orange nodes in figure 2*B* illustrate a subsection of this cluster from the Golgi apparatus: All the orange nodes are N-acetylglucosaminyl transferases belonging to keratan sulfate biosynthesis, sphingolipid metabolism, or blood group biosynthesis pathways. These nodes are linked to each other and are connected to other pathways and represent a group of genes that are present in higher copy numbers in *B. taurus* than in *H. sapiens* (except for nodes also belonging to sphingolipid metabolism).

Discussion

Using an approach that allows us to map more than 94% of the *H. sapiens* metabolic network onto other mammalian species, we have explored the patterns of CNAs across these mammalian networks. Despite the fact that mammalian genomes have significant differences in gene content and organization (Murphy et al. 2001), the metabolic network topology is relatively conserved across this group (using *H. sapiens* as reference). Nonetheless, there are reasonably large numbers of CNAs observed (table 1): many of these variations appear to involve transporter proteins (fig. 5).

Of course, one important caveat of our analysis is that we have only the expertly curated metabolic networks from *H. sapiens* to use as the basis of our analyses. Thus, we cannot compare the networks from the other seven species directly but must instead contrast their evolutionary path with that in humans. Having a second out-group metabolic network would clarify the evolutionary history of the CNAs. However, we note that although two metabolic

Table 3. Details of the Clustering Analysis.

	Pan troglodytes	*Macaca mulatta*	*Mus musculus*	*Rattus norvegicus*	*Bos taurus*	*Equus caballus*	*Canis familiaris*
Ma et al. (2007)							
Bigger maximal component		●	●	●	●		●
Fewer components		●	●	●	●	●	●
Higher average in-degree						●	
Higher average out-degree in the real data						●	
Duarte et al. (no compartment)[a]							
Bigger maximal component		●	●	●	●		
Fewer components		●		●	●		
Higher average in-degree	●	●	●	●	●	●	●
Higher average out-degree in the real data	●	●	●	●	●	●	●
Duarte et al. (2007)							
Bigger maximal component				●			●
Fewer components				●			●
Higher average in-degree				●	●		●
Higher average out-degree in the real data				●	●		●

[a] The network of Duarte et al. (2007) was decompartmentalized for comparison with the network of Ma et al. (2007).
"●," Results significant ($P < 0.001$).

networks for *M. musculus* have been published (Sheikh et al. 2005; Selvarasu et al. 2010), they are less than ideal for our purposes as they only cover a small proportion of the set of enzyme-coding genes (473 and 724, respectively, supplementary fig. S4, Supplementary Material online), as compared with more than 1,400 genes for both human networks and more than 1,800 genes placed in the *M. musculus* network by our techniques above.

Our results clearly show that the CNAs in metabolism are not all selectively neutral: they cluster in the metabolic network, creating a large interconnected subnetwork within the core metabolic network. Random distributions of CNAs do not mimic this pattern, indicating that some form of natural selection has acted to preserve duplications (or to favor gene losses) in the network. This result might seem to conflict with the known importance of genetic drift in preserving eukaryotic duplicate genes (Lynch and Conery 2003). However, we suggest that this difficulty is probably mostly one of perspective, especially as the comparisons being made here tend to be over larger evolutionary distances where selection may play a more prominent role. As mentioned, many of the CNAs involve transporter genes that show significantly different patterns of evolution in copy number than does the remainder of the metabolic network. These transporter alterations are not uniform across the mammalian phylogeny but vary by cellular compartment according to the lineage in question (fig. 5). Together with the presence of very large clusters interconnected by these transporters (table 3 and fig. 2*B*), the results may indicate that transporter duplication is favored in more central regions of the network, leading to the higher in- and out-degrees of the CNA-associated transporters in figure 5.

Similarly, we find that core sugar metabolism (glycolysis/gluconeogenesis, and glycerophospholipid metabolism) show an excess of CNAs, recalling known patterns of duplication both in vertebrates (Steinke et al. 2006) and in other organisms (Conant and Wolfe 2007). Curiously, although the peroxisome (as defined in KEGG) mainly consists of membrane proteins and transporters, this organelle actually possesses fewer CNAs than does the network at large. We attribute this difference to the relative isolation of this region of the metabolic network: closeness centrality among these reaction nodes is also low (fig. 5).

