PHARMACOPHYLOGENOMICS: GENES, EVOLUTION AND DRUG TARGETS

David B. Searls

Phylogenomics, which advocates an evolutionary view of genomic data, has been useful in the prediction of protein function, of significant sequence and structural elements, and of protein interactions and other relationships. Although such information is important in characterizing individual pharmacological targets, evolutionary analyses also indicate new ways to view the overall space of gene products in terms of their suitability for therapeutic intervention. This view places increased emphasis on the comprehensive analysis of the evolutionary history of targets, in particular their orthology and paralogy relationships, the rate and nature of evolutionary change they have undergone, and their involvement in evolving pathways and networks.

PHYLOGENOMICS
The application to genomics of principles and techniques from evolutionary biology, to achieve a better understanding of gene function. 'Pharmacophylogenomics' is the use of phylogenomics in aid of drug discovery, through improved target selection and validation.

Bioinformatics Division, Genetics Research, GlaxoSmithKline Pharmaceuticals, 709 Swedeland Road, P.O. Box 1539, King of Prussia, Pennsylvania 19406, USA.
e-mail: David_B_Searls@gsk.com
doi:10.1038/nrd1152

In coining the term PHYLOGENOMICS some five years ago, Eisen suggested that genomics had lagged behind other biological disciplines in deriving benefit from the molecular fossil record and the vast natural experiment of evolution[1,2]. Phylogenomic analysis involves a comparison of genes and gene products across a number of species, generally in the context of whole genomes, characterizing HOMOLOGUES and seeking further insights arising from the evolutionary process itself. Such an approach, in its simplest form, has long been useful in detecting conserved functional residues in multiple alignments of homologous proteins, a theme that has been elaborated to encompass ever-more complex patterns of conservation[3]. This principle has been extended to such applications as finding key regulatory elements in non-coding genomic regions (BOX 1) and delineating specificity determinants in proteins[4]. Such analyses are not limited to primary sequence data; phylogenomics encompasses non-homology-based inferences[5], and essentially the same principles can be extended to structures, pathways, expression patterns and so forth. More broadly, evolutionary thinking has offered fresh viewpoints to a number of fields that are relevant to drug discovery, including physiology[6], immunology[7], neurosciences[8], epidemiology[9], and what is sometimes called 'Darwinian medicine'[10], which places human health and disease within an evolutionary perspective.

The drug-discovery enterprise has long had a keen interest in the ORTHOLOGUES and PARALOGUES of putative targets (BOX 2), as well as the pathways in which they participate. What might be called the traditional view of orthologues, though, has tended to focus on pharmacologically well-studied species such as the rat, in the interest of developing assays and disease models. At the same time, paralogues have been studied primarily to collect families of known tractable targets and to outline selectivity issues. Interest in pathways in model organisms has extended to gaining an understanding of pathophysiology and to seeking routes for expansion from biologically interesting but problematic targets to more tractable ones.

By contrast, it will be seen that a phylogenomic view of orthologues extends beyond the usual model organisms to embrace a wider swath of evolutionary history using full PHYLOGENETIC RECONSTRUCTIONS and related techniques, all of which are better suited to the determination of function and, most significantly, of changes in function over time (FIG. 1). Similarly, the study of paralogues and pathways in an evolutionary context can provide insights into broader issues of PLEIOTROPY and functional REDUNDANCY that are of particular concern for drug discovery.

Box 1 | **Footprinting and shadowing**

During World War II, the mathematician Abraham Wald was asked to analyse patterns of bullet holes in aircraft returning from combat missions. Legend has it that the military proposed to add extra armour at those points where the most holes were found. Wald pointed out that in all likelihood the density of hits was uniform, and that in areas where fewer hits were observed, it was because the planes hit there were not returning. So, he argued, the crucial points were where the planes were (apparently) hit less often[132].

Substitute mutations for bullets and Darwinian selection for the fortunes of war, and one can discern the essence of phylogenetic footprinting as well as many related forms of analysis. Although multiple alignments of proteins have long been used to detect conserved, and therefore functionally significant, residues, only more recently have non-coding nucleotide sequences been systematically examined for the same purpose[133]. In a typical footprinting experiment, human and mouse sequences upstream of related genes are aligned, and regions of higher conservation are searched for consensus regulatory elements; although ordinarily the latter produce many false positives, when such signals coincide with regions of high interspecies similarity they have been shown to be far more reliable[134].

Phylogenetic footprinting requires that species be at sufficient evolutionary distance for peaks of conservation to stand out from a divergent background. Primates, for example, are too closely related for this purpose, and this is obviously a disadvantage when one is interested in biological traits unique to primates. However, a new technique called phylogenetic shadowing can take advantage of the additive collective divergence of a large number of primate species, together with knowledge of the precise phylogenetic relationships among them, to extract sufficient signal to identify primate-specific functional elements; this was done, for example, for the recently evolved gene encoding apolipoprotein A, a biomarker for cardiovascular disease[135]. Such an experiment strikingly demonstrates the general principle that the greater the number and diversity of genomes available, the more information that can be derived — and this fact is the foundation of the pharmacophylogenomic approach.

Target orthology

A strong motivation for the further study of orthology of drug targets is the fact that species differences of various kinds — for instance, in pathophysiology or drug metabolism — frequently hamper the progression of targets and compounds, often after quite significant investment. This indicates that even a marginally improved understanding of species differences could have a major impact on the cost of developing medicines. The sequencing of the genomes of new model organisms, and in particular additional mammalian genomes, will make feasible the construction of complete orthology maps among relevant species, similar to the efforts already undertaken in simpler organisms[11]. Such orthology maps, combined with expression data and annotated with pathway information, will serve as frameworks for reasoning about species differences — for example, supporting efforts in predictive toxicology based on expression profiles. However, any such effort must go beyond the popular notion of orthologues as the 'corresponding' genes in different species.

Establishing orthology. A common and often successful method for finding orthologues is to identify pairs of genes that constitute each others' highest-scoring BLAST hits between the species in question — in other words, based on straightforward sequence similarity. However, not only does this approach assume that the respective genomes are correct and complete in their sequencing and assembly, but also that the genes themselves have been correctly identified and delineated, including splice variants. (Since homology information is used in many gene-calling procedures, there is the potential for a dangerous circularity, as has also been noted with regard to gene annotation[12].) Similarity searching itself can be quite challenging, particularly over greater evolutionary distances[13] and when multiple protein domains are involved[14]; either situation might require even more complex analyses of structural similarity, which can be important for accurate alignment[15], for the proper interpretation of conserved elements such as active sites[16], and for placing similarity in the context of an emerging understanding of protein-fold space[17]. A particular complicating factor in this regard is INCONGRUENT EVOLUTION (BOX 3), as when different domains of the same protein, such as the ligand-binding and DNA-binding domains of nuclear receptors, seem to have a disparate evolutionary history[18].

Not only does reducing similarity to a single numeric score fail to account for the fine structures of both genes and gene products, it does not really address the question of how an ensemble of present-day homologues could have been derived by a plausible evolutionary history[19]. The simplistic 'top BLAST hit' approach can be confounded, for example, when the true orthologue has been lost or duplicated since speciation (BOX 2), or when differing rates of evolution distort relationships[2]. Not only are protein families well known for such rate variations, but paralogues occurring in repetitive multigene families can be susceptible to a variety of homogenizing influences collectively termed CONCERTED EVOLUTION[20]. The occurrence of similar genes in corresponding positions within regions of conserved SYNTENY between species can add strong evidence for orthology, but still is not absolute proof; for instance, human and mouse major histocompatability complex (MHC) class I genes that are clearly not orthologues nevertheless occupy the same chromosomal framework[21].

Pairwise BLAST comparisons can be considerably improved by large-scale clustering of similarities among sets of homologues from whole genomes[11], thereby accounting for the information available from many genes and species. However, such clusters still do not represent the actual evolutionary relationships among homologues[2]. A full phylogenetic reconstruction, incorporating as many homologues and intervening species as possible, can provide a much more reliable and informative orthologue call with appropriate statistical support. A number of techniques and tools, such as the popular PHYLIP and PAUP packages, are available to perform phylogenetic reconstruction[22], and though such analyses can be laborious, several new programs have been designed specifically to characterize orthologues with a much higher degree of automation[23,24].

Added to the many challenges in establishing orthology is the most significant issue of all, the fact that the strict definition of orthology says nothing at all about function; yet function is the crucial relationship for target validation, and in particular for anticipating species differences. By no means does orthology guarantee common function (nor, for that matter, does common

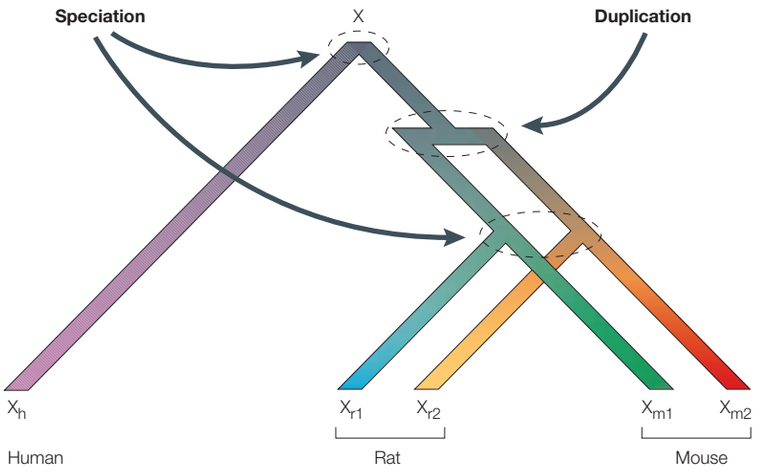
HOMOLOGUES
Genes that are similar by virtue of having derived from the same ancestral gene. The similarity might be evident in the DNA sequences of the genes, or in the sequence and/or structure of the gene products. Similarity does not guarantee homology, as unrelated sequences can undergo convergent evolution.

