Comparative Genomics

19

Ross C. Hardison

Abstract Comparative genomics harnesses the power of sequence comparisons within and between species to deduce not only evolutionary history but also insights into the function, if any, of particular DNA sequences. Changes in DNA and protein sequences are subject to three evolutionary processes: drift, which allows some neutral changes to accumulate, negative selection, which removes deleterious changes, or positive selection, which acts on adaptive changes to increase their frequency in a population. Quantitative data from comparative genomics can be used to infer the type of evolutionary force that likely has been operating on a particular sequence, thereby predicting whether it is functional. These predictions are good but imperfect; their primary role is to provide useful hypotheses for further experimental tests of function. Rates of evolutionary change vary both between functional categories of sequences and regionally within genomes. Even within a functional category (e.g. protein or gene regulatory region) the rates vary. A more complete understanding of variation in the patterns and rates of evolution should improve the predictive accuracy of comparative genomics. Proteins that show signatures of adaptive evolution tend to fall into the major functional categories of reproduction, chemosensation, immune response and xenobiotic metabolism. DNA sequences that appear to be under the strongest evolutionary constraint are not fully understood, although many of them are active as transcriptional enhancers. Human sequences that regulate gene expression tend to be conserved among placental mammals, but the phylogenetic depth of conservation of individual regulatory regions ranges from primate-specific to pan-vertebrate.

Contents

R.C. Hardison (✉)
Center for Comparative Genomics and Bioinformatics,
Huck Institutes of Life Sciences,
Department of Biochemistry and Molecular Biology,
The Pennsylvania State University,
PA 16802, USA
e-mail: rch8@psu.edu

M.R. Speicher et al. (eds.), *Vogel and Motulsky's Human Genetics: Problems and Approaches,*
DOI 10.1007/978-3-540-37654-5_19,

19

19.1 Goals, Impact, and Basic Approaches of Comparative Genomics

Comparative genomics uses evolutionary theory to glean insights into the function of genomic DNA sequences. By comparing DNA and protein sequences between species or among populations within a species, we can estimate the rates at which various sequences have evolved and infer chromosomal rearrangements, duplications and deletions. This evolutionary reconstruction can then be used to predict functional properties of the DNA. Sequences that are needed for functions common to the species being compared are expected to change little over evolutionary time, whereas sequences that confer an adaptive advantage when altered are expected to have greater divergence between species. Furthermore, sequence comparisons can help in predicting what role is played by a particular functional region, e.g., coding for a protein or regulating the level of expression of a gene.

These insights from comparative genomics are having a strong impact on medical genetics, and their role is expected to become more pervasive in the future. When profound mutant phenotypes lead to the discovery of genes in model organisms (bacteria, yeast, flies, etc.), the human genome is immediately searched for homologs, which frequently are discovered to be involved in similar processes. Control of the cell cycle [76] and defects in DNA repair associated with cancers [24, 47] are particularly famous examples. In studies of the noncoding regions of the human genome, conservation has become almost a proxy for function [20, 26, 64], and we will explore the power and limitations of this approach more in this chapter. The mapping and genotyping of millions of polymorphisms in humans [32] coupled with the availability of genome sequences of species closely related to humans [14, 70] has stimulated great interest in discovering genes and control sequences that are adaptive in humans, which may provide clues to the genetic elements that make us uniquely human (see Chaps. 8 and 16). As more and more loci are implicated in disease and susceptibility to diseases, identifying strong candidates for the causative mutations becomes more challenging. Research in comparative genomics is helping to meet this challenge by generating estimates across the human genome of sequences likely to be conserved for functions common to many species as well as sequences showing signs of adaptive change. Finding disease-associated markers in either type of sequence could rapidly narrow the search for mutations that cause a phenotype.

19.1.1 How Biological Sequences Change Over Time

All DNA sequences are subject to change, and these changes provide the fuel for evolution. Replication is highly accurate but not perfect, and despite the correction of many replication errors by repair processes during S-phase, a small fraction is retained as altered sequences. Mutagens in the environment can damage DNA, and some of these induced mutations escape repair. In addition, DNA bases can change spontaneously, for example, oxidative deamination of cytosine to produce uracil. The *mutation rate* is the number of sequence changes escaping correction and repair that

accumulate per unit of time. The average mutation rate in humans has been estimated to be about 2 changes in 10^8 sites per generation [43, 57]. Thus for a diploid genome of 6×10^9 bp, about 120 new mutations arise in each generation. As will be discussed later in more detail, the mutation rate varies among loci and depends on the context, with transitions at CpG dinucleotides occurring about ten times more frequently than other mutations.

Mutations can be substitutions of one nucleotide for another, deletions of strings of nucleotides, insertions of nucleotides, or rearrangements of chromosomes, including duplications of DNA segments. Substitutions are about ten times as frequent as the length-changing alterations, with transitions greatly favored over transversions.

Mutations occur in individuals, and it is instructive to consider how an alteration in a single individual can eventually lead to a sequence difference between two species, which we call a *fixed difference*. Of course, only mutations arising in the germ-line can be passed along to progeny and have some possibility of fixation. Initially, the allele carrying a mutation has a low frequency in the population, i.e., $1/(2N_e)$ for a diploid organism, where N_e is the effective population size. All the mating individuals in a population contribute to the pool of new alleles. Mutant alleles that are disadvantageous will be cleared out of the population quickly, whereas those that confer a selective advantage rapidly will go to fixation (occurrence in most members of a population). However, many of the new mutations will have no effect on the individual; we call these mutations with no functional consequence *polymorphisms* or *neutral changes*. The frequency of these polymorphisms will increase or decrease depending on the results of matings and survival of progeny. The vast majority will be transitory in the population, with most headed for loss. However, the stochastic fluctuations in allele frequencies will allow some to eventually increase to a high frequency. Thus, some of the neutral changes lead to fixed differences. In fact, Kimura [40] and others have argued that such neutral changes are the major contributors to the overall evolution of the genome.

In order for a sequence change to have an effect on an organism, the change has to occur in a region that is involved in some function. Examples of such regions are an exon encoding part of a protein or a promoter or enhancer involved in gene regulation. The rapid removal of disadvantageous alleles results from *negative* or *purifying selection* (Fig. 19.1). The rapid fixation of advantageous alleles is *adaptive evolution* resulting from positive selection. Biological function is inferred from evidence of selection. Thus, the aim of comparative genomics to identify functional sequences can be stated as a goal of finding DNA sequences that show significant signs of positive or negative selection.

In addition to mutations of single bases, strings of nucleotides can be inserted or deleted as a result of replication errors or recombination. Often, the direction

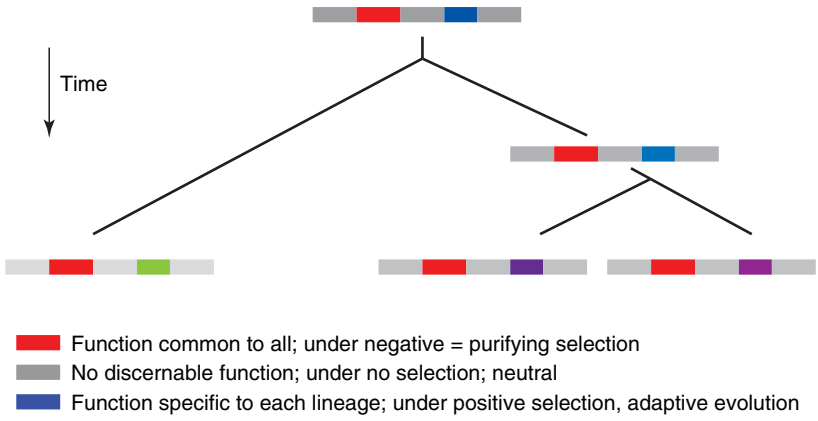

Fig. 19.1 Three modes of evolution, two of which are associated with function. The red line indicates a functional DNA sequence whose role has remained the same from ancestor to contemporary sequences, and thus it has been subject to purifying selection. The blue line represents a sequence that was functional in the ancestor, but changes in separate lineages (illustrated by different shades of blue, green, and purple) are adaptive and hence are subject to positive selection. The gray lines represent sequences of no known function, i.e., neutral DNA

of the event is not known because it is inferred from a gap in an alignment of only two sequences. In these cases the event is called an *indel*. Adding a third sequence to the alignment as an outgroup allows one to conclude with some confidence whether the event is an insertion or a deletion. Indels are less frequent than nucleotide substitution, and their frequency declines sharply with the size of the insertion or deletion. However, a single insertion or deletion can involve tens of thousands of nucleotides. Thus, they account for the majority of the nucleotides that differ between closely related species.

Rearrangements of chromosomes, such as intrachromosomal duplications and inversions or interchromosomal translocations, also lead to large-scale changes both in contemporary populations and over evolutionary time. Some chromosomal rearrangements are associated with human disease (see Chap. XX). In comparisons over evolutionary time, e.g., between mammalian orders, the history of chromosomal rearrangements can be reconstructed with some accuracy.

19.1.2 Purifying Selection

DNA sequences that encode the same function in contemporary species and in the last common ancestral species have been subject to *purifying* selection. The DNA sequence carried out some function in the ancestor, and any changes to this successful invention are more likely to break it than to improve it. Mutations in the sequence tend to work less well than the original one, and those mutations are cleared from the population. Hence the selective pressure to maintain a function prevents the DNA sequence from accumulating many changes, and the selection is referred to as purifying. This type of selective pressure tends to decrease the number of changes observed, and thus it is also called *negative* selection. The sequence under purifying selection is *constrained* by its function to remain similar to the ancestor. Saying that a sequence is subject to constraint is the equivalent of saying that it is subject to purifying selection. Examples of sequences under constraint include most protein-coding regions and many DNA sequences that regulate the level of expression of a gene.

In this chapter, we distinguish between conserved and constrained elements. A feature (e.g., a segment of DNA, a protein, an anatomical structure) that is found in contemporary species and that is inferred as being derived from a similar feature in the last common ancestor is conserved. In particular, a DNA sequence that reliably aligns between two species is considered to be conserved. That does not necessarily mean that it is functional. Evidence of *constraint*, i.e., alignment with a level of similarity greater than expected for neutral DNA, is taken as an indicator of function common to the two species.

The hallmark of purifying selection is a rate of change that is slower than that of neutral DNA. The next section (Sect. 19.2) will delve more deeply into how rates of evolution are determined, but for now assume that we can align related sequences with reasonable accuracy and can use that alignment to measure how frequently mismatches occur. Then the problem of finding sequences under purifying selection becomes one of determining the substitution rate in a segment that is a candidate for being functional and comparing it to the rate in neutral DNA. DNA segments whose inferred rate of evolutionary change is significantly lower than neutral will show a peak of similarity for comparisons at a sufficient phylogenetic distance (e.g., human versus mouse in Fig. 19.2).