These general observations support a role for gene dosage as one factor in preserving duplications in the mammalian metabolic network. The association of CNAs and transporters is especially intriguing given that *Saccharomyces cerevisiae* (bakers' yeast) cells under selection from a glucose-limited environment undergo multiple tandem duplications of their high-affinity glucose transporters (Brown et al. 1998). A particularly interesting illustration of the related phenomenon in mammals is in the metabolism of cattle milk production. As shown by the figure 3, a significant excess of extracellular transporters from *B. taurus* involved in milk production possess CNAs. The pathways involved include milk production itself (Jensen 1995), as well as its regulation (Ingvartsen and Friggens 2005; Hammon et al. 2007) and the induction of mammary angiogenesis (Spitsberg 2005; Nakajima et al. 2009). Notably, one of the transporters showing CNAs between *H. sapiens* and *B. taurus* (DGAT1) is also the site of a quantitative trait locus for milk production (Grisart et al. 2002). These results are also consistent with Lemay et al. (2009). We hypothesize that natural or artificial selection for milk production has shaped these CNAs; indeed, it may be the case that alteration of transporter gene dosage represents one of the more evolutionarily "easy" adaptations. That these CNAs are cattle-specific amplifications is clear from the fact that no similar alterations are seen when comparing other mammalian networks to the human network.

CNAs between species likely represent a complex mixture of dosage-related adaptations, cases of enzymatic "neo-functionalization" through gene duplication, artifacts of genetic drift, and likely other processes we have yet to identify. Given the ability to not only identify copy number changes between sequenced genomes but also to put them into the functional context of a biological network, it

should soon be possible to tease apart the relative contributions of these various mechanisms and even potentially exploit copy number alteration in fields such as agriculture and medicine.

Supplementary Material

Supplementary tables S1–S3 and figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank Patrick P. Edger, Corey M. Hudson, and J. Chris Pires for helpful discussions. This work was supported by the Reproductive Biology Group of the Food for the 21st Century program at the University of Missouri and by a Research Board grant from the University of Missouri to G.C.C.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

Bada JL, Lazcano A. 2002. Origin of life. Some like it hot, but not the first biomolecules. *Science* 296:1982–1983.

Bastian M, Heymann S, Jacomy M. 2009. *Gephi: an Open Source Software for Exploring and Manipulating Networks. International AAAI Conference on Weblogs and Social Media.*

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: evolution of new genes under continuous selection. *Proc Natl Acad Sci U S A.* 104:17004–17009.

Birchler JA, Veitia RA. 2007. The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell.* 19:395–402.

Blank LM, Lehmbeck F, Sauer U. 2005. Metabolic-flux and network analysis of fourteen hemiascomycetous yeasts. *FEMS Yeast Res.* 5:545–558.

Brandes U. 2001. A faster algorithm for betweenness centrality. *J Math Sociol.* 25:163–177.

Brown CJ, Todd KM, Rosenzweig RF. 1998. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol Biol Evol.* 15:931–942.

Caetano-Anollés G, Kim HS, Mittenthal JE. 2007. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc Natl Acad Sci U S A.* 104:9358–9363.

Coleman TF, More JJ. 1983. Estimation of sparse Jacobian matrices and graph coloring blems. *SIAM J Num Anal.* 20:187–209.

Conant GC. 2009. Neutral evolution on mammalian protein surfaces. *Trends Genet.* 25:377–381.

Conant GC, Wagner A. 2002. GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res.* 30:3378–3386.

Conant GC, Wolfe KH. 2007. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Mol Syst Biol.* 3:129.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet.* 9:938–950.

de Oliveira Dal'Molin CG, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK. 2010. AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiol.* 152:579–589.

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO. 2007. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A.* 104:1777–1782.

Duarte NC, Herrgård MJ, Ø B, Palsson . 2004. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* 14:1298–1309.

Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* 17:699–717.

Flicek P, Aken BL, Ballester B, et al. 57 co-authors. 2010. Ensembl's 10th year. *Nucleic Acids Res.* 38:D557–D562.

Francino MP. 2005. An adaptive radiation model for the origin of new gene functions. *Nat Genet.* 37:573–577.

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 16:805–814.

Fruchterman TMJ, Reingold EM. 1991. Graph drawing by force-directed placement. *Software: Pract Exp.* 21:1129–1164.

Gonzalez E, Kulkarni H, Bolivar H, et al. (21 co-authors). 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 307:1434–1440.

Grisart B, Coppieters W, Farnir F, et al. (13 co-authors). 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Res.* 12:222–231.