ORTHOLOGUES
Homologous genes in different species arising from a common ancestral gene at the time of speciation (BOX 2). Orthology does not guarantee common function, as function can change over time and vary in different evolutionary lineages.

PARALOGUES
Homologous genes in the same species arising by duplication (BOX 2).

Box 2 | **Orthology and paralogy**

Using the original definition of Walter Fitch[136], orthologues are genes in different species that arose from a single gene in the most recent common ancestor of those species — that is, by a process of speciation. Paralogues, on the other hand, are genes in the same species that arose from a single gene in an ancestral species by a process of duplication. In the phylogenetic tree depicted, an ancestral gene X gives rise to a gene X_h in modern humans. In the line leading to rodents, X undergoes a duplication, after which there is a speciation event so that two 'versions' are now present in each modern rodent species; X_{r1} and X_{r2} are paralogues in the rat, as are X_{m1} and X_{m2} in the mouse. Note that the human gene X_h therefore has two orthologues in each rodent species — it is a common misconception that orthologues must be unique. X_{r1} is orthologous to X_{m1} but not to X_{m2}, however similar they might be, because the latter did not arise from the same gene in the most recent common ancestor of rats and mice. If by chance the X_{m1} gene were lost during evolution (a not uncommon occurrence), X_{m2} might well be the most similar gene to X_{r1} in the mouse despite not being its orthologue, and if X_{r2} were lost as well there would be no way to tell that the remaining genes were not orthologues, except perhaps by information derived from additional species. Such eventualities, and others described in the text, can often complicate the assignment of orthology, and highlight the importance of detailed phylogenetic reconstructions with as many species as possible.



PHYLOGENETIC RECONSTRUCTION
The attempt to recreate the evolutionary history of a set of orthologues and/or paralogues (or, more generally, any set of measurable characters) and portray it in tree form. A number of different methods and algorithms are used for this purpose, and are the subject of much technical debate, but in the final analysis certainty as to ancestral forms is not possible.

PLEIOTROPY
The property of a gene or gene product by which it exhibits multiple phenotypic effects or possesses multiple functions.

REDUNDANCY
The property by which more than one gene or gene product is able to produce a given phenotype or function.

function require orthology, even within common pathways[25]). Protein functional shifts in the course of evolution are common, yet recognizing them from sequence data alone is not straightforward; experience from protein engineering shows that protein function is in some cases exquisitely sensitive to changes in just a few key amino acids. However, functional shifts in natural evolution are not so directed, taking place as they do against the background of the mutational 'MOLECULAR CLOCK,' which affords techniques for assessing the likelihood of changes in function having occurred.

Detecting functional shifts. Extensive sequence divergence between orthologues might raise suspicion of a functional shift, but simple pairwise comparisons are not generally useful because of the highly variable rates of evolution in different protein families[26]. However, phylogenetic reconstructions across a number of species can add an extra dimension of information, which is revealed by the topology of the tree and comparative histories of related genes. For example, a reconstruction of the CYP2A family of cytochrome P450 enzymes indicates that the rat liver isoform Cyp2a1 has diverged considerably from the human CYP2A6 and mouse Cyp2a4 (as well as the rat lung isoform Cyp2a3), occupying a lone long branch of the tree rooted outside the rest of the family (FIG. 2). This marked divergence correlates with a well-known functional shift, insofar as the rat enzyme metabolizes the substrate coumarin to an hepatotoxic epoxide, whereas the human and mouse enzymes act on the same substrate by way of a more innocuous hydroxylation[27].

Phylogenetic reconstructions need not be so dramatically divergent to be useful in the prediction of functional shifts. By examining ratios of NON-SYNONYMOUS to SYNONYMOUS nucleotide substitution rates one can estimate the nature and extent of evolutionary selection acting on a gene. Low ratios indicate a negative or purifying selection, typical of a gene whose function has remained stable over evolutionary time, whereas high ratios indicate positive or adaptive selection, quite possibly driven by a functional shift that proves advantageous[28] (but see BOX 3). As a result, one can annotate trees with measures of selection reflecting the likelihood of functional shifts having occurred, as has been done, for example, to demonstrate episodic adaptive evolution of primate lysozymes[29]; phylogenetic analysis software packages such as PAML perform the necessary calculations[30]. Of particular pharmacological interest, an analysis of the hormone leptin from a number of mammals found indications of accelerated adaptation in the primate lineage, indicative of the known functional shift whereby leptin acts directly as a satiety signal in rodents but not in humans[31].

For longer evolutionary timescales, synonymous mutations eventually become saturated and ratios are no longer useful. However, 'site-specific rate shifts', in which only non-synonymous substitutions are examined but in relation to each other within the same gene, offer a means of extending this form of analysis over a broader evolutionary span[32]. Like rate ratios, variations across phylogenies in the residues undergoing change can also indicate specific functional determinants, though such variation seems to be widespread and is not always associated with obvious functional shifts[33].

Selective sweeps. For shorter timescales, as within the human lineage, there might not have been sufficient non-synonymous substitutions to provide a statistically meaningful ratio. In this case, population genetics offers techniques based on the detection of 'selective sweeps' affecting selectively neutral polymorphisms even outside the coding region in question[34]. When strong selection arises for some variant, it can move toward fixation in a population so rapidly that it carries with it adjacent markers in what is called a 'hitchhiking' effect[35]. This produces a telltale signature consisting of a polymorphism 'trough' and related phenomena[36]. As an example, it was recently observed that chimpanzees have reduced levels of polymorphism in introns of their MHC class I genes, which could reflect a selective sweep 2–3 million years ago. Given the role of these genes in immune defense against intracellular infection, it was proposed

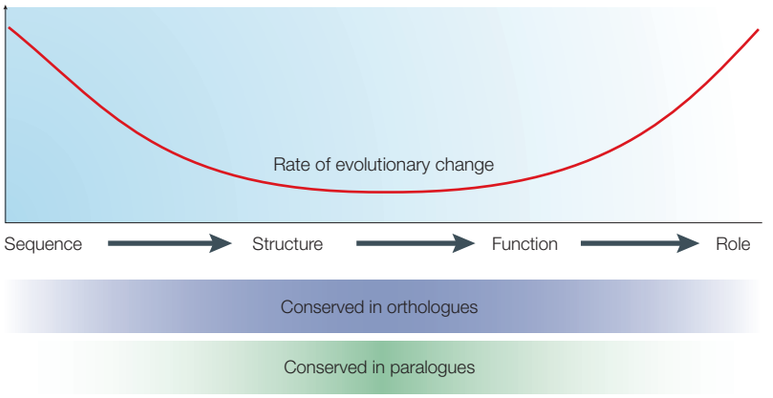


Figure 1 | **Relationship of orthology and paralogy to the rate and nature of evolutionary change.** As a rule, the structure of a protein is better conserved through time than its primary sequence, as is its biochemical function in comparison to its physiological role. A family of enzymes, for example, might possess a structural homology that is no longer detectable in sequence data, and might share a common reaction mechanism that is applied in many different cellular roles. Just as individual residues in sequence and structure can range from neutral to highly selected, there is often a gradation from well-conserved mechanism to somewhat less-conserved binding specificities to even more variable patterns of expression. Orthologues (genes in different species arising from a common ancestral gene during speciation) are usually better conserved than paralogues (genes in the same species arising by duplication), and in that difference there might be useful information, recoverable by phylogenomic methods. (As is common practice, the distinction between function and role will be blurred in the remainder of this paper, but should be borne in mind.)

that this paucity of variation might have resulted from a pandemic infection by human immunodeficiency virus-1 (HIV-1), which would help to explain the resistance of modern chimpanzees to the progression of HIV infections to full-blown AIDS[37].

The genetic signals produced by selection can be confounded by demographic effects, including rapid population growth known to have occurred in the human lineage, as well as more complex forms of selection, but new techniques promise to allow these effects to be better distinguished[34]. The detection of selection signatures in the human genome is presently benefiting from the rapid accumulation of polymorphism data; initial analyses have putatively identified more than a hundred human genes as candidates for selection, including a number of disease-related genes, such as the *cystic fibrosis transmembrane conductance regulator* (*CFTR*) gene and the *peroxisome proliferator activated receptor-γ* gene (*PPAR-γ*), a drug target for type 2 diabetes[38].

So, there is an armamentarium of techniques now available for assessing the likelihood of functional shifts at various evolutionary distances. These methods can also be combined to good effect, as in recent work with a transcription factor gene, *FOXP2*, which in several cases has been found to be mutated in severe speech and language disorders[39]. Aside from two polyglutamine tracts, FOXP2 is among the 5% of proteins that are most-conserved between rodents and humans; of only three amino acid changes since the mouse–human divergence, two have occurred very recently, since humans split from other primates. Not only did non-synonymous-to-synonymous codon ratios provide evidence of positive selection, but also the pattern of neutral alleles at this site indicated a recent selective sweep, raising the intriguing possibility that FOXP2 has evolved rapidly in the human lineage as part of the development of a capacity for language[39]. This hypothesis is especially interesting given a proposed connection between the evolution of human language capabilities and schizophrenia[40].