In order to distinguish neutral from constrained DNA, sequences of divergent species must be compared. The choice of species to compare will depend on the questions being examined, but enough sequence change must have occurred to distinguish signal from noise. In practical terms, human comparisons with chimpanzee are too close (too similar) to effectively find constrained sequences, but multiple alignments among many primates do have considerable power [8]. Many studies have used comparisons between mammalian orders, such as primate (human) with rodent (mouse), to see the constrained sequences (Fig. 19.2).

19.1.3 Models of Neutral DNA

Although the concept of DNA that has no function is very useful and has led to much insight in molecular evolutionary genetics, it is difficult to establish that any DNA is truly neutral. Several models for neutral DNA are in common use. One of the earliest is the set of nucleotides in protein-coding regions that can be altered without changing the encoded amino acid [41]. The nucleotides are called *synonymous* or *silent* sites. They are neutral with respect to coding capacity, but alterations in particular synonymous sites can affect

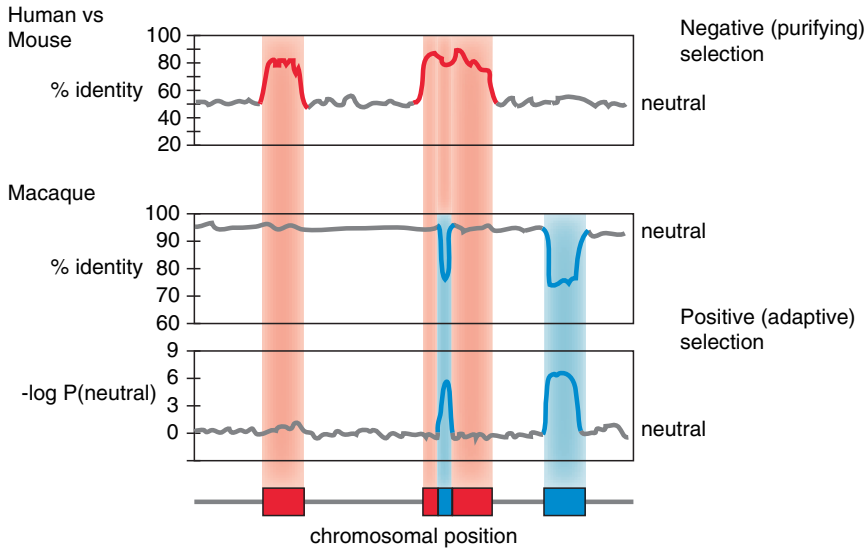

Fig.19.2 Ideal cases for interpretation of sequence similarity. Idealized graphs of levels of sequence similarity (as percent identity) for a segment of a human chromosome compared with mouse (top) and rhesus macaque (middle), and of the likelihood that the DNA interrogated by the human-macaque comparison is not neutral (negative logarithm of the probability that the sequence similarity comes from the distribution of values for comparisons of neutral DNA, third graph). In the graphs, values that are close to those observed for a model of neutral DNA are shown in gray, those that indicate the action of negative selection are red, and those that indicate positive selection are blue. The bottom map is an interpretation of the graphs as discrete segments of DNA either under negative (red boxes) or positive (blue boxes) selection on a background of neutral DNA (gray line). Note that one segment shows evidence of negative selection since the separation of primates from rodents (red in top graph) but positive selection since the separation of human and Old World monkey (macaque) lineages (blue in middle and bottom graphs)

translation efficiency, splicing, or other processes. The latter appear to be a minority of synonymous sites, and as a group the synonymous sites are the most frequently used neutral model.

Another useful model for neutral DNA are *pseudogenes*. These are copies of functional genes, but the copies no longer code for protein because of some disabling mutation, such as a frameshift mutation or a substitution that generates a translation termination codon. For the period of time since the inactivating mutation, the pseudogene has likely been under little or no selective pressure. The rate of divergence of pseudogenes after inactivation is clearly higher than that of the homologous functional genes, and they have been used successfully as neutral models in many studies of particular gene families (e.g., [48]). One limitation of using pseudogenes as a neutral model is the uncertainty of determining when the inactivating mutation(s) occurred. Also, they are rather sparse for genome-wide studies.

For comparisons in mammalian genomes, *ancestral repeats* (Fig. 19.3) have proved effective, albeit imperfect, models for neutral DNA [27, 85]. The interspersed DNA repeats in the genomes of humans and other mammals are derived from transposable elements, mostly *retrotransposons* that move via an RNA intermediate. Members of an interspersed repeat family generated by recent transposition (on an evolutionary time-scale) are quite similar to each other because they have not had sufficient time to diverge. These are restricted to particular clades, such as the *Alu* repeats that are prevalent in primate genomes. Considerably more differences are observed among members of repeat families that are derived from transposons active in an ancestral species because of the longer divergence time. The members of these older repeat families are present in all the descendant species. Examples include *LINE2* and *MIR* repeats, which are present in the genomes of all eutherian mammals examined. Interestingly, all the members of these ancestral repeat families are quite divergent from each other, indicating that they have not been actively transposing since the separation of the descendant species. Thus, most ancestral repeats appear to be relics of ancient transposable elements, and they are not active even for

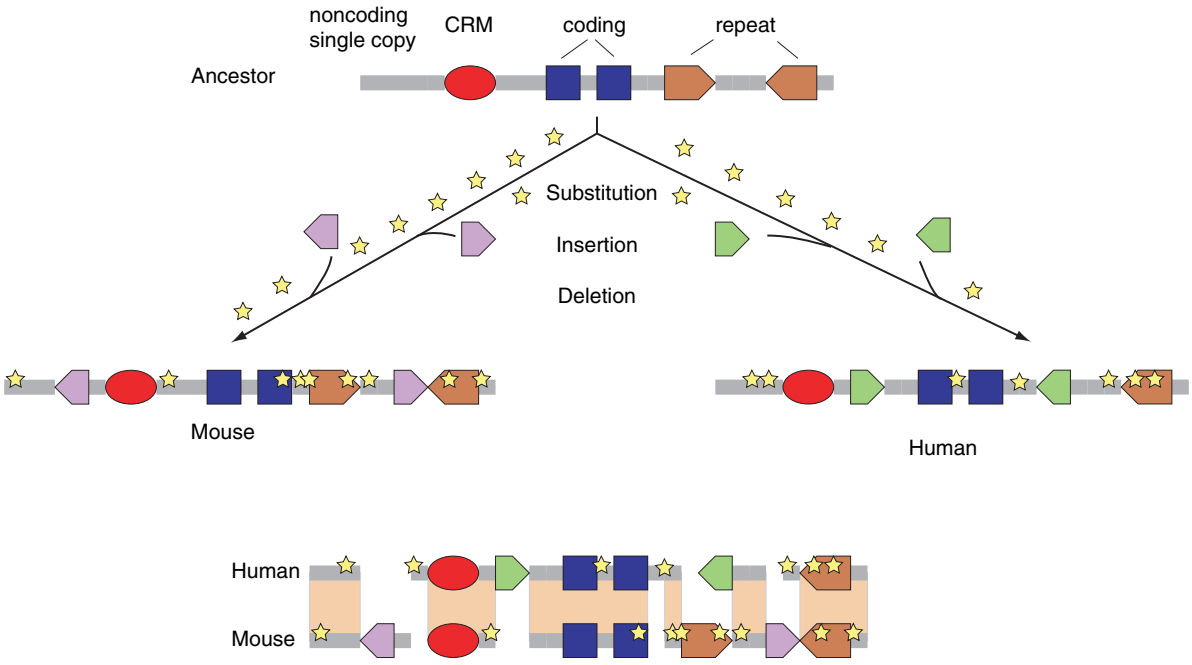

Fig. 19.3 Substitutions, insertions of transposable elements and deletions in the evolution of genomes. (**a**) Illustration of functional regions such as protein-coding exons (blue boxes), cis-regulatory modules (CRMs, red ovals), such as enhancers and promoters, and ancestral repeats (brown pointed boxes). After divergence of rodents and primates, sequences diverge by substitutions (gold stars), insertion of lineage-specific transposable elements (purple and green pointed boxes), and deletions. (**b**) Alignments of the contemporary species allow some of the evolutionary history to be reconstructed, including deletions inferred from the nonaligning portions

transposition. The vast bulk of these ancestral repeats have no apparent function. They are found frequently in eutherian mammals, and thus provide a neutral model with many sites.

When interpreting any measurement or study involving a comparison with a neutral model, it is important to keep in mind that the deduced absence of function is limited by contemporary knowledge. Experimental tests and molecular evolutionary studies have shown that some individual synonymous sites and ancestral repeats are not neutral. They do not constitute the bulk of the sites in these neutral models, and of course the known functional sites can be removed from the neutral set. However, future studies could reveal additional function, which will affect interpretations based on these neutral models.

19.1.4 Adaptive Evolution

The functions of some DNA segments and proteins have changed along the evolutionary lineages to contemporary species. Some sequence changes confer a new function on the DNA or protein that helps the organism adapt to a new environment or condition. These advantageous mutations increase in frequency in a population, leading to fixation (i.e., becoming the predominant allele in the population). The selective pressure favoring these changes is called *positive* selection, since it tends to increase the frequency of changes. This leads to *adaptive evolution*, i.e., a change in a DNA or protein sequence that favors survival and procreation of an organism. The positive selection for new functionality is also referred to as *Darwinian* selection.

The hallmark of adaptive evolution is a rate of sequence change that is faster than that of neutral DNA. Sequences subject to adaptive evolution may change so much that they will not align reliably at greater phylogenetic distances (Fig. 19.1). Also, the selective pressure leading to adaptive changes may apply only recently or in limited clades, such as among humans or among humans and great apes. Thus, sequence comparisons to find adaptive changes are usually done for closely related, recently diverged sequences (Fig. 19.2). The signal for positive selection

may be captured as a significant decrease in similarity between species or an increase in the probability that a sequence has not evolved neutrally (Fig. 19.2).

19.2 Alignments of Biological Sequences and Their Interpretation

Biological sequence comparisons are most commonly done with protein sequences (strings of amino acids) or DNA sequences (strings of nucleotides). The comparisons begin with an alignment, which is a mapping of one sequence onto another with insertions of gaps (often indicated by a dash) to optimize a similarity score (Fig. 19.3). The score can be determined in a variety of ways, but in all cases matching symbols (for amino acids or nucleotides as appropriate) are favored, whereas mismatches are not favored and gaps are penalized. The gap penalty frequently takes the form of a gap-open penalty plus an additional, smaller penalty for each position included in the gap. The latter are referred to *as affine gap penalties*.