Hammon HM, Bellmann O, Voigt J, Schneider F, Kuhn C. 2007. Glucose-dependent insulin response and milk production in heifers within a segregating resource family population. *J Dairy Sci.* 90:3247–3254.

Huss M, Holme P. 2007. Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. *IET Syst Biol.* 1:280–285.

Ingvartsen KL, Friggens NC. 2005. To what extent do variabilities in hormones, metabolites and energy intake explain variability in milk yield? *Domest Anim Endocrinol.* 29:294–304.

Jensen RG. 1995. Handbook of milk composition. San Diego (CA): Academic Press.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L. 2000. The large-scale organization of metabolic networks. *Nature* 407:651–654.

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38:D355–D360.

Kondrashov FA, Kondrashov AS. 2006. Role of selection in fixation of gene duplications. *J Theor Biol.* 239:141–151.

Lemay DG, Lynn DJ, Martin WF, et al. (19 co-authors). 2009. The bovine lactation genome: insights into the evolution of mammalian milk. *Genome Biol.* 10:R43.

Le Novere N, Bornstein B, Broicher A, et al. (12 co-authors). 2006. BioModels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.* 34:D689–D691.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* 302:1401–1404.

Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, Goryanin I. 2007. The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol.* 3:135.

Morowitz HJ, Kostelnik JD, Yang J, Cody GD. 2000. The origin of intermediary metabolism. *Proc Natl Acad Sci U S A.* 97:7704–7708.

Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res.* 17:413–421.

Murphy WJ, Stanyon R, O'Brien SJ. 2001. Evolution of mammalian genome organization inferred from comparative gene mapping. *Genome Biol.* 2:REVIEWS0005.

Nair S, Miller B, Barends M, Jaidee A, Patel J, Mayxay M, Newton P, Nosten F, Ferdig MT, Anderson TJ. 2008. Adaptive copy number evolution in malaria parasites. *PLoS Genet.* 4:e1000243.

Nakajima K-I, Nakamura M, Ishisaki A, Kozakai T. 2009. Synergistic effect of dexamethasone and prolactin on VEGF expression in bovine mammary epithelial cells via p44/p42 MAP kinase (vascular endothelial growth factor, mitogen-activated protein kinase). *Australas J Anim Sci*. 22:788–795.

Newman ME. 2006. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*. 103:8577–8582.

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

Opsahl T, Agneessens F, Skvoretz J. 2010. Node centrality in weighted networks: generalizing degree and shortest paths. *Soc Netw*. 32:245–251.

Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424:194–197.

Papp B, Pál C, Hurst LD. 2004. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* 429:661–664.

Payer B, Lee JT. 2008. X chromosome dosage compensation: how mammals keep the balance. *Annu Rev Genet*. 42:733–772.

Perry GH, Dominy NJ, Claw KG, et al. (13 co-authors). 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet*. 39:1256–1260.

Prasad AB, Allard MW, Green ED. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol Biol Evol*. 25:1795–1808.

Raymond J, Segre D. 2006. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* 311:1764–1767.

Roth JR, Andersson DI. 2004. Adaptive mutation: how growth under selection stimulates Lac(+) reversion by increasing target copy number. *J Bacteriol*. 186:4855–4860.

Sabidussi G. 1966. The centrality index of a graph. *Psychometrika* 31:581–603.

Selvarasu S, Karimi IA, Ghim GH, Lee DY. 2010. Genome-scale modeling and in silico analysis of mouse cell metabolic network. *Mol Biosyst*. 6:152–161.

Sheikh K, Forster J, Nielsen LK. 2005. Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus. *Biotechnol Prog*. 21:112–121.

Smith E, Morowitz HJ. 2004. Universality in intermediary metabolism. *Proc Natl Acad Sci U S A*. 101:13168–13173.

Spitsberg VL. 2005. Invited review: bovine milk fat globule membrane as a potential nutraceutical. *J Dairy Sci*. 88:2289–2294.

Steinke D, Hoegg S, Brinkmann H, Meyer A. 2006. Three rounds (1R/2R/3R) of genome duplications and the evolution of the glycolytic pathway in vertebrates. *BMC Biol*. 4:16.

Stranger BE, Forrest MS, Dunning M, et al. (17 co-authors). 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315:848–853.

Vitkup D, Kharchenko P, Wagner A. 2006. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol*. 7:R39.