Targets and disease. These examples serve to highlight the fact that phylogenetics combined with complete genomes will be especially powerful in the analysis of known differences in phenotypes and disease susceptibilities in various species, such as those between humans and chimpanzees[41]. Such differences often govern the choice of disease model organisms, but phylogenomics opens up new possibilities for correlating those phenotypes with the evolutionary behaviour of genes, and could usher in what amounts to interspecies disease genetics.

Another challenge and opportunity in this arena will be the adaptation of these techniques to comparisons of regulatory regions, which do not afford any straightforward notion of synonymous versus non-synonymous change[42], but which might benefit from phylogenetic footprinting techniques, as well as correlation with gene expression data from platform technologies (BOX 4). In fact, even synonymous codon changes can affect gene expression through, for example, codon bias, RNA secondary structure or splicing signals, and thereby show evidence of selection in specialized metrics[43]. A recent study of 35 G-protein-coupled receptors (GPCRs) implicated in psychiatric and neurological disorders detected such selection in the dopamine D_2 receptor, and demonstrated marked functional effects of supposedly silent variants[44]. (Note that purifying selection acting on synonymous codon changes will paradoxically increase non-synonymous-to-synonymous ratios, as has been demonstrated in the *BRCA1* gene[45].)

Target paralogy

As important as orthology is in assessing drug targets, paralogy might be even more so. Many genes of pharmacological interest occur in large families for which phylogenetic analyses have provided a classification framework and key insights, especially the nuclear receptors and GPCRs. Even beyond these cases of extensive paralogy, there is evidence that vertebrate genomes have undergone, by various and controversial accounts, one, two or more duplications in their entirety, thereby producing a general background level of paralogy[46]. Newer evidence indicates the importance of very recent expansions by tandem or segmental duplications of >90% similarity that could account for 5% of the euchromatic genome[47–49]. Indeed, there have lately been instances in which adjacent or nearby duplications of genes have provided possible alternatives to drug targets already under study — for example, vanilloid receptor ion channels[50] and nicotinic acid receptors[51]. Moreover, certain therapeutic areas might call for multifunctional or 'broad spectrum' compounds that affect two or more paralogues. For example, in the treatment of cancer and related diseases it might be desirable to intervene at more

BLAST
Basic Local Alignment Search Tool, the most widely used bioinformatics algorithm[130]. It efficiently searches sequence databases for the entries most similar to a query sequence. Recent, more advanced, versions and related tools are specially adapted to finding distant homologues, for which sequence similarity is not obvious but typically some structural similarity is retained.

INCONGRUENT EVOLUTION
Apparent topological differences in the phylogenetic trees of individual genes relative to that of the species, or of individual domains or regions within genes relative to each other. This can arise from phenomena such as domain shuffling or horizontal transmission of genes between species.

CONCERTED EVOLUTION
Greater-than-expected similarity seen in members of gene families within a species relative to that seen between species. This can arise from phenomena related to physical mechanisms of replication and recombination that tend to maintain uniformity between (often tandem) copies.

Box 3 | **Co-evolution and covariation**

"Now, here, you see", said the Red Queen to Alice in Lewis Carroll's *Through the Looking Glass*, "it takes all the running you can do, to keep in the same place". This passage furnished the name for a principle of evolutionary biology called the Red Queen effect, which states that species in competition must each continuously evolve just to maintain their respective fitness, much less advance it[137]. In relationships such as those between pathogens and their hosts, this can produce signs of adaptive selection that indicate a functional shift, when in fact the pathogen is merely evolving to preserve its virulence in the face of selection due to the similarly evolving immune system of the host[138]. So, a gene might need to change in order to remain the same, in terms of its function in a wider context.

The Red Queen effect demonstrates the concept of co-evolution[139], which, however, is not limited to competitive situations but extends to cases of mutualism between species and even to a complementary interplay of gene products within a species. Evidence of co-evolution can be found in congruence (topological similarity) between phylogenetic trees, either of species or of individual genes that are evolving in concert because of interactions, such as those between receptors and ligands. Even within single gene products, co-variation between residues might point to a physical interaction, as seen most obviously in compensatory mutations that preserve base pairing in RNA secondary structure; on the other hand, incongruence of trees for different domains of the same protein could reflect a complex evolutionary history, for instance, one involving domain shuffling. Just as patterns of evolutionary conservation indicate important functional features at many levels, patterns of co-variation connect related features and can add value to a pharmacophylogenomic approach.

than one point in a pathway or process, as with dual inhibitors of topoisomerases I and II[52], the receptor tyrosine kinases epidermal growth factor receptor (EGFR) and ERBB2 (also known as HER2/neu)[53], farnesyl- and geranylgeranyl-protein transferases[54], and the two subtypes of 5α-reductase in the prostate (SRD5A1 and SRD5A2)[55]. The same theme occurs in antibiotics targeting multiple paralogous components of fatty acid biosynthesis in bacteria, in which the differential distribution of these paralogues across bacterial phylogeny is an important extra consideration[56]. Psychiatric disorders involving a constellation of paralogous monoamine receptor subtypes might require tuning of drug 'receptor profiles' to address complex symptomology simultaneously with side effects[57]. So, cataloguing and thoroughly understanding paralogy is important for new target identification and functional characterization, as well as for delineating selectivity challenges in lead optimization and opportunities for multifunctional intervention.

SYNTENY
The property of genes of being found on the same chromosome. The ordering of orthologues on chromosomes is often conserved between related species over extended segments, indicating a common ancestry of those segments; this phenomenon is referred to as conservation of synteny. (To describe the orthologues or regions of the different species as being syntenic to each other is a common misuse of the term.)

Pleiotropy and redundancy. By analogy with orthology, paralogy is best understood when considered in a full phylogenetic context that accounts for intermediate states, possible functional shifts, incongruence and so on. Beyond these factors, paralogy bears on issues of pleiotropy and redundancy that can profoundly affect the suitability of targets (FIG. 3). In genetic terms, pleiotropy occurs when a gene affects more than one trait, and redundancy when a trait is affected by more than one gene. At the level of gene products, there are many senses in which a protein can have more than one function[58], ranging from multifunctionality associated with multiple domains, to relaxed substrate or ligand specificities, to the accretion of unrelated physiological roles by what have been called 'moonlighting proteins'[59,60]. The latter, which can arise by GENE SHARING or RECRUITMENT, include proteins such as lens crystallins that often serve radically different functions in some other tissue[61]; others, such as 4α-carbinolamine dehydratase/DCoH (DCOHM), whose function depends on the cellular compartment in which it finds itself (enzymatic activity in the cytoplasm, transcriptional control in the nucleus)[62]; and yet others, such as phosphoglucose isomerase (GPI), a glycolytic enzyme that also serves in several different extracellular roles, for instance as the cytokines neuroleukin and autocrine motility factor[63]. Such 'reuse' of proteins is another reason that assessing function on the basis of a single top BLAST hit can miss important information[64]. These are perhaps extreme cases, but even unremarkable monofunctional enzymes can assume disparate physiological roles in different cell types and developmental stages, in varying environments of pathway utilization, substrate and cofactor availability, pH and redox conditions, and so forth. Signalling molecules can be expected to be even more polymorphous in their roles.

The potential for such pleiotropy occurring in drug targets is of obvious importance, as when the need arises to dissect physiological effects such as the genotropic and non-genotropic actions of the oestrogen receptor[65]. Pleiotropy is related to paralogy in an evolutionary sense, insofar as the former affords multiple functions from a single gene locus, whereas the latter affords multiple functions by divergence following a gene duplication. In fact, recent evolutionary theory suggests that pleiotropy can actually precede paralogy as a rule; that is, genes can first acquire multiple functions before being duplicated and then specializing, in a process called subfunctionalization[66]; common mechanisms might include the divergence of multiple related enzymatic activities from ancestral enzymes with lower substrate specificity, and duplication and divergence of transcription factors to control different subsets of genes originally controlled as a group. In this view, alternative transcription, such as that embodied in splice variants, can be seen as a kind of intermediate between paralogy and pleiotropy; indeed, as a 'paralogy in place' that considerably increases the effective size of the genome[67] (FIG. 3).

The converse of pleiotropy is redundancy, which describes a situation in which more than one gene product can serve or contribute to the same function.