19.2.1 Global and Local Alignments

A *global* alignment maps each symbol in one sequence onto a corresponding symbol in another sequence. The result is an alignment of the two (or more) sequences from their beginnings to their ends, with any length differences accommodated by gaps that are introduced. This is an appropriate strategy for sequences are related to each other over their entirety. That is the case for many proteins and many mRNAs. The earliest computer program for aligning two biological sequences, written by Needleman and Wunsch [58], generates global alignments. Popular contemporary programs for aligning proteins, such as *ClustalW* [80], also compute global alignments. Global aligners for DNA sequences include *VISTA* [54], *MAVID* [9], and *LAGAN* [10].

A frequent task in comparative genomics is to find matches between two or more sequences that are not related over their entire lengths. For instance, two protein sequences may be related only in one or a few domains, but be different in other parts. The protein-coding portions of genes are frequently divided into short exons that are separated by introns. Exons tend to be under constraint, whereas much of the intronic DNA may be neutral, and thus at a sufficient phylogenetic distance introns can be so divergent that they no longer align, whereas exons will match well. The most common use of comparative genomics is to search a large database of all compiled DNA or protein sequences with a query sequence of interest. In this case, the goal is to find a match that may comprise only one part in billions of the database. When a match between only a portion of two or more sequences is desired, then a *local* alignment should be generated. One of the earliest computer programs for finding local alignments came from Smith and Waterman [75]. The *blast* family of programs (Basic Local Alignment Search Tool, [1]) is used for database searches. One variant, called *blastZ*, has been adapted to compute local alignments of long genomic DNA sequences [72].

19.2.2 Aligning Protein Sequences

Proteins are composed of 20 amino acids, so that for any position in one sequence the possibilities for alignment with a position in a comparison sequence are 1 match, 19 mismatches, or a gap. However, the likelihood for each of the 19 mismatches is not the same. Replacement of an amino acid by a chemically similar amino acid occurs much more frequently than does replacement with a distinctly different amino acid. These different frequencies of amino acid substitutions can be captured as a *scoring matrix*, in which matches are given the highest similarity score and mismatches that occur frequently in protein sequences are given positive scores, decreasing with declining frequencies of the substitution. These scoring matrices are determined by the frequency with which mismatches are observed in well-aligned sequences. Several effective matrices have been generated, beginning with the pioneering work of Dayhoff et al. [18] and continuing on to the BLOSSUM matrices of Henikoff [29].

Alignments can be used to organize relationships among the large number of sequenced proteins. Large compilations of aligned protein sequences are analyzed to find clusters of proteins that appear to share a common ancestor and to find blocks of aligned sequences that are distinctive for various protein domains. Indeed, when genes and their encoded proteins

are predicted or identified in genome sequences, the primary basis for making inferences about their function is sequence similarity to known proteins.

Sequence similarity between proteins can be found with considerably greater sensitivity than can be found using a DNA sequence. The reason is that the 20 amino acids found in proteins constitute a much more complex group of characters, or alphabet, than the four nucleotides found in DNA. Thus, alignments between distantly related proteins may only match at a very small percentage of positions, but these are still statistically significant and they can be biologically meaningful.

19.2.3 Aligning Large Genome Sequences

The smaller alphabet for DNA sequences, consisting of only four nucleotides (A, C, G, T), means that the threshold for statistical significance is considerably higher than that used for protein sequences. For random sequences of equal nucleotide composition, any position in one sequence should have a 25% chance of matching any position in the other. However, sufficiently long runs of matching sequences are much less likely, and reliable alignment can be generated between related sequences. Just like for alignments of protein sequences, some substitutions are more likely to occur than others. For example, transitions are much more frequent than transversions. These preferences can be incorporated into the alignment process by using scoring matrices that were deduced from the empirical frequencies of matches and substitutions in reliable alignments, similar to the process that generated scoring matrices for protein alignments.

The portions of DNA sequences that code for proteins tend to be more similar and to have many fewer indels than the rest of a genome for comparisons at a sufficient phylogenetic distance. Hence these are relatively easy to align and different alignment strategies tend to give similar results for coding regions. Other parts of the genome are more likely to have mismatches or to have undergone insertion or deletion, which requires introduction of gaps into the alignment. In these noncoding regions, choice of an alignment strategy is expected to have an impact on the result. Global aligners are expected to have somewhat greater sensitivity, but they may include more inaccurate alignments. Local aligners will not align sequences that are too dissimilar, even if they occur in analogous positions in the two genomes. More calibration of the various methods is needed to clarify these issues, but at this point there is no consensus on whether the regions that fail to align by local aligners are not homologous, or whether they are homologs that have changed so much that the similarity is not recognizable by these programs [53].

Chromosomal rearrangements complicate the construction of comprehensive alignments between genomes. Genes that are on the same chromosome in one species are *syntenic*. Groups of genes that are syntenic in humans are frequently also syntenic in mouse, and thus these groups of genes display *conserved synteny*. In addition, they frequently maintain a similar order and orientation, indicating *homology*, which is similarity because of common ancestry. The homologous segments between distantly related species rarely extend for entire chromosomes, but rather one human chromosome will align with several homology blocks in mouse, many of which are on different chromosomes in mouse (Fig. 19.4). For genome comparisons, the goal is to find all the reliable alignments within the homology blocks and deduce how the various homology blocks are connected in the genomes of the species being compared. This requires additional steps to the alignment procedure. For local aligners, it means that the large number of individual alignments needs to be organized along chromosomes. For global aligners, it means that homology blocks must be identified prior to execution of a global alignment.

Local alignments are restricted to the DNA segments between rearrangement breakpoints. A collection of local alignments can be organized into *chains* to maintain the order of DNA segments along the chromosome. In this case, local alignment A is connected to local alignment B in a chain if the beginnings of the aligned sequences in B follow the ends of the aligned sequences in A. The chains can be nested in a group, called a *net* [39], and these are used to navigate local alignments through rearrangements (Fig. 19.4). On a large scale, these nets can be used to illustrate chromosomal rearrangements between species, and on a smaller scale they can reveal multiple events associated with rearrangement breakpoints.

Global aligners can be used in genomic regions that have not been rearranged. In practice, for whole-genome alignments, homology blocks are initially identified using a rapid local alignment procedure. Then a global

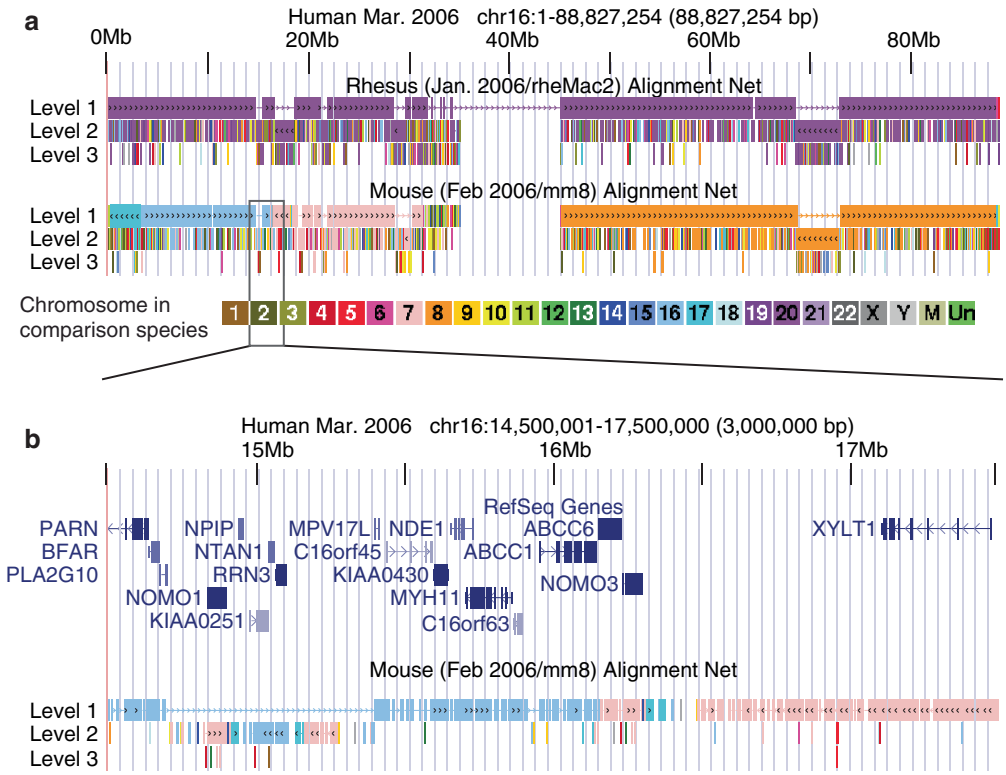

Fig. 19.4 Blocks of conserved synteny and chromosomal rearrangements with human chromosome 16 as the reference sequence. (**a**) Almost all of human chromosome aligns with rhesus chromosome 20, indicated by the purple boxes, but portions of human chromosome 16 align to different chromosomes in mouse, which are color coded by the aligning chromosome in the comparison species. For both comparisons, levels 1, 2, and 3 of a nested set of chained alignments (called a net) are shown. Local alignments form a chain when the start positions of the sequences in one alignment follow the end positions of the sequences in the preceding alignment. The level 1 chain is the highest scoring (usually longest) set of local alignments; the level 1 chain with rhesus covers almost all of rhesus chromosome 20. Gaps in the level 1 chain are filled with the highest scoring additional chains to make level 2 chains, and so on for up to six levels. Inversions are evident by changes in the directions of the arrowheads on the chain maps. (**b**) A higher resolution view of a portion of human chromosome 16 that encompasses a major change in conserved synteny from mouse chromosome 16 (light blue) to mouse chromosome 7 (pink). The diagram illustrates the results of a complex rearrangement history, including an inversion and interlacing of matches to the two mouse chromosomes. Many genes are present in this region despite the complex rearrangements of the chromosome between human and mouse

aligner such as *LAGAN* is run on the sequences in the regions that have not been rearranged [10].

Several powerful Web-servers are available for running these alignment programs on chosen sequences. Often it is prudent to use precomputed alignments because of the complexity of these alignment pipelines and the need for careful adjustment of alignment parameters for different comparisons. Nets and chains of local alignments generated by *blastZ* are available from the UCSC Genome Browser [45] and Ensembl [31]. Precomputed alignments of whole genomes generated by pipelines using *LAGAN* and *VISTA* are also available. As discussed in the next section, analyzes of these alignments can be used to predict function in genomic DNA sequences. Table 19.1 lists a selection of network servers for making and viewing alignments.