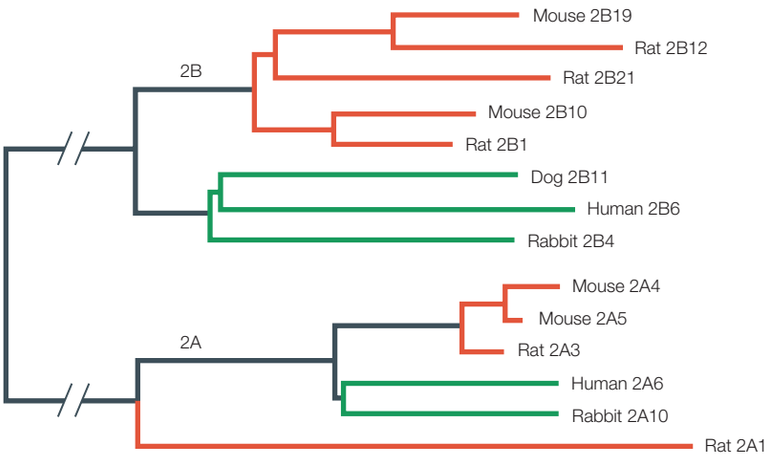


Figure 2 | **Phylogenetic reconstruction of the CYP2 family of cytochrome P450s.** The tree was constructed from selected CYP2A and B isoforms by a simple neighbour-joining procedure. The CYP2B subfamily shows a characteristic clustering of rodent orthologues and paralogues, well separated from other mammals. The CYP2A subfamily, however, isolates the rat 2A1 isoform on its own long branch, which accords well with a known functional shift in the metabolism by 2A1 of the substrate coumarin (see text).

The relationship to paralogy is direct, when paralogues provide either a total or partial redundancy of function; it is perhaps most graphic in the many observed cases of robustness to gene knockouts or null mutations that, notwithstanding the need to consider the full range of phenotypic effects, environmental influences, responses to stress, and so on, reveal at least the potential for overlapping gene function. (Notably, it is thought that pleiotropy might contribute to preserving redundancy in gene duplications that might otherwise diverge rapidly[68].) Results from systematic yeast gene ablation studies confirm a correlation between functional redundancy and duplicated genes, and, as might be expected, the correlation increases according to sequence similarity[69]. Recent demonstrations of redundancy that are of particular pharmacological interest include apparent partial redundancies of dopamine transporters for serotonin transporters in adjacent neurons[70], PPAR-δ for PPAR-α in skeletal muscle in which the former is highly expressed[71], caspase-9 for caspase-2 in apoptosis[72], COX-1 for COX-2 and vice versa[73], the nuclear receptor PXR for FXR in bile acid signalling[74], and butyrylcholinesterase for acetylcholinesterase in central cholinergic pathways[75].

Crosstalk and heteromery. Such examples alone provide an argument for a careful assessment of the 'paralogy space' of any drug target, but phenomena such as CROSSTALK and HETEROMERY, which often involve paralogy, further underline this need (FIG. 3). Crosstalk can be seen as a combination of pleiotropy and redundancy, an archetypal example being the action of cytokines such as interleukins on multiple immune cell types, each of which is in turn affected by multiple cytokines in an "interdigitating, redundant network [that has] crucial significance in the development of therapeutic strategies…in cytokine-mediated inflammatory processes"[76]. Intracellular and paracrine crosstalk, on the other hand, might be largely 'controlled' in nature by compartmentalization in time and space[77], but, because of tendencies toward compensatory behaviours in response to perturbation by disease or intervention, the potential must still be carefully considered. The recent literature is replete with examples of signalling crosstalk[78–81].

The formation of heteromers, as a rule between paralogues, is increasingly recognized as a key aspect of function in a number of proteins of pharmacological

MOLECULAR CLOCK
The hypothesis that, except for the effects of functional constraints on gene products, sequence substitutions occur at a constant rate on an evolutionary timescale. It is closely tied to the 'neutral theory' of evolution, which asserts that most such mutations are selectively neutral and driven only by random drift. Although subject to certain caveats and continuing debate, the notion of the molecular clock has proven to be an important and useful tool in many contexts[131].

NON-SYNONYMOUS SUBSTITUTION
A nucleotide substitution that results in an amino acid change.

SYNONYMOUS SUBSTITUTION
A 'silent' nucleotide substitution, often in the third codon position, that does not result in an amino acid change.

GENE SHARING (RECRUITMENT)
An adaptation of a gene to serve an additional unrelated function, generally in a different tissue and presumably by the incorporation of alternative regulatory elements at the same locus. It is one proposed mechanism for establishing pleiotropy.

Box 4 | **Expression and interaction**

Although much can be accomplished by means of pharmacophylogenomic analysis of genomes, far greater strides can be expected through integration with genomics and proteomics platform data. Such observables as gene expression patterns and protein interactions are, after all, evolving phenotypic characters in which selective pressures can be detected, just as in the sequence that encodes them. A notable recent illustration is the comparison of rates of change in overall patterns of gene expression in primates. This study demonstrated that humans are more similar to chimpanzees than either are to macaques in liver and blood cell gene expression patterns, conforming well to the known species tree; however, in the brain there is evidence of a rapid acceleration of change unique to the human lineage[140]. Genes that have evolved to possess similar expression patterns can be expected to have acquired common regulatory elements, and indeed these have been shown to be accessible to footprinting[141].

Differences in tissue distributions of orthologues are *prima facie* evidence of functional shifts. Expression patterns of paralogues can indicate whether a functional redundancy is likely, or whether the gene products are segregated in space or time so as to circumvent redundancy; for example, knockout of the Myf5 transcription factor in the mouse results in a rib cage defect, despite the fact that this defect can be rescued by placing a paralogue, myogenin, under control of the regulatory region of Myf5 (REF. 142). As noted in the text, such segregation of expression can also control the potential for crosstalk and heteromery.

Interaction networks can be probed both phylogenomically and by platform technologies, and the combination can provide insights into pathway evolution, compensatory mechanisms and so forth. Just as evolutionarily conserved regulatory elements can be discovered by footprinting upstream regions of co-expressed genes, consensus sequences of peptide recognition elements can be determined by phage display, then used to predict whole-genome interaction maps that can be tested by yeast two-hybrid methods[143]. Both phylogenomic and platform technology data can be beset by distinctive forms of noise and uncertainty — all the more reason to exploit the mutual information they offer, and in particular the organizing framework inherent in an evolutionary view of whole genomes.

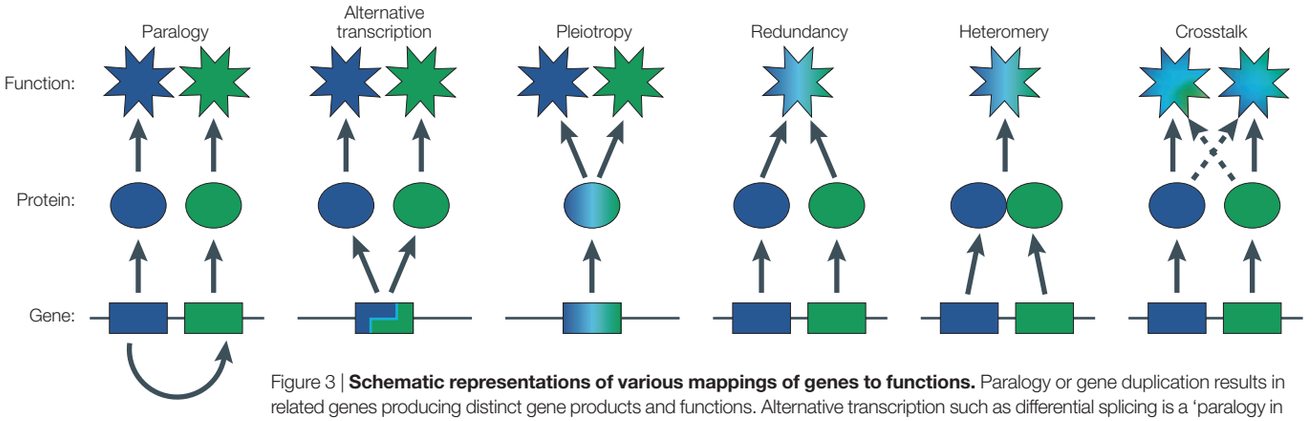


Figure 3 | **Schematic representations of various mappings of genes to functions.** Paralogy or gene duplication results in related genes producing distinct gene products and functions. Alternative transcription such as differential splicing is a 'paralogy in place' that also produces distinct (but related) gene products and functions. Pleiotropy manifests when a single gene product has more than one function. Conversely, redundancy exists when more than one gene product possesses or contributes to the same function. In heteromery, distinct (but often paralogous) gene products associate to serve a single function. Crosstalk is a combination of pleiotropy and redundancy that might or might not involve paralogy.

interest, including at least three major classes of drug targets: the GPCRs, beginning with the $GABA_B$ (γ-aminobutyric acid B) receptors[82] but now thought to extend to other cases and even to larger oligomers[83]; the nuclear receptors, which form not only homodimers but heterodimers with retinoid X receptors and in a number of other combinations[18]; and many types of ion channels[50,84,85]. Note that homomery can be mediated by mechanisms such as symmetric oligomerization domains (for instance, in DNA-binding proteins that recognize palindromic sequences) and DOMAIN SWAPPING[86], indicating a natural route for the evolution of heteromery through gene duplications that maintain these mechanisms after divergence.

So there are a number of different mechanisms that serve to lend combinatoric diversity to gene products at many levels: at the genome level, in multidomain proteins; at the transcriptional level, in alternative splicing, for example; at the post-transcriptional and post-translational levels in the many forms of modification that can occur; and at the physiological level in various types of interaction, as embodied in heteromery and crosstalk (FIG. 3). As with orthology, pharmacophylogenomics can offer insights into these complexities, by tracking paralogy and selective pressures across species to indicate where potentials might have come and gone for combinatoric interactions. Such efforts will be most valuable when undertaken in close coordination with expression studies and other genomic platform technologies (BOX 4).