19.3 Assessment of Conserved Function from Alignments

Many of the sequences that are conserved between species can be found in the portions of genomes that align. As discussed above, alignment algorithms are good, but imperfect, and no one can guarantee that all

19

Table 19.1 Selected network servers for making and viewing alignments of genome sequences

Program or pipeline	Name	URL
blastZ, nets and chains	UCSC Genome Browser	http://genome.ucsc.edu/
blastZ, nets and chains	Ensembl	http://www.ensembl.org/
VISTA	VISTA Tools	http://genome.lbl.gov/vista/index.shtml
LAGAN	LAGAN alignment toolkit	http://lagan.stanford.edu/lagan_web/index.shtml
MAVID	MAVID Server	http://baboon.math.berkeley.edu/mavid/
blastZ and others	DCODE.org NCBI	http://www.dcode.org/
blastZ	PipMaker	http://pipmaker.bx.psu.edu/pipmaker/

the conserved sequences will align, especially as the phylogenetic distance between the species increases. Nevertheless, the portions that align should have much of the conserved DNA. Within that conserved DNA is a subset that has a function common to the species being compared; that is the portion that shows evidence of constraint, i.e., purifying selection. Thus, searching genome alignments for evidence of constraint is a major, powerful approach for finding functional DNA sequences.

19.3.1 Phylogenetic Depth of Alignments

The longer two species have been separated, the more divergent their genomes become, and thus one indicator of constraint operating on a sequence is that it aligns with sequences in distantly related species. Several insights can be gleaned by examining the phylogenetic distance at which a particular sequence or class of genomic features continues to align.

As expected, most of the human genome aligns with the genomes of our closest relative, the chimpanzee, and an Old World monkey (the rhesus macaque). The genomes of the comparisons species are not finished for the most part, and thus the values for portion aligning (Table 19.2) will be underestimated, but

Table 19.2 Portions of the human genome conserved and constrained between various species

	Distance from human		Fraction of human aligning to comparison species[d]			
Comparison species[a]	Divergence time (Myr)[b]	Substitutions per synonymous site[c]	Total genome[e]	Coding exons[f]	Regulatory regions[g]	UCEs[h]
Chimpanzee	5.40	0.015	0.95	0.96	0.97	0.99
Macaque	25.0	0.081	0.87	0.96	0.96	0.99
Dog	92.0	0.35	0.67	0.97	0.87	0.99
Mouse	91.0	0.49	0.43	0.97	0.75	1.00
Rat	91.0	0.51	0.41	0.95	0.70	1.00
Opossum	173	0.86	0.10	0.82	0.32	0.95
Chicken	310	1.2	0.037	0.67	0.06	0.95
Zebrafish	450	1.6	0.023	0.65	0.03	0.76
Number			2.858×10^9 nucleotides	250,607	1,3	481

Notes:
[a]Sources of genome sequences are: human: [33]; chimpanzee: [14]; macaque: [70]; dog: [49]; mouse: [85]; rat: [25]; opossum: Broad Institute; chicken: [30]; zebrafish: Zebrafish Sequencing Group at the Sanger Institute
[b]Divergence times for separation from the human branch to the branch leading to the indicated species are from [46]
[c]Estimated substitutions per synonymous site are from [53]
[d]The human genomic intervals in each dataset were examined for whether they aligned with DNA from each comparison species in whole-genome *blastZ* alignments [42]. An interval that is in an alignment for at least 2% of its length was counted as aligning, but in the vast majority of cases the entire interval was aligned.
[e]The number of nucleotides in the human genome that align with each species was divided by the number of sequenced nucleotides in human (given on the last line)
[f]Coding exons are from the RefSeq collection of human genes [68]
[g]Putative transcriptional regulatory regions were determined by high-throughput binding assays and chromatin alterations in the ENCODE regions [79]; the set compiled by King et al. [42] was used here
[h]Ultraconserved elements (UCEs) are the ones with at least 200 bp with no differences between human and mouse [4]

they are still informative. Since almost all of the genome aligns, of course virtually all known functional regions align between human and apes or Old World monkeys. This includes coding exons [68] and putative transcriptional regulatory regions, which are deduced from high-resolution studies on occupancy of DNA by regulatory proteins [79].

When the comparison is made with genomes of eutherian mammals outside the primate lineage, considerably less of the human genome aligns (Table 19.2). Within the 37–57% of the genome that does align, however, we find almost all of the coding exons (95–97%) and putative regulatory regions (74–89%). Even less of the genome aligns with the marsupial opossum (about 13%). At this phylogenetic distance, the alignments of coding exons tend to persist, but only 39% of the putative regulatory regions still align. Only a small fraction of the human genome aligns to more distant species, such as chickens and fish. At this distance, the estimated substitution rate in neutral DNA (synonymous sites) is so high that a segment of neutral DNA is no longer expected to align, and thus it is highly likely that all the alignments between human and chicken or fish are in functional regions.

The insights about conservation of functional elements are easier to visualize when presented as a function of phylogenetic distance (Fig. 19.5). No single comparison is adequate for all goals. Some are particularly good for one purpose, such as using human-opossum alignments for examining coding regions. Almost all the coding regions still align at this distance, but only 13% of the genome aligns. Most comparisons involve a trade-off between sensitivity (the ability to find the desired feature) and specificity (the ability to reject undesired sequences). One may want to examine alignments at a sufficient distance such that no neutral DNA is aligning, but at that distance (e.g., human-chicken) a third of the coding exons and about 90% of the putative regulatory regions no longer align. This means that the specificity is excellent but the sensitivity is lower than usually desired. In practice, it is common to examine comparisons among multiple species that have given good sensitivity, such as alignments among eutherian mammals, and to apply some discriminatory function to better ascertain the regions that are constrained or show some other evidence of function. Alignments to more distant species can be included as well, but they should not be used as an exclusive filter.

The utility and limitations of examining multiple eutherian species has been studied extensively. About 1,000 Mb align among human, mouse, and rat [25], illustrated by the central portion of the Venn diagram in Fig. 19.6. A similar study of human, dog, and mouse revealed about 812 Mb conserved in all three [49]. This approximately 1 Gigabase of genome sequence found in common can be considered the core of the genome of placental mammals. The DNA sequences needed for functions common to all eutherians are expected to be in this core, and indeed virtually all coding exons and putative regulatory regions are found in it (Table 19.2). However, it seems unlikely that this entire core is under constraint. About 162 Mb of the core consists of repetitive DNA

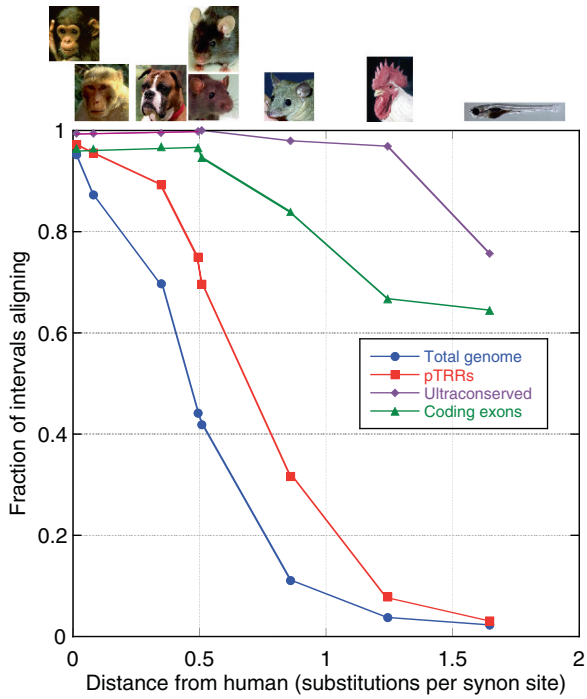

Fig. 19.5 The fraction of genomic intervals that align with comparison species at increasing phylogenetic distance. The fractions of intervals in putative regulatory regions (pTRRs, red squares), coding exons from RefSeq (green triangles) and ultraconserved elements (purple diamonds) substantially exceed the fraction of the human genome (blue circles) that aligns with each species in almost all comparisons. The comparison species in increasing order of distance from human are chimpanzee, rhesus macaque, dog, mouse, rat, opossum, chicken, and zebrafish (pictured above the graph). The distance is the estimated number of substitutions per synonymous site along the path in a tree from human to each species [53]. This measures takes into account faster rates on some lineages, and thus it places mouse and rat more distant from human than dog, despite the earlier divergence of carnivores

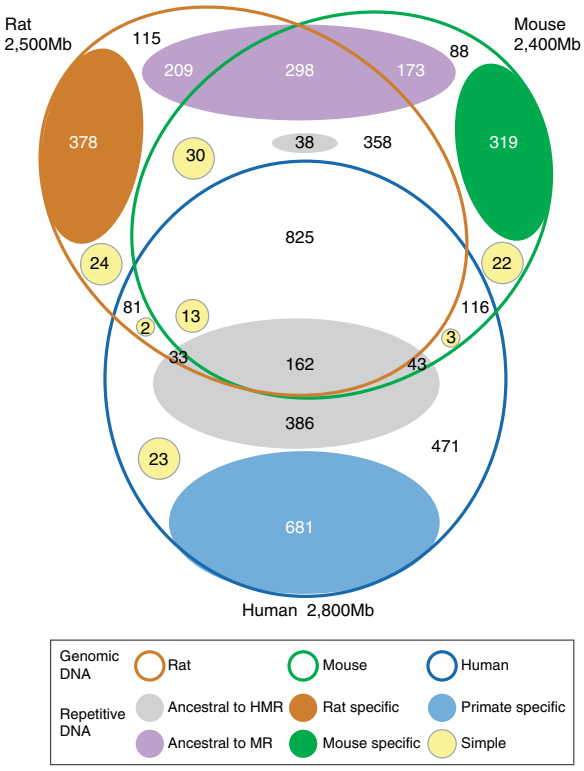

Fig. 19.6 Venn diagram showing common and distinctive sequences in humans and two rodents. As summarized in the key (box under the diagram), the outlined ellipses represent the DNA in each genome, and the overlaps show the amount of sequence aligning in all three species (rat, mouse, and human) or in only two species. Portions of the ellipses that do not overlap represent sequences that do not align. Different types of repetitive DNA are shown as colored disks, and are classified by their ancestry. Those that predate the divergence between rodents and primates are gray, and those that arose on the rodent lineage before the divergence between rat and mouse are lavender. Disks for repeats specific to each species are colored orange for rat, green for mouse, and blue for human; and disks for simple repeats are colored yellow. The disks for the repeats are placed to illustrate the approximate amount of each type in each alignment category. Uncolored areas represent nonrepetitive DNA; the bulk is assumed to be ancestral to the human–rodent divergence. The numbers of nucleotides (in Mb) are given for each sector (type of sequence and alignment category). (Reprinted from Gibbs et al. [25], with permission from Nature Publishing Group)

that is ancestral to primates and rodents (Fig. 19.6). As discussed above, most of this ancestral repetitive DNA can be considered neutral. Granted that some of these ancestral repeats may indeed be functional, it is unlikely that all of them are. Hence, even in the approximately 800 Mb of the core that is nonrepetitive, it is expected that some, and maybe much, also lack a function conserved in all eutherians. This illustrates the need for further discrimination of constrained sequences from those that are conserved but are apparently neutral. Figure 19.6 also shows that the rat and mouse genomes share many DNA sequences that are not in human, and about 358 Mb are nonrepetitive. One may expect to find rodent-specific functional sequences in these portions of the mouse and rat genome. Genomic DNA sequences that are found only in rat or only in mouse are dominated by lineage-specific interspersed repeats.

19.3.2 Portion of the Human Genome Under Constraint

Within the subset of the human genome that aligns with other species, we want to know what fraction of it appears to be under constraint (covered in this section), and then to be able to identify the constrained sequences (covered in the next section). One way to estimate the portion of the human genome under constraint is to evaluate all the segments that align with a comparison species for a level of similarity higher than that seen for neutral DNA. This would be a straightforward approach if we knew all the neutral DNA (which we do not; see Sect. 19.1.3), and if the neutral DNA diverged at the same rate at all positions in the chromosome (illustrated by the ideal case in Fig. 19.2). However, the estimated neutral rates show substantial local variation across the human genome (Fig. 19.7). This has been seen for comparison of the human genome with mouse [27, 85], dog [49], and chimpanzee [14]. Thus, estimates of constraint need to take into account the local rate variation.

For comparison of the human and mouse genomes [85], alignments throughout the genomes were evaluated for a level of similarity that exceeds the similarity expected from the amount of divergence in ancestral repeats in the vicinity. The distribution of similarity scores in ancestral repeats is normal, and many similarity scores in the bulk of the genome overlap with those in the neutral distribution (Fig. 19.8). Notably, a pronounced shoulder of alignments presents a score higher than the scores for a vast majority of ancestral repeats. The broad distribution of alignment scores through the genome can be interpreted as the combination of two distributions, one for neutral DNA and one for DNA that is under constraint. Various models lead to the conclusion that about 5% of the human genome falls into the latter distribution. A similar estimate has been obtained for alignments of the human and dog genomes [49]. In support of the idea

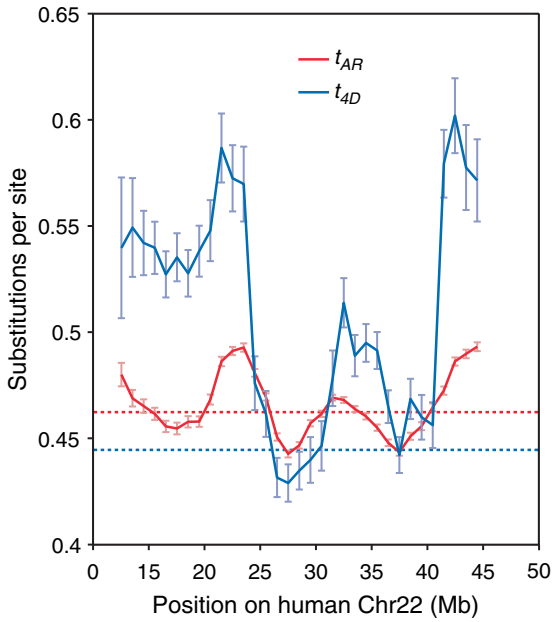

Fig. 19.7 Variation in the rate of human-mouse divergence in neutral DNA along human chromosome 22. The substitutions per site in ancestral repeats (t_{AR}, red) and in and in the subset of synonymous sites that are fourfold degenerate (t_{4D}, blue) were estimated in 5 Mb windows, overlapping by 4 Mb. The horizontal dotted lines indicate the estimates of t_{AR} and t_{4D} across the entire human genome. The confidence intervals are shown as brackets; the places where the confidence interval lies outside the genome-wide estimate are those with significant differences in evolutionary rate. (Reprinted from Waterston et al. [85], with permission from Nature Publishing Group)

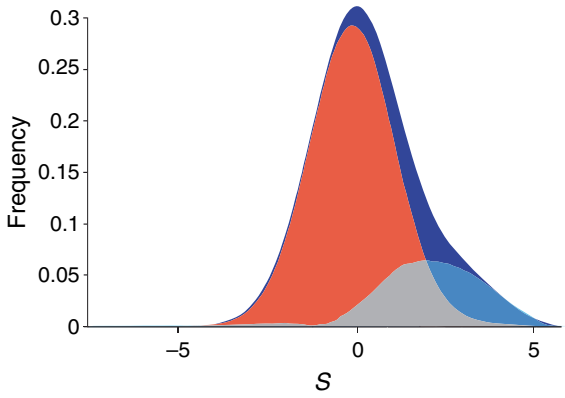

Fig. 19.8 Decomposition of conservation score into neutral and likely selected portions. *S* is the conservation score adjusted for variation in the local substitution rate. The frequency of the S scores for all 50 bp windows in the human genome, after alignment with mouse, is shown as the blue distribution. The frequency of S scores for ancestral repeats is shown in red. The inferred distribution of scores for regions under constraint is shown in light gray and light blue. This represents about 5% of the human genome. (Reprinted from Waterston et al. [85], with permission from Nature Publishing Group)

of a conserved eutherian core genome that encompasses the sequences with common function, the human sequences inferred to be under constraint are the same whether the comparison is with dog or mouse [49].

This result tells us that about 5% of the human genome has been under continuous purifying selection since the divergence of primates from carnivores and rodents, approximately 85–100 million years ago. The functions that would be subject to the continuous selection are those that were present in a eutherian ancestor and continue to play those roles in contemporary primates, rodents, and carnivores (and likely all eutherians). This is a lower bound estimate of the portion of the human genome that is functional. DNA sequences that have diverged for new functions in different lineages are not included in this estimate, nor are sequences that have acquired function recently through adaptive evolution. Thus, the portion of the human genome that is functional is certainly higher than 5%, but it is not possible with current knowledge to place an upper bound on the estimate.

The lower bound estimate of the portion under continuous constraint is a remarkable number. The portion of the human genome needed to code for proteins has been estimated at about 1.2%, with another 0.7% corresponding to untranslated regions of mature mRNA [33], giving an estimate of about 2% of the genome devoted to coding for mRNA. This leaves about 3% of the human genome with sequences that do not code for protein but still carry out functions common to eutherian mammals. Among these additional sequences under constraint should be genes for noncoding RNAs and DNA sequences that regulate the level of expression of genes. It is striking that the fraction of the genome devoted to the conserved noncoding functions is greater than the fraction needed to code for proteins.

19.3.3 Identifying Specific Sequences Under Constraint

In order to find particular functional sequences, it is necessary to identify specific sequences whose alignments are likely to be in the portion under constraint. In principle, it is a matter of finding segments with a similarity score above the neutral background (Fig. 19.2). Of course, it is important to adjust the analysis for variation in local substitution rate, as just discussed. For example, from the distribution of *S* scores in ancestral repeats (Fig. 19.8) based on pairwise human–mouse

alignments, one can compute a probability that a given alignment could result from the locally adjusted neutral rate. Those that are unlikely to result from neutral evolution between humans and nonprimates are likely to be under constraint.

Other measures have been developed to utilize the greater amount of information in multiple sequence alignments to identify constrained sequences. One measure is based on modeling the genome as having two states of "conservation," one that is effectively neutral and one that is the slowly changing, constrained state. By combining phylogenetic models with Hidden Markov models of those states, a score called *phastCons* is computed, which gives the posterior probability that any aligned position came from the constrained state [74]. This measure is routinely computed genome-wide for several sets of genome alignments, and is accessed as the "Conservation" track on the UCSC Genome Browser (Fig. 19.9). Note that it has a form similar to the idealized case in Fig. 19.2, with higher peaks associated with a greater likelihood of being constrained.

A constrained sequence is one that had an opportunity to change because it was mutated in an individual in a population, but the mutation was not fixed in the genome sequence of the species because of selective pressure against the change. Thus, there could have been a substitution, but purifying selection rejected it. Another measure of constraint, called genomic evolutionary rate profiling or GERP [16], explicitly models this process and estimates the number of "rejected substitutions" (Fig. 19.9). Another method, binCons, models the substitution frequency as a binomial distribution, with the contribution of alignments of different species weighted according to their phylogenetic distance from the reference species [52].

In a region evaluated by these methods, some segments are identified as being under constraint by all three, and others are found by only one. Each approach has value, and each has some unique advantages and some idiosyncratic problems. Thus, it is useful to combine the output of each to generate sets of "multispecies conserved sequences" [53, 79]. The strict, moderate, and relaxed sets correspond to the MCSs found by intersection, inclusion in at least two, or the union of the three sets. The example shown in Fig. 19.9 illustrates strong constraint not only in the coding exons but also in the introns. Experimental tests on two of these intronic constrained elements show that they affect the level of expression from a linked promoter [71].

19.4 Evolution Within Protein-Coding Genes

Comparative analysis of protein-coding genes requires several steps. First, a set of protein-coding genes must be defined in each species, and then a set of orthologous genes shared among the species is examined. With this, the rates of change among proteins can be computed and then one can study how those differences in rates correlate with function. Most protein-coding genes are under significant constraint over the course of mammalian evolution. However, genes whose products have roles in reproduction, chemosensation, immunity, and metabolism of foreign compounds are found consistently to be changing more rapidly than other genes. Thus, these are some of the functional classes that determine species-specific functions.

19.4.1 *Comparative Genomics in Gene Finding*

One of the most important tasks in genomics is to identify the segments of DNA that code for a protein. As covered in Chap. XX, most eukaryotic genes are composed of *exons*, which code for mRNA, and *introns*, which are transcribed but spliced out of the mature mRNA. Most internal exons encode a portion of the protein product of the gene, whereas the initial and terminal exons also contain untranslated regions of the mRNA. Most protein-coding exons can be identified by a variety of approaches. However, combining the exons into genes, including accurate determination of the initial exon (or multiple initial exons), is more of a challenge.