Consequences of pleiotropy. Phylogenomics, for instance, could aid in recognizing pleiotropy, which theory predicts will result in lower levels of variation and lower substitution rates in a gene[87]. Intuitively, pleiotropy creates more constraints on a protein, attributable to its more diverse function involving more functional residues, such that the degree or location of purifying selection might be informative[26]. There is evidence that the remarkable conservation of complex developmental patterns across phylogeny is primarily due to pleiotropy resulting in such selection[88]. Highly conserved proteins in a number of species tend to be larger, with a wider size distribution, than less conserved proteins[89], an observation consistent with a view of large multifunctional proteins evolving more slowly.

As previously noted, expression of a gene in multiple tissues can be associated with pleiotropy, and in fact there is a marked negative correlation in mammals between breadth of expression and evolutionary rates[90]. Although it has been suggested that pleiotropy is most likely to be observed in the middle ground between narrowly and ubiquitously expressed genes[58] (FIG. 4), in fact, housekeeping genes that must maintain the same function in many different tissue types, with varying interactions and physical/chemical conditions, might thus experience selective pressures that are indistinguishable from those associated with true pleiotropy[91].

Functional shifts, pleiotropy, and redundancy have the potential to constitute both good news and bad news for drug discovery. A functional shift in a target might be bad news when it means that an animal model is unavailable or misleading, but it can also be good news if it indicates that a troubling animal toxicity is irrelevant to humans[27]. Similarly, pleiotropy can evoke unintended drug side effects, but might also create opportunities to pursue multiple indications[92,93]. Redundancy would be a liability if it meant that a disease process was resistant to intervention, yet might be offset if timely recognition of paralogous functional overlaps allowed for lead optimization toward the necessary compound multifunctionality; it could even indicate possibilities for highly selective intervention in complex disorders, particularly when the functional overlaps are partial[57].

Pathways and networks

Concerns about crosstalk and heteromery raise the question of whether pathways and interaction networks are also amenable to pharmacophylogenomic

CROSSTALK
The interaction of elements of distinct signalling or regulatory pathways such that an input to one pathway has some effect on the output of the other.

HETEROMERY
The physical association of distinct but often similar macromolecules, as when a pair of protein subunits combine to form a heterodimer. A combination of identical subunits is called homomery.

DOMAIN SWAPPING
The symmetric exchange of portions of polypeptides (ranging up to entire domains), by partial unfolding, between subunits of a multimeric (usually dimeric) assemblage, such that the exchanged portions occupy positions in their counterpart subunits analogous to those they would assume in the monomers.

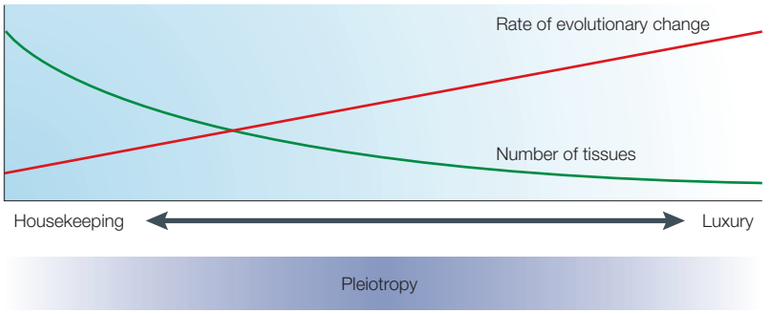


Figure 4 | **Phylogenomics and expression patterns.** "Pleiotropy, the condition in which a single gene affects multiple traits, may well be the rule rather than the exception in higher organisms. In the past, geneticists have usually preferred to focus on genes with a single well-defined function… Most 'housekeeping' genes (ubiquitously expressed), and many 'luxury' genes (expressed in only one tissue) fall into this category, but most genes in animal genomes are expressed in some but not all tissues, and probably act differently in each situation"[58]. There seems to be an inverse correlation between breadth of expression and rates of evolution of proteins[90]. As a rule, it might be desirable to seek drug targets that avoid both pleiotropy and ubiquity.

approaches in their own rights. Indeed, early theories of pathway evolution suggested that paralogy might have played a key role, with metabolic pathways in particular arising by way of gene duplication and divergence of enzymes whose substrate recognition sites were similar by virtue of binding successive metabolites in a reaction sequence[94]. There are intimations of such a mechanism, for example, in apparent paralogy (at least at the level of structural homology) seen within amino acid synthetic pathways such as those for methionine[95], tryptophan[96], and histidine[97], as well as in the aforementioned bacterial fatty acid synthetic pathways[56]. More generally, a genome-wide study has shown that homologous enzymes statistically tend to be situated close to each other in metabolic networks[98]. On the other hand, another phylogenomic analysis indicates that this evolutionary motif is less prevalent than recruitment of enzymes from parallel, related pathways[99], in which case a generalized notion of 'pathway paralogy' might prove fruitful. Recent work has begun to establish a theoretical framework for the extension of phylogenetic analysis to metabolic networks[100].

Compensation and interaction. As noted in previous text, paralogy giving rise to functional redundancy can account for robustness to gene ablation; so too can compensatory changes in pathways (with or without paralogy), for example, by differential regulation of related pathways or other components of the same pathway. Large metabolic networks can compensate in this way to maintain an optimal flux of metabolites, and developmental mechanisms in model organisms also seem to be 'buffered' against mutation[101]. Such compensation can be at a molecular, physiological or even structural level; in mouse skeletal muscle, knockout of myoglobin is compensated by expression-related changes in angiogenesis, nitric oxide metabolism and vasomotor regulation[102], whereas knockout of creatine kinase results in redirection of metabolic pathways, for instance, through upregulation of myoglobin and genes related to ATP production, and even changes in ultrastructure[103]. Parallelism of pathways, such as that seen in apoptosis, might predispose to such compensatory effects, which must therefore be considered in therapeutic intervention[104] and which also highlight the potential contribution of crosstalk[105].

Phylogenomic approaches to pathways and networks demonstrate how evolutionary inferences can be made without consideration of sequence homology. By examining a number of different genomes for recurring gene fusions, it is possible to discover many sets of gene products that participate in the same pathway or that otherwise interact in the cell. For example, the bifunctional human enzyme δ-1-pyrroline-5-carboxylate synthetase comprises a fusion of domains that in *Escherichia coli* exist as separate gene products — γ-glutamyl phosphate reductase and glutamate-5-kinase, which catalyse the first two steps in proline synthesis. Once again, the bacterial fatty acid synthases offer another instance, in that they form a large multifunctional polypeptide in eukaryotes[56]. Such tendencies toward fusions into what have been dubbed 'Rosetta Stone proteins' thereby allow for an *in silico* form of pathway or interaction analysis[106]. More generally, common function[107] or subcellular location[108] of proteins can be inferred by simply counting the presence or location of genes across many genomes in a technique called phylogenetic profiling that gains in statistical power with each new genome examined.

Co-evolution. For proteins that are both interacting and evolving, such as receptors and peptide ligands or enzymes with macromolecular substrates, one can expect to see evidence of co-evolution (BOX 3), as has been shown, for example, between the chemokines and their GPCRs[109], and between a variety of other ligand–receptor pairs[110]. Such co-evolution is reflected in similarities in the detailed topologies of their phylogenetic trees, which with appropriate metrics can allow for the *de novo* prediction of interactions[111,112]. It follows that pathways and networks as a whole must co-evolve in the complex interactions of their components, interactions that can be direct through contact or indirect through the influence of metabolites. For example, there is evidence for co-evolution in the close congruence of phylogenetic trees of elements of bacterial two-component signal transduction pathways[113]. Note that some of the same interactions that leave their traces in the phylogenetic record of co-evolution are probably at play in 'real time' in the compensatory responses described previously.

As has been noted, pleiotropy can be associated with slower evolution. One way that pleiotropy could manifest itself is in greater numbers of interactions with other proteins, and indeed, the topology of yeast interaction networks indicates an inverse relationship between degree of interaction and evolutionary rates (FIG. 5). In this organism, proteins with greater numbers of interactions have evolved more slowly as a rule; moreover, interacting proteins evolve at similar rates, as would be predicted from co-evolution[114]. Although

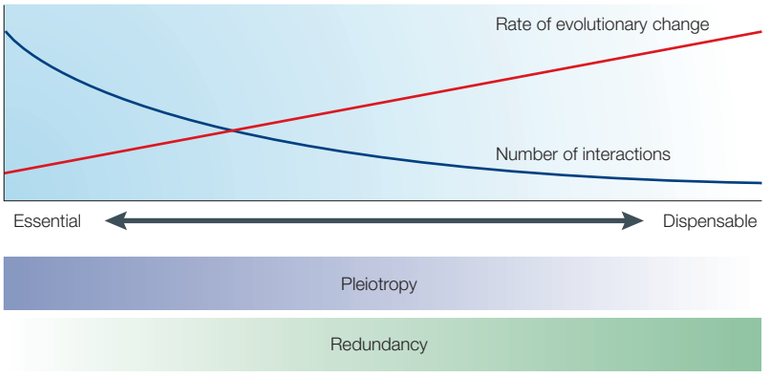


Figure 5 | **Phylogenomics and interaction patterns.** Various threads of evidence indicate that pleiotropic genes and those whose gene products have the greatest numbers of interactions evolve relatively slowly (see text). Highly pleiotropic genes or those at the 'hubs' of interaction networks can be expected to be essential as a rule, whereas duplicated and therefore redundant genes are classically assumed to be dispensable and released from selective pressure, allowing for rapid change. Combining these themes as shown is purely a schematic representation of trends that are probably much more complex, noisy, and higher-dimensional in nature, but it nevertheless underscores the need to evaluate potential drug targets in phylogenomic terms.

it has been suggested that the former effect might be limited only to the most highly interacting 'hubs' of interaction networks[115], a more recent study with larger datasets tends to confirm the generality of the observation[116]. It is interesting to note that highly interacting proteins tend not to interact with each other, which could serve to damp crosstalk; this property seems to be inherent in the topology of interaction maps in nature, which, in common with metabolic and regulatory networks, tend to assume the form of so-called scale-free networks that are inherently robust to random node removal because most nodes make few connections[117,118].