The several approaches for finding exons and genes can be divided into two categories: evidence-based and *ab initio*. Evidence-based methods find genomic DNA segments that align almost exactly with known protein sequences (after translating the genomic sequence) or complete mRNA sequences. Most evidence-based methods also incorporate data on

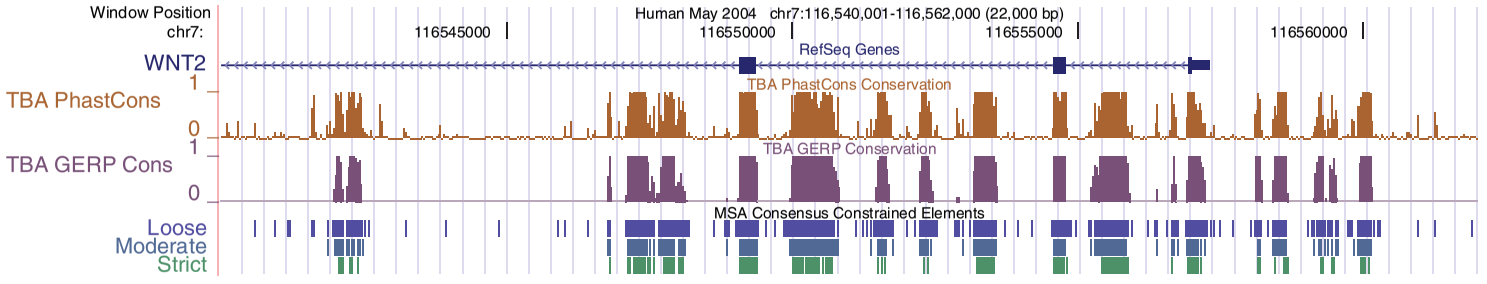

Fig. 19.9 PhastCons and GERP in a portion of ENCODE region ENm001. The first three exons (blue boxes) and introns (lines with arrows showing the direction of transcription from right to left) of the gene *WNT2* are shown on the top line. The next two panels plot the *phastCons* and *GERP* scores, respectively, with higher values indicating a higher probability that a sequence is under constraint. The bottom panel shows the levels of multispecies conserved sequences (see text)

expressed sequence tags (ESTs), which are short sequences containing portions of a very large number of mRNAs, and tags of sequence derived from the 5′ capped ends of mRNAs. The mRNA-coding segments of genomic DNA are grouped, using rules about pre-mRNA splicing signals, to find strings of exons that after splicing gives the mRNA sequence, or after splicing and translation gives the protein sequence. In order to find likely exons of genes whose mRNA sequences are not in the databases, *ab initio* methods based on models derived from basic knowledge about gene structure are applied. The genetic code and rules for splice junctions (Chaps. XX) provide the rules that make up the basic grammar for encoding proteins. Hidden Markov models such as those in the programs *genscan* [11] and *genmark* [28] are used to find likely exons and likely arrangements for these exons in genes.

Adding alignments of sequences of other species can improve gene prediction. Two commonly used methods are *Twinscan* [89] and *SGP* [87]; these build on the models in *genscan* but also apply rules from comparative approaches, such as allowing mismatches at degenerate sites in the genetic code. Another program, *exoniPhy* [73], uses the grammar of protein coding and a phylogenetic analysis of multispecies alignments to improve exon finding.

Often the initial and final exons do not code for protein, and thus the *ab initio* predictors no longer benefit from the well-known rules for encoding proteins. Furthermore, it is not uncommon for a gene to have multiple initial exons, with some used at particular times of development or in certain tissues. Thus, the accuracy of fully assembling genes from exons is enhanced by evidence such as mRNA sequences and tags derived from the 5′ ends of mRNA. Powerful pipelines for gene annotations have been developed that combine both evidence-based and *ab initio* methods; one of the most widely used is the Ensembl automatic gene annotation system [17].

In the current assembly of the human genome (NCBI build 36, March 2006, hg18), the Ensembl pipeline predicts 270,239 exons. These are arranged into 44,537 mRNAs from 21,662 genes. Most genes code for multiple mRNAs, thereby greatly increasing the diversity of proteins encoded in the human genome. Of these exons and genes, how many are found in other species, and which contribute to lineage-specific characteristics?

19.4.2 Sets of Related Genes

When discussing genes that are shared among species, we usually want to find the genes that are derived from the same gene in the last common ancestor. Homologous genes that separated because of a speciation event are *orthologous*. When there is a simple 1:1 relationship between orthologous genes, such as for *RRM1* in Fig. 19.10a, then any differences between the genes can be interpreted as changes since the time of divergence of the species.

When homologous genes are members of multigene families, then it is important to distinguish genes that have separated as a result of gene duplication (*paralogous* genes) from the orthologous genes, which separated by speciation events (Fig. 19.10a). For instance, the beta-like globin genes in humans arose by duplication in mammals. Within this gene family, each gene is paralogous to the other. For example, *HBE1* and *HBB* are paralogs that resulted from an earlier duplication, whereas *HBG1* and *HBG2* are paralogs that duplicated recently. Each of the four beta-like globin genes in chickens is paralogous to the other three, again because of the duplication history.

When gene duplications have occurred independently in both lineages, then all the duplicated genes in one species are orthologous to each of the genes in the other lineage. This is a many-to-many orthologous relationship. The human *HBB* gene is equally distant from each of the chicken beta-like globin genes, and it is orthologous to each.

Frequently a comparison will involve multigene families in species that share a duplication history, such as the beta-like globin gene clusters in human and macaque (Fig. 19.10b). The gene duplications outlined in panel A pre-date the catarrhine ancestor (ancestor to Old World monkeys, apes and humans). Thus, the *HBB* gene in humans is orthologous to the *HBB* gene in macaque, but it is paralogous to the other macaque beta-like globin genes, such as *HBD*, *HBG1*, etc. Likewise, the human *HBE1* gene is orthologous to the *HBE1* gene in macaque, but paralogous to the others. Comparisons between the orthologs reflect changes that have occurred since the separation of Old World monkeys and humans, whereas comparisons between the paralogs will reflect changes over a much greater phylogenetic distance, i.e., back to the gene duplications that generated the ancestors to the genes being

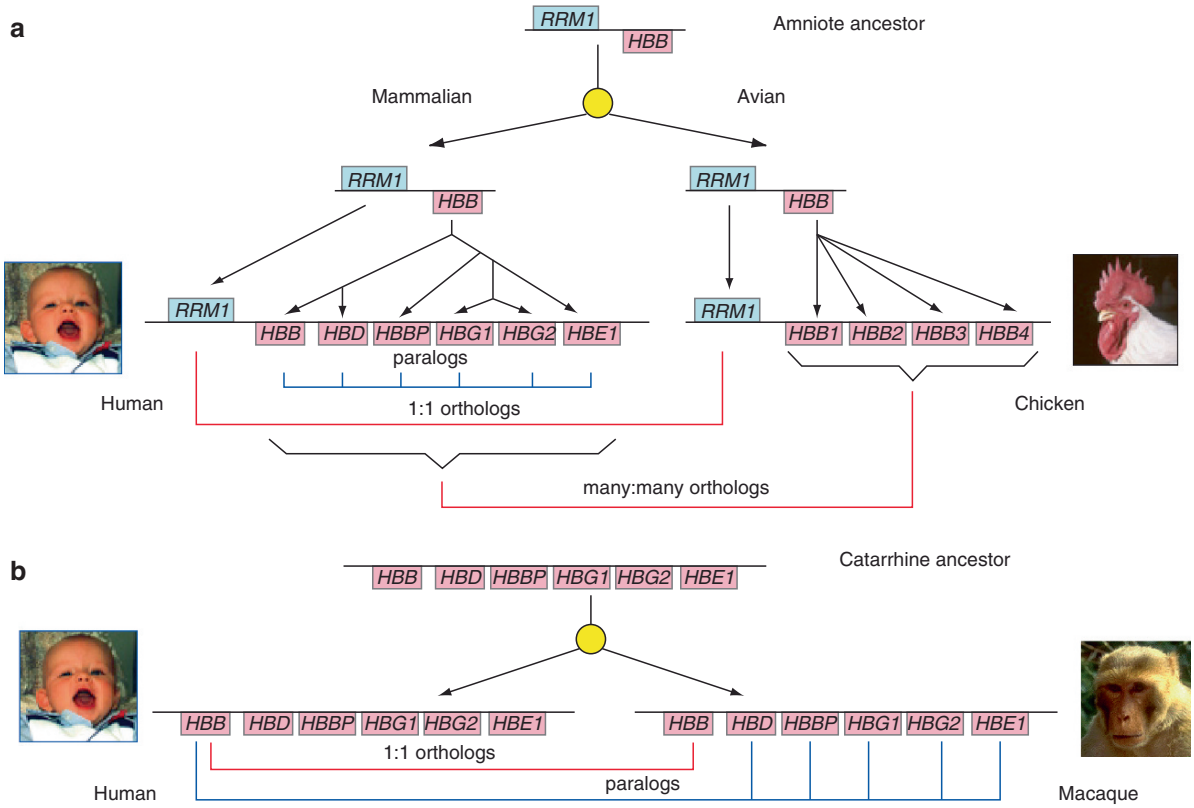

Fig. 19.10 Orthologous and paralogous relationships among genes. Speciation events are shown as yellow disks, and gene duplications are denoted by bifurcating arrows or multiple arrows with a single source. Red lines between genes in contemporary species connect orthologous genes, whereas blue lines connect paralogous genes. (**a**) Illustration of the phylogenetic history of the *RRM1* gene (encoding ribonucleotide reductase M subunit) and the *HBB* gene (encoding beta-globin) and genes related to it by duplication since the divergence of mammalian and avian lineages from the amniote ancestor. The gene duplications in the beta-like globin gene family occurred separately in the mammalian and avian lineages, leading to paralogous relationships within a species and many-to-many orthologous relationships between the species. (**b**) Illustration of the the phylogenetic history of the beta-like globin gene cluster over the much shorter time since humans and macaques (an Old World monkey) diverged from the catarrhine ancestor. The gene duplications predate the ancestor, and thus the speciation event resulted in 1:1 orthologous relationships between human and macaque *HBB*, human and macaque *HBD*, etc. Other relationships, e.g., between human *HBB* and macaque *HBD* are paralogous

compared. In this situation, correct assignments of paralogous and orthologous relationships are particularly important. For instance, an incorrect assignment of paralogous genes as being orthologous between human and macaque would lead to a conclusion of greater sequence change since speciation than would a truly orthologous comparison.

Once gene sets have been defined in two or more species, then orthologous gene sets can be determined. For the cases of 1:1 orthologs, reciprocal highest similarity is a good guide to orthologous relationships. The more complicated cases for multigene families can be summarized as many-to-many orthologous relationships. Figure 19.11 shows the results of comparisons of protein-coding genes among human (*Homo sapiens*), chicken (*Gallus gallus*), and the teleost fish *Fugu rubripes* [30]. Of the almost 22,000 genes annotated in humans in this study, about a third are in 1:1:1 orthologous relationships with chicken and *Fugu*, and about 5% are in many-to-many relationships. About a third of the genes have clear homologs but cannot be definitively assigned as orthologous. Intriguingly, about 4,000 human genes do not have a clear homolog in either chicken or fish. These may encode mammal-specific functions.

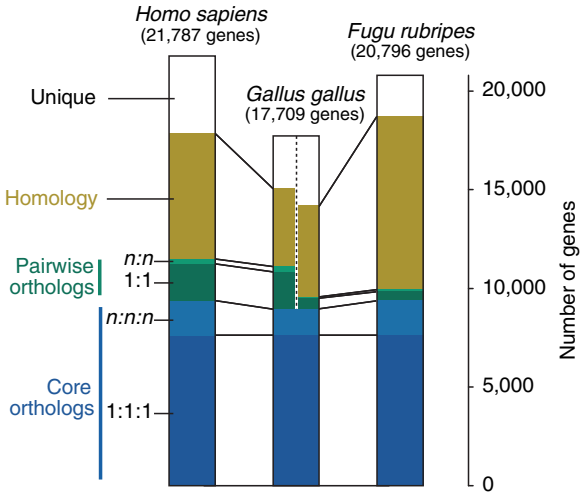

Fig. 19.11 Homology relationships among protein-coding genes in human (*Homo sapiens*), chicken (*Gallus gallus*), and the fish *Fugu rubripes*. Genes in the three species are grouped by their orthology relationships among the three species (1:1:1 or *n:n:n* for many:many:many) or between two species if the gene is not detected in a third species. Genes that are clearly related between species but for which clear orthology relationships cannot be determined are placed in the 'Homology' class. Genes not falling in the orthology or 'homology' classes are considered 'Unique'. (Reprinted from Hillier et al. 2004, with permission from Nature Publishing Group)

19.4.3 Rates of Sequence Change in Different Parts of Genes

Within the set of 1:1 orthologous genes, the amount of sequence similarity can be determined in each of the basic parts of a gene. One of the first genome-wide studies in mammals compared human genes with mouse genes [85], and it confirmed many insights from smaller scale studies. The protein-coding exons are the most similar between human and mouse, showing about 85% identity (Fig. 19.12). The regions adjacent to the splice junctions show peaks of higher identity, reflecting the selection on both coding potential and on splicing function. The introns have the lowest similarity, but they are considerably more similar than is DNA in ancestral repeats (the neutral model in this study), which are about 60% identical. The untranslated regions of exons are about 75% identical. The higher percent identity in the untranslated regions and introns, than in the neutral model, indicate that some portion of these sequences is under constraint. Intronic regions that provide important functions include splicing enhancers and transcriptional enhancers. In the 3′ untranslated region can be found targets for regulation by miRNAs as well as the polyadenylation signals. These short segments can be subject to stringent constraint. If all the intronic and untranslated sequences were subject to such stringent constraint, then their overall percent identity would be closer to that of the coding regions. Thus, one interpretation of these results is that intronic and untranslated regions contain short constrained segments interspersed within larger regions with little or no signature of purifying selection.

19.4.4 Evolution and Function in Protein-Coding Exons

From the earliest comparisons of homologous protein sequences, it was recognized that some proteins change little between species. A classic example is histone H4,

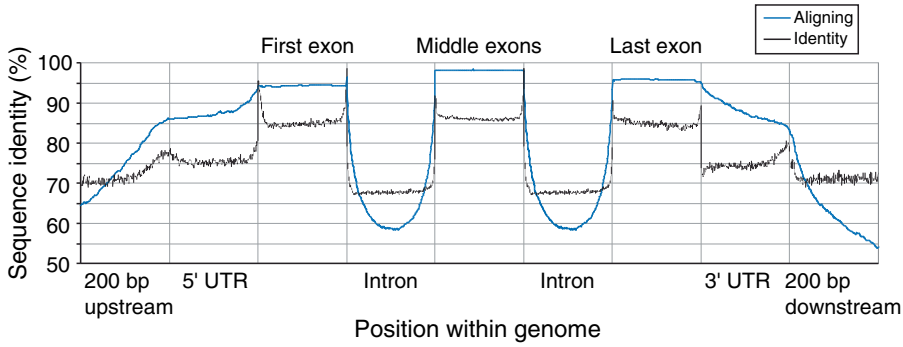

Fig. 19.12 Sequence identity between human and mouse in a generic gene. Within a group of 3,165 RefSeq genes that aligned between the mouse and human genomes, 200 evenly spaced bases across each of the variable-length regions were sampled between human and mouse. The blue line shows the average percentage of bases aligning and the black line shows the average base identity. (From Waterston et al. [85], with permission from Nature Publishing Group)

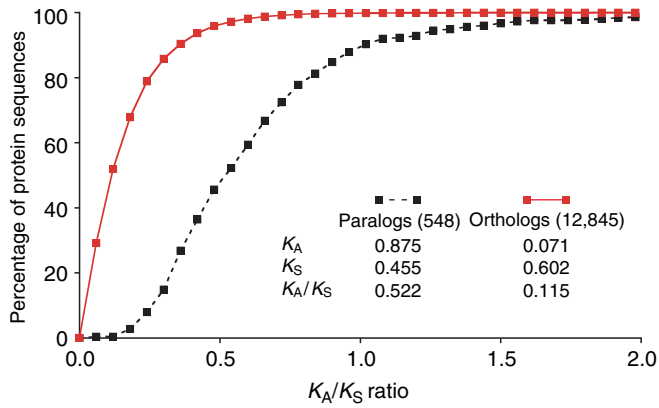

Fig. 19.13 Cumulative distribution of K_A/K_S values for mouse proteins compared with human homologs. The distribution of scores for proteins that are clearly orthologous between human and mouse is shown by the red points and line. The distribution of scores for proteins encoded by locally duplicated, paralogous mouse-specific gene clusters is shown by the black points and line. (From Waterston et al. [85], with permission from Nature Publishing Group)

which has only one amino acid replacement between peas and cows. Other proteins change rapidly. Among the most rapidly changing proteins are the fibrinopeptides, which are segments of fibrinogen molecules that are cleaved off by thrombin during blood clotting. It appears that the amino acid sequence of the fibrinopeptides is not critical for their function, and they are under little or no selective pressure. Interspecies comparisons of even a modest number of proteins showed that the rate of changes in amino acids ranged over 100-fold [60]. Some proteins, such as histones, are under stringent selection over most of their sequence, whereas others seem to be free to change extensively – or have been adapted to new function.

Comparisons of the protein-coding genes for entire mammalian genomes provide the opportunity to examine these issues more comprehensively. The sets of related genes between species can be analyzed to show which genes are under strong purifying constraint and which show signs of adaptive evolution. For protein-coding genes, it is common to consider substitutions at synonymous sites to be neutral. The number of synonymous substitutions per synonymous sites in two species is called K_S. This can be used as an estimate of the neutral rate. Then the number of nonsynonymous substitutions per nonsynonymous site, or K_A, can be compared with K_S to obtain an estimate of the stringency of the purifying selection or the strength of adaptive evolution. As a rule of thumb, a K_A/K_S ratio of 0.2 for human–mouse comparisons is indicative of constraint, whereas ratios of 1 or greater indicate adaptive evolution.

In a study of orthologous genes aligned between mouse and human [85], about 80% show an overall signal for constraint (Fig. 19.13). Very few show evidence of positive selection over their entire length. Thus, at the phylogenetic distance of mouse and human, evolution of protein-coding sequences in orthologous genes is dominated by constraint. This result indicates that the matching, orthologous segments code for proteins that provided a function in the ancestor, and their descendant sequences provide a similar function in contemporary species. Many changes in the encoded amino acid sequences have been selected against because they did not improve the function of the protein. We note that short segments or single codons under positive selection would not be detected in this test.

In contrast, the set of paralogous genes compared between mouse and human are shifted to higher K_A/K_S ratios. Thus, the paralogous genes are more likely to be undergoing adaptive evolution (positive or diversifying selection) than are the orthologous genes. The multigene families are major contributors to lineage-specific function. Duplication of genes leaves at least one copy free to accumulate changes that can provide an adaptive advantage. In contrast, genes that remain as single copies are constrained to fulfill the role that they have played since they arose in some distant ancestor.

19.4.5 *Fast-Changing Genes That Code for Proteins*

The families of fast-changing genes appear to be adapting to new pressures in a lineage-specific manner. An examination of the types of gene families with this property should provide insights into the types of pressures that lead to adaptive changes. A remarkably consistent result has been found in multiple studies

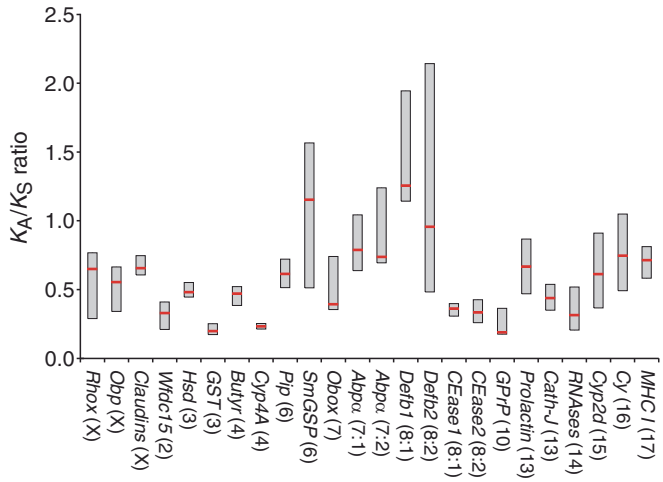

Fig. 19.14 Distributions of K_A/K_S values for duplicated mouse-specific gene clusters. The chromosome on which the clusters are found is indicated in brackets after the abbreviated cluster name. The K_A/K_S values for each sequence pair in the cluster were calculated from aligned sequences. The box plots summarize the distributions of these values, with the median indicated by the red horizontal line and the boxes extending from the 16th and 83 rd percentiles and hence covering the middle 67% of the data. (From Waterston et al. [85], with permission from Nature Publishing Group)

of this question. The four general categories of reproduction, chemosensation, immune response, and xenobiotic metabolism (breakdown of drugs, toxins, and other compounds not produced in the body) encompass many of the genes and gene families subject to positive selection. Thus, these are the major physiological functions in which rapid sequence change leads to adaptive evolution.