The complementary inference would be that redundancy should lead to faster change. This is certainly compatible with the venerable notion that gene duplication allows for divergence through release of one copy from stabilizing selection[119], and, to the extent that redundant genes are dispensable, it has long been predicted that they would evolve faster than essential genes[120]. In bacteria[121] and in yeast[122], gene-ablation studies indicate that dispensability of genes does indeed correlate with rate of evolution (FIG. 5), though the effect in yeast might be small[123,124]. Although the evidence in rodents points to an inverse relationship between evolutionary rates and severity of knockout phenotypes, it seems that this can be largely accounted for by an over-representation of immune-related genes that might be under co-evolutionary selection[125] (BOX 3). As the dispensability of yeast genes does correlate with their degree of duplication, as previously noted[69], one might expect that evolutionary rates would therefore also correlate directly with extent of paralogy. It does seem to be the case that larger gene families in yeast support higher amino acid substitution rates, perhaps due to a 'buffering' of such mutations by paralogues, but this is not seen in selected multicellular organisms[126]. Such differences between single-cell and multicellular organisms in the relationships among dispensability, paralogy and evolutionary rates could be the result of certain mathematical effects of population size[68], but a more intriguing possibility is that tissue compartmentalization of gene expression in more complex organisms effectively segregates paralogues that might otherwise create redundancy[126] (BOX 4).

Target evolution. In general, potential phylogenomic indicators of phenomena such as pleiotropy and redundancy still require validation, especially in mammals, but at least raise the possibility that such properties of

Box 5 | **Developability and druggability**

The developability of compounds — that is, their predicted *in vivo* behaviour in terms of absorption, distribution through the body, metabolism, probable toxicities and so forth, independent of their mechanism of action — is increasingly being addressed at earlier stages of discovery. The 'drug-like' character of compounds has been assessed by means ranging from the intuition and experience of chemists to sophisticated computational methods; the latter include machine learning algorithms that generalize from various chemical descriptors of known 'good' drugs[144] and expert systems that adopt a rule-based approach using easily measured properties[145]. The most widely used set of metrics has been the Lipinski 'rule-of-five' property filters for absorption, which establish windows of 'drug-likeness' within ranges of molecular mass, lipophilicity and hydrogen-bonding potential[146]; lately, these have been extended and refined with parameters such as number of rotatable bonds[147].

To date there have been few such general heuristics for predicting the 'target-likeness' or inherent tractability of targets to intervention, independent of their disease relevance. The suitability of targets is largely assessed through the intuition and experience of biologists and on the basis of membership in classes with proven track records as drug targets, which in turn often relates to such factors as subcellular localization. Beyond this, analyses are mostly *ad hoc*, and not based on general principles *à la* Lipinski. To be sure, there are important differences between compounds and targets in assessing tractability. For one, compounds can be designed, whereas targets are a given. Also, the potential number of compounds is staggering compared with the size of the genome; drug-like compound scaffolds and basic protein folds can both be restricted sets, but the diversity around them is of a fundamentally different character.

Even so, recent studies have begun to consider the set of targets comprising the 'druggable genome' in aggregate terms, such as their drug-binding domain content[148]. The evolutionary and systems view provided by pharmacophylogenomics suggests a number of possible target 'property filters,' for example, the likelihood of functional shift, degree and nature of paralogy, and factors reflecting pleiotropy such as size, breadth of expression, interaction potential, and evolutionary rates, all of which could soon allow for systematic guidelines regarding the druggability of targets.

targets could be analysed much like developability properties of compounds (BOX 5). In any case, a pharmacophylogenomic approach in assessing targets can already add considerable value through a better understanding of where, in evolutionary terms, a target has been and even where, in selective terms, it is headed. Viewing genes as potentially being in the midst of change, can provide new insights, for instance, in the interpretation of structure[127], function[128] and polymorphism[129]. Pharmacogenetics is teaching us that targets cannot be regarded as homogeneous entities, while systems and pathway biology are demonstrating that they cannot be considered in isolation. Pharmacophylogenomics will show in closely related ways that targets should not be considered as static, but rather in the context of a still-unfolding biological history that can inform drug discovery in important ways.