For example, the locally duplicated gene families with relatively high K_A/K_S values fall into distinct functional classes (Fig. 19.14). Members of the major categories for adaptive evolution (reproduction, chemosensation, immune response, and xenobiotic metabolism) are apparent. For example, the mouse *Rhox* genes on chromosome X are homeobox genes expressed in male and female reproductive tissue, and targeted disruption of the *Rhox5* gene leads to reduced male fertility [51]. Another example is the oocyte-specific homeobox gene *Obox* on mouse chromosome 7. The *Obp* gene cluster encodes odorant-binding proteins such as lipocalins and aphrodisin, involved in both chemosensation and reproduction. Immune response genes include the *MHC I* genes on chromosome 17, which regulate the immune response, the *Wfdc15* gene, which encodes an antibacterial protein, and the *Defb* genes on chromosome 8 encoding beta-defensins. Several adaptive genes are involved in xenobiotic metabolism, including members of the cytochrome P450 gene family, *Cyp4a* and *Cyp2d*, and a glutathione-*S*-transferase gene (*GST*).

Additional studies of lineage-specific expansions of gene families in comparisons of rat and mouse [25] and of humans and chickens [30] identify the same general categories of reproduction, chemosensation, immune response, and xenobiotic metabolism. Thus, along multiple lineages, these gene families are implicated in adapting to unique pressures on each species. Enrichment of these functional categories for genes implicated in adaptive evolution can be readily rationalized. Changes in genes involved in reproduction and chemosensation could lead to or maintain the differences that cause divergence of species. Adaptation of immune function and the ability to metabolize foreign compounds are important for survival in the distinctive environment of each species. Other families with rapid changes between species include keratins, which are involved in making feathers in birds but hair in mammals.

19.4.6 Recent Adaptive Selection in Humans

In addition to improving our understanding of the evolution of humans within the context of other vertebrates, comparative genomics also provides insights

into recent adaptive changes that may eventually tell us what genome sequences make us distinctively human. Comparisons to close relatives such as the chimpanzee and analysis of human polymorphisms drive these new studies.

As was the case for human–mouse comparisons discussed above, the K_A/K_S ratio was computed in genome-wide comparison of the human and chimpanzee gene sets [12, 14, 15, 61]. The ratio for human–chimpanzee comparisons is significantly higher than that seen for mouse–rat comparisons, showing more changes in amino acids in proteins (normalized to synonymous substitutions) in the hominid lineages than in rodents. This does not, however, indicate an overall stronger positive selection in hominids, but rather it reflects the relaxation of purifying selection in species with a small population size. Estimates of effective population size for rodents far exceed those for humans and chimpanzees, and it is well recognized that the severity of selection increases with population size. However, despite this relaxed selection, examination of the orthologous genes with the most extreme ratios of amino acid-changing substitutions to presumptive neutral changes reveals interesting candidates for hominid-specific adaptive evolution. One is the gene for glycophorin C, which is the membrane protein used for invasion of the malarial parasite *Plasmodium falciparum* into human erythrocytes. Others include granulysin, which is needed for defense against intracellular parasites, and semenogelins, which are involved in reproduction. A stronger signal for positive selection can be observed when genes are grouped together, either by physical proximity (often as duplicated genes) or by functional category. For human-chimpanzee comparisons, the sets of genes changing most rapidly include the now-familiar categories of reproduction (e.g., spermatogenesis, fertilization, and pregnancy), chemosensation (olfactory receptors, taste receptors), immunity (immunoglobulin lambda, immunoglobulin receptors, complement activation), and xenobiotic metabolism, plus additional categories such as inhibition of apoptosis.

The distribution of human polymorphisms along chromosomes and their frequency in populations can be analyzed for insights into very recent selection (reviewed in [5, 44]). Positive selection is expected to drive mutations quickly to fixation, so loci under positive selection should be characterized by a skew in the allele frequency distribution toward rare alleles. One measure of that skew is Tajima's D [77]. Also, the rapid fixation of an advantageous allele will bring along linked polymorphisms. These polymorphisms will not have had time to be separated from the selected allele by recombination, and thus linkage disequilibrium will extend further around positively selected alleles than is expected from neutral evolution. Various tests of properties such as these have been developed, and have traditionally been applied to a small number of loci. A major limitation to these studies is that changes in population demographics can generate the same signals. For example, recent expansion in population size, such as that experienced by humans, will also lead to an excess of rare alleles or extended linkage disequilibrium. Thus, it is difficult to disentangle the confounding effects of population demographics and positive selection when only a few genetic loci are examined. However, the recent availability of genome-wide data on polymorphisms [32] provides one solution. Changes in population size should affect all loci in the genome, whereas selection should act on only a few. Thus, when the distribution of values for Tajima's D, long-range haplotype, or related measures are examined for a large number of loci, then it is likely that the outliers are undergoing adaptive evolution [5].

Recent genome-wide studies have identified significant outliers based on frequency of rare alleles (Tajima's D, [13, 37]) and linkage disequilibrium [82, 83]. For example, Carlson et al. [13] calculated Tajima's D in sliding windows across the human genome for populations descended from Africans, Europeans, or Chinese. Several extended regions with consistently negative values for Tajima's D were identified, with most observed in only one of the populations (Fig. 19.15). Negative values for Tajima's D are associated with positive selection if population expansion is not a factor, and the study design to identify outliers in a genome-wide analysis should greatly reduce the confounding effect of such an expansion. Thus, results such as those in Fig. 19.15 indicate that at least one genetic element in the roughly one megabase region with reduced Tajima's D has been under positive selection in humans of European ancestry. Resequencing of targeted genes within these regions has supported the conclusion of positive selection, and in some cases (e.g., *CLSPN* in Fig. 19.15) it has revealed a polymorphism that alters the encoded amino acid sequence [13]. Such a change in amino

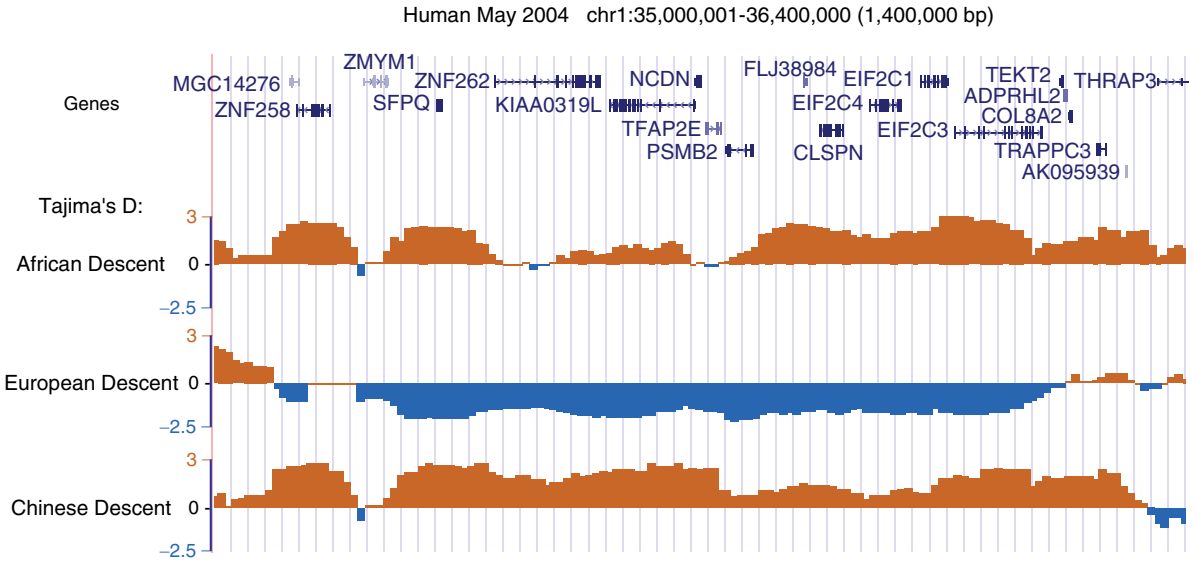

Fig. 19.15 An extended region with an excess of rare alleles indicative of positive selection. The region from human chromosome 1 is one of several identified in the study by Carlson et al. [13] showing an excess of rare alleles in at least one of three human populations (those of European descent in this case) as measured by Tajima's D [77]. Negative values of Tajima's D can be explained by positive selection or population expansion; the design of genome-wide studies favors the former explanation. The full data from the study are available on the UCSC Genome Browser [39]; this figure was generated from the Browser output

acid sequence is a candidate for the functional variant under selection.

A third type of test for recent selection utilizes both human polymorphism data and interspecies divergence between human and close relative, such as chimpanzee. The McDonald-Kreitman [55] test compares the ratio of polymorphisms to divergence (r_{pd}) at nonsynonymous sites (leading to amino acid changes in the protein product) with that ratio in synonymous sites, which do not change the amino acid sequence and are expected to be largely neutral. If the changes in nonsynonymous sites had no selective advantage or disadvantage, then r_{pd} at these sites would not be significantly different from r_{pd} at neutral sites. Deviation from neutral expectation can be evaluated with a chi-square or related statistic. Bustamante et al. [12] applied this test to over 11,000 human genes (with polymorphsims determined in three different populations) compared with chimpanzee. They found that 9% had a significant signal for positive selection and 14% had a significant signal for negative selection.

Each method for finding loci under recent selection in humans has its distinctive strengths and weaknesses. Much effort is currently devoted to examining overlaps and differences in the results. Among the several studies reviewed by Biswas and Akey [5], a total of 2,316 human genes were found to have at least one signature for positive selection. Almost a third of these, including *EDAR, SLC30A9*, and *HERC1*, are found in more than one genome-wide study. Other candidate genes for positive selection are found by only one approach, such as *TRPV5* and *TRPV6*. At least to some extent, the failure to overlap reflects the different types of selective events being assayed in the different tests. The features examined by one approach, such as low frequency alleles, are not contributing to other tests, such as linkage disequilibrium measurements based on common alleles [5].

Some genes that are candidates for human-specific selection lead to intriguing and exciting possibilities, such as alterations in *FOXP2* implicated in language acquisition [22] and *MCPH1* and *ASPM* implicated in brain size [23, 56]. Further studies of recent selection in humans should lead to critical new insights into human biology and disease.

19.4.7 Human Disease-Related Genes

Comparative genomics can be used to study the origins and implications of genetic variants associated with human disease. Disadvantageous mutations should be

cleared from a population quickly, so why are some genetic diseases rather common?

One factor is the relaxed selection against mildly deleterious alleles resulting from population expansion. A common estimate of the effective population size of humans is about 10,000 individuals, and of course the population has expanded dramatically to the current level of over 6 billion. This would tend to