1. Eisen, J. A., Kaiser, D. & Myers, R. M. Gastrogenomic delights: a moveable feast. *Nature Med.* **3**, 1076 (1997).
2. Eisen, J. A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* **8**, 163–167 (1998).
The first full description of the phylogenomic approach.
3. Casari, G., Sander, C. & Valencia, A. A method to predict functional residues in proteins. *Nature Struct. Biol.* **2**, 171–178 (1995).
4. Mirney, L. A. & Gelfand, M. S. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.* **321**, 7–20 (2002).
5. Eisen, J. A. & Wu, M. Phylogenetic analysis and gene functional predictions: phylogenomics in action. *Theor. Popul. Biol.* **61**, 481–487 (2002).
6. Hochachka, P. W. & Monge, C. Evolution of human hypoxia tolerance physiology. *Adv. Exp. Med. Biol.* **475**, 25–43 (2000).
7. Barclay, A. N. Ig-like domains: evolution from simple interaction molecules to sophisticated antigen recognition. *Proc. Natl Acad. Sci. USA* **96**, 14672–14674 (1999).
8. Jaaro, H., Beck, G., Conticello, S. G. & Fainzilber, M. Evolving better brains: a need for neurotrophins? *Trends Neurosci.* **24**, 79–85 (2001).
9. Wilson, D. R. Evolutionary epidemiology and manic depression. *Br. J. Med. Psychol.* **71**, 375–395 (1998).
10. Gammelgaard, A. Evolutionary biology and the concept of disease. *Med. Health Care Philos.* **3**, 109–116 (2000).
11. Tatusov, R. L. *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22–28 (2001).
12. Gilks, W. R. *et al.* Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics* **18**, 1641–1649 (2002).
13. Jones, D. T. & Swindells, M. B. Getting the most from PSI-BLAST. *Trends Biochem. Sci.* **27**, 161–164 (2002).
14. George, R. A. & Heringa, J. Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins* **48**, 672–681 (2002).
15. Holm, L. & Sander, C. Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.* **27**, 244–247 (1999).
16. Todd, A. E., Orengo, C. A. & Thornton, J. M. Plasticity of enzyme active sites. *Trends Biochem. Sci.* **27**, 419–426 (2002).
17. Hou, J., Sims, G. E., Zhang, C. & Kim, S. H. A global representation of the protein fold space. *Proc. Natl Acad. Sci. USA* **100**, 2386–2390 (2003).
18. Thornton, J. W. & DeSalle, R. A new method to localize and test the significance of incongruence: detecting domain shuffling in the nuclear receptor superfamily. *Syst. Biol.* **49**, 183–201 (2000).
19. Koski, L. B. & Golding, G. B. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* **52**, 540–542 (2001).
20. Liao, D. Concerted evolution: molecular mechanism and biological implications. *Am. J. Hum. Genet.* **64**, 24–30 (1999).
21. Amadou, C. Evolution of the MHC class I region: the framework hypothesis. *Immunogenetics* **49**, 362–367 (1999).
22. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. in *Molecular Systematics* (eds Hillis, D. M., Moritz, C. & Mable, B. K.) 407–514 (Sinauer Associates, Sunderland, 1996).
23. Storm, C. E. & Sonnhammer, E. L. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* **18**, 92–99 (2002).
24. Zmasek, C. M. & Eddy, S. R. Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* **3**, 14 (2002).
25. Koonin, E. V., Mushegian, A. R. & Bork, P. Non-orthologous gene displacement. *Trends Genet.* **12**, 334–336 (1996).
26. Brookfield, J. F. What determines the rate of sequence evolution? *Curr. Biol.* **10**, R410–R411 (2000).
27. Lake, B. G. Coumarin metabolism, toxicity and carcinogenicity: relevance for human risk assessment. *Food Chem. Toxicol.* **37**, 423–453 (1999).
28. Li, W.-H. *Molecular Evolution* (Sinauer Associates, Sunderland, 1997).
29. Messier, W. & Stewart, C. B. Episodic adaptive evolution of primate lysozymes. *Nature* **385**, 151–154 (1997).
30. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
31. Benner, S. A. *et al.* Functional inferences from reconstructed evolutionary biology involving rectified databases — an evolutionarily grounded approach to functional genomics. *Res. Microbiol.* **151**, 97–106 (2000).
32. Gaucher, E. A. *et al.* Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.* **27**, 315–321 (2002).
33. Lopez, P., Casane, D. & Philippe, H. Heterotachy, an important process in protein evolution. *Mol. Biol. Evol.* **19**, 1–7 (2002).
34. Bamshad, M. & Wooding, S. P. Signatures of natural selection in the human genome. *Nature Rev. Genet.* **4**, 99–111 (2003).
An extensive and accessible review of evidence for selection in the human genome.
35. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res. Camb.* **23**, 23–35 (1974).
36. Przeworski, M. The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179–1189 (2002).
37. de Groot, N. G. *et al.* Evidence for an ancient selective sweep in the MHC class I gene repertoire of chimpanzees. *Proc. Natl Acad. Sci. USA* **99**, 11748–11753 (2002).
38. Akey, J. M. *et al.* Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**, 1805–1814 (2002).
39. Enard, W. *et al.* Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* **418**, 869–872 (2002).
Demonstrates the use of measures of selection to suggest a recent functional shift in a gene also associated with an inherited disorder.
40. DeLisi, L. E. Speech disorder in schizophrenia: review of the literature and exploration of its relation to the uniquely human capacity for language. *Schizophr. Bull.* **27**, 481–496 (2001).
41. Olson, M. V. & Varki, A. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nature Rev. Genet.* **4**, 20–28 (2003).
Makes a strong case for the utility of primate genomes in the study of human disease.
42. Rockman, M. V. & Wray, G. A. Abundant raw material for *cis*-regulatory evolution in humans. *Mol. Biol. Evol.* **19**, 1991–2004 (2002).
43. Akashi, H. Gene expression and molecular evolution. *Curr. Opin. Genet. Dev.* **11**, 660–666 (2001).
44. Duan, J. *et al.* Synonymous mutations in the human dopamine receptor D_2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.* **12**, 205–216 (2003).
45. Hurst, L. D. & Pal, C. Evidence for purifying selection acting on silent sites in BRCA1. *Trends Genet.* **17**, 62–65 (2001).
46. Durand, D. Vertebrate evolution: doubling and shuffling with a full deck. *Trends Genet.* **19**, 2–5 (2003).
47. Samonte, R. V. & Eichler, E. E. Segmental duplications and the evolution of the primate genome. *Nature Rev. Genet.* **3**, 65–72 (2002).
48. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
49. Friedman, R. & Hughes, A. L. The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol. Biol. Evol.* **20**, 154–161 (2003).
50. Smith G. D. *et al.* TRPV3 is a temperature-sensitive vanilloid receptor-like protein. *Nature* **418**, 186–190 (2002).
51. Wise, A. *et al.* Molecular identification of high and low affinity receptors for nicotinic acid. *J. Biol. Chem.* **278**, 9869–9874 (2003).
52. Vicker, N. *et al.* Novel angular benzophenazines: dual topoisomerase I and topoisomerase II inhibitors as potential anticancer agents. *J. Med. Chem.* **45**, 721–739 (2002).
53. Xia, W. *et al.* Anti-tumor activity of GW572016: a dual tyrosine kinase inhibitor blocks EGF activation of EGFR/erbB2 and downstream Erk1/2 and AKT pathways. *Oncogene* **21**, 6255–6263 (2002).
54. Lobell, R. B. *et al.* Evaluation of farnesyl:protein transferase and geranylgeranyl:protein transferase inhibitor combinations in preclinical models. *Cancer Res.* **61**, 8758–8768 (2001).
55. Foley, C. L. & Kirby, R. S. 5α-reductase inhibitors: what's new? *Curr. Opin. Urol.* **13**, 31–37 (2003).
56. Heath, R. J., White, S. W. & Rock, C. O. Lipid biosynthesis as a target for antibacterial agents. *Prog. Lipid Res.* **40**, 467–497 (2001).
57. Goldstein, J. M. The new generation of antipsychotic drugs: how atypical are they? *Int. J. Neuropsychopharmacol.* **3**, 339–349 (2000).
58. Hodgkin, J. Seven types of pleiotropy. *Int. J. Dev. Biol.* **42**, 501–505 (1998).
A thorough review and catalogue of manifestations of pleiotropy from a genetic perspective.
59. Jeffery, C. J. Moonlighting proteins. *Trends Biochem. Sci.* **24**, 8–11 (1999).
60. Copley, S. D. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol.* **7**, 265–272 (2003).
61. Wistow, G. & Piatigorsky, J. Recruitment of enzymes as lens structural proteins. *Science* **236**, 1554–1556 (1987).
62. Citron, B. A. *et al.* Identity of 4α-carbinolamine dehydratase, a component of the phenylalanine hydroxylation system, and DCoH, a transregulator of homeodomain proteins. *Proc. Natl Acad. Sci. USA* **89**, 11891–11894 (1992).
63. Sun, Y. J. *et al.* The crystal structure of a multifunctional protein: phosphoglucose isomerase/autocrine motility factor/neuroleukin. *Proc. Natl Acad. Sci. USA* **96**, 5412–5417 (1999).
64. Gomez, A., Domedel, N., Cedano, J., Pinol, J. & Querol, E. Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins? *Bioinformatics* **19**, 895–896 (2003).
65. Kousteni, S. *et al.* Nongenotropic, sex-nonspecific signaling through the estrogen or androgen receptors: dissociation from transcriptional activity. *Cell* **104**, 719–730 (2002).
66. Hughes, A. L. Adaptive evolution after gene duplication. *Trends Genet.* **18**, 433–434 (1994).
Suggests that pleiotropy might precede paralogy in the evolution of novel gene function.
67. Brett, D. *et al.* Alternative splicing and genome complexity. *Nature Genet.* **30**, 29–30 (2002).
68. Wagner, A. The role of population size, pleiotropy and fitness effects of mutations in the evolution of overlapping gene functions. *Genetics* **154**, 1389–1401 (2000).
69. Gu, Z. *et al.* Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63–66 (2003).
70. Zhou, F. C., Lesch, K. P. & Murphy, D. L. Serotonin uptake into dopamine neurons via dopamine transporters: a compensatory alternative. *Brain Res.* **942**, 109–119 (2002).
71. Muoio, D. M. *et al.* Fatty acid homeostasis and induction of lipid regulatory genes in skeletal muscles of peroxisome proliferator-activated receptor (PPAR)-α knock-out mice. Evidence for compensatory regulation by PPAR-δ. *J. Biol. Chem.* **277**, 26089–26097 (2002).
72. Troy, C. M. *et al.* Death in the balance: alternative participation of the caspase-2 and -9 pathways in neuronal death induced by nerve growth factor deprivation. *J. Neurosci.* **21**, 5007–5016 (2001).

73. Zhang, J. *et al.* The tissue-specific, compensatory expression of cyclooxygenase-1 and -2 in transgenic mice. *Prostaglandins Other Lipid Mediat.* **67**, 121–135 (2002).
74. Wang, L. *et al.* Redundant pathways for negative feedback regulation of bile acid production. *Dev. Cell* **2**, 721–731 (2002).
75. Mesulam, M. M. *et al.* Acetylcholinesterase knockouts establish central cholinergic pathways and can use butyrylcholinesterase to hydrolyze acetylcholine. *Neuroscience* **110**, 627–639 (2002).
76. Haddad, J. J. Cytokines and related receptor-mediated signaling pathways. *Biochem. Biophys. Res. Commun.* **297**, 700–713 (2002).
77. Dumont, J. E., Pecasse, F. & Maenhaut, C. Crosstalk and specificity in signalling. Are we crosstalking ourselves into general confusion? *Cell Signal.* **13**, 457–463 (2001).
78. Iwamoto, T. *et al.* STAT and SMAD signalling in cancer. *Histol. Histopathol.* **17**, 887–895 (2002).
79. Takayanagi, H. *et al.* T-cell-mediated regulation of osteoclastogenesis by signalling cross-talk between RANKL and IFN-γ. *Nature* **408**, 600–605 (2000).
80. Stork, P. J. & Schmitt, J. M. Crosstalk between cAMP and MAP kinase signaling in the regulation of cell proliferation. *Trends Cell Biol.* **12**, 258–266 (2002).
81. Schwartz, M. A. & Ginsberg, M. H. Networks and crosstalk: integrin signalling spreads. *Nature Cell Biol.* **4**, E65–E68 (2002).
82. Marshall, F. H. *et al.* $GABA_B$ receptors function as heterodimers. *Biochem. Soc. Trans.* **27**, 530–535 (1999).
83. Angers, S., Salahpour, A. & Bouvier, M. Biochemical and biophysical demonstration of GPCR oligomerization in mammalian cells. *Life Sci.* **68**, 2243–2250 (2002).
84. North, R. A. Molecular physiology of P2X receptors. *Physiol. Rev.* **82**, 1013–1067 (2002).
85. Czirjak, G. & Enyedi, P. Formation of functional heterodimers between the TASK-1 and TASK-3 two-pore domain potassium channel subunits. *J. Biol. Chem.* **277**, 5426–5432 (2002).
86. Liu, Y. & Eisenberg, D. 3D domain swapping: as domains continue to swap. *Protein Sci.* **11**, 1285–1299 (2002).
87. Waxman, D. & Peck, J. R. Pleiotropy and the preservation of perfection. *Science* **279**, 1210–1213 (1998).
88. Galis, F., van Dooren, T. J. & Metz, J. A. Conservation of the segmented germband stage: robustness or pleiotropy? *Trends Genet.* **18**, 504–509 (2002).
89. Lipman, D. J. *et al.* The relationship of protein conservation and sequence length. *BMC Evol. Biol.* **2**, 20 (2002).
90. Duret, L. & Mouchiroud, D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17**, 68–74 (2000).
91. Hastings, K. E. M. 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 (1996).
92. Moskowitz, D. W. Is angiotensin I-converting enzyme a "master" disease gene? *Diabetes Technol. Ther.* **4**, 683–711 (2002).
93. Viner, J. L., Umar, A. & Hawk, E. T. Chemoprevention of colorectal cancer: problems, progress, and prospects. *Gastroenterol. Clin. North Am.* **31**, 971–999 (2002).
94. Horowitz, N. H. in *Evolving Genes and Proteins* (eds Bryson, V. & Vogel, H. J.) 15–23 (Academic Press, New York, 1965).
95. Belfaiza, J. *et al.* Evolution of biosynthetic pathways: two enzymes catalyzing consecutive steps in methionine biosynthesis originate from a common ancestor and possess a similar regulatory region. *Proc. Natl Acad. Sci. USA* **83**, 867–871 (1986).
96. Wilmanns, M. *et al.* Structural conservation in parallel β/α-barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis. *Biochemistry* **30**, 9161–9169 (1991).
97. Fani, R., Lio, P., Chiarelli, I. & Bazzicalupo, M. The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the *hisA* and *hisF* genes. *J. Mol. Evol.* **38**, 489–495 (1994).
98. Alves, R., Chaleil, R. A. & Sternberg, M. J. Evolution of enzymes in metabolism: a network perspective. *J. Mol. Biol.* **320**, 751–770 (2002).
99. Copley, R. R. & Bork, P. Homology among $(\beta\alpha)_8$ barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.* **303**, 627–641 (2000).
100. Forst, C. V. & Schulten, K. Phylogenetic analysis of metabolic pathways. *J. Mol. Evol.* **52**, 471–489 (2001).
101. Wagner, A. Robustness against mutations in genetic networks of yeast. *Nature Genet.* **24**, 355–361 (2001).
102. Grange, R. W. *et al.* Functional and molecular adaptations in skeletal muscle of myoglobin-mutant mice. *Am. J. Physiol. Cell Physiol.* **281**, C1487–C1494 (2001).
103. de Groof, A. J., Oerlemans, F. T., Jost, C. R. & Wieringa, B. Changes in glycolytic network and mitochondrial design in creatine kinase-deficient muscles. *Muscle Nerve* **24**, 1188–1196 (2001).
104. Zheng, T. S. *et al.* Deficiency in caspase-9 or caspase-3 induces compensatory caspase activation. *Nature Med.* **6**, 1241–1247 (2001).
105. Putcha, G. V. *et al.* Intrinsic and extrinsic pathway signaling during neuronal apoptosis: lessons from the analysis of mutant mice. *J. Cell Biol.* **157**, 441–453 (2002).
106. Marcotte, E. M. *et al.* Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
Shows that products of genes that fuse in the course of evolution also tend to interact or participate in common pathways in species where they remain unfused.
107. Pellegrini, M. *et al.* Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96**, 4285–4288 (1999).
108. Marcotte, E. M., Xenarios, I., van der Bliek, A. M. & Eisenberg, D. Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **97**, 12115–12120 (2000).
109. Goh, C. S. *et al.* Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**, 283–293 (2000).
110. Goh, C. S. & Cohen, F. E. Co-evolutionary analysis reveals insights into protein–protein interactions. *J. Mol. Biol.* **324**, 177–192 (2002).
111. Bafna, V., Hannenhalli, S., Rice, K. & Vawter, L. Ligand-receptor pairing via tree comparison. *J. Comput. Biol.* **7**, 59–70 (2000).
112. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.* **14**, 609–614 (2001).
113. Koretke, K. K. *et al.* Evolution of two-component signal transduction. *Mol. Biol. Evol.* **17**, 1956–1970 (2000).
114. Fraser, H. B. *et al.* Evolutionary rate in the protein interaction network. *Science* **296**, 750–752 (2002).
115. Jordan, I. K., Wolf, Y. I. & Koonin, E. V. No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol. Biol.* **3**, 1 (2003).
116. Fraser, H. B., Wall, D. P. & Hirsh, A. E. A simple dependence between protein evolution rate and the number of protein–protein interactions. *BMC Evol. Biol.* **3**, 11 (2003).
117. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002).
118. Featherstone, D. E. & Broadie, K. Wrestling with pleiotropy: genomic and topological analysis of the yeast expression network. *Bioessays* **24**, 267–274 (2002).
119. Ohno, S. *Evolution by Gene and Genome Duplication* (Springer, Berlin, 1970).
The classic statement of the theory that duplicated genes are released from selective pressure and are therefore free to rapidly evolve new function.
120. Wilson, A. C., Carlson, S. S. & White, T. J. Biochemical evolution. *Annu. Rev. Biochem.* **46**, 573–639 (1977).
121. Jordan, I. K., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* **12**, 962–968 (2002).
122. Hirsh, A. E. & Fraser, H. B. Protein dispensability and rate of evolution. *Nature* **411**, 1046–1049 (2001).
123. Pal, C., Papp, B. & Hurst, L. D. Genomic function: rate of evolution and gene dispensability. *Nature* **421**, 496–497 (2003).
124. Hirsh, A. E. & Fraser, H. B. Genomic function: Rate of evolution and gene dispensability. *Nature* **421**, 497–498 (2003).
125. Hurst, L. D. & Smith, N. G. C. Do essential genes evolve slowly? *Curr. Biol.* **9**, 747–750 (1999).
126. Conant, G. C. & Wagner, A. GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res.* **30**, 3378–3386 (2002).
127. Schrag, J. D., Winkler, F. K. & Cygler, M. Pancreatic lipases: evolutionary intermediates in a positional change of catalytic carboxylates? *J. Biol. Chem.* **267**, 4300–4303 (1992).
128. Zhang, J., Dyer, K. D. & Rosenberg, H. F. Evolution of the rodent eosinophil-associated RNase gene family by rapid gene sorting and positive selection. *Proc. Natl Acad. Sci. USA* **97**, 4701–4706 (2000).
129. Wooding, S. P. *et al.* DNA sequence variation in a 3.7-kb noncoding sequence 5' of the *CYP1A2* gene: implications for human population history and natural selection. *Am. J. Hum. Genet.* **71**, 528–542 (2002).
130. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
131. Bromham, L. & Penn, D. The modern molecular clock. *Nature Rev. Genet.* **4**, 216–224 (2003).
132. Mangel, M. & Samaniego, F. J. Abraham Wald's work on aircraft survivability. *J. Amer. Statistical Assoc.* **79**, 259–270 (1984).
133. Hardison, R. C., Oeltjen, J. & Miller, W. Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* **8**, 959–966 (1997).
134. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human–mouse genome comparisons to locate regulatory sites. *Nature Genet.* **26**, 225–228 (2000).
135. Bofelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
136. Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113 (1970).
The origin of the terms 'orthologue' and 'paralogue'.
137. Van Valen, L. A new evolutionary law. *Evol. Theory* **1**, 1–30 (1973).
138. Black, C. G. & Coppel, R. L. Synonymous and non-synonymous mutations in a region of the *Plasmodium chabaudi* genome and evidence for selection acting on a malaria vaccine candidate. *Mol. Biochem. Parasitol.* **111**, 447–451 (2000).
139. Woolhouse, M. E., Webster, J. P., Domingo, E., Charlesworth, B. & Levin, B. R. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genet.* **32**, 569–577 (2002).
140. Enard, W. *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340–343 (2002).
Introduces the notion of phylogenetic analysis of overall gene expression patterns.
141. Tavazoie, S. *et al.* Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285 (1999).
142. Wang, Y., Schnegelsberg, P. N., Dausman, J. & Jaenisch, R. Functional redundancy of the muscle-specific transcription factors Myf5 and myogenin. *Nature* **379**, 823–825 (1996).
143. Tong, A. H. *et al.* A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**, 321–324 (2002).
144. Ajay, A., Walters, W. P. & Murcko M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **41**, 3314–3324 (1998).
145. Muegge, I., Heald, S. L. & Brittelli, D. Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* **44**, 1841–1846 (2001).
146. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 4–25 (1997).
147. Veber, D. F. *et al*. Molecular properties that influence oral bioavailability of drug candidates. *J. Med. Chem.* **45**, 2615–2623 (2002).
148. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Rev. Drug Discov.* **1**, 727–730 (2002).
An influential review that helps establish a view of targets as having measurable properties (their drug-binding domain content) making them generally suitable for therapeutic intervention.

Acknowledgements

The author thanks J. R. Brown, K. Rice, and N. Odendahl for many helpful comments on the manuscript.

Online links

DATABASE

The following terms in this article are linked online to:

LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/

DCOHM | BRCA1 | CFTR | Cyp2a1 | Cyp2a3 | Cyp2a4 | CYP2A6 | dopamine D_2 | EGFR | ERBB2 | FOXP2 | GPI | PPAR-γ | SRD5A1 | SRD5A2

FURTHER INFORMATION

PHYLogeny Inference Package (PHYLIP):
http://evolution.genetics.washington.edu/phylip.html

Phylogenetic Analysis Using Parsimony (PAUP):
http://paup.csit.fsu.edu/index.html

Resampled Inference of Orthologs (RIO):
http://www.rio.wustl.edu

Phylogenetic Analysis by Maximum Likelihood (PAML):
http://abacus.gene.ucl.ac.uk/software/paml.html

Access to this interactive links box is free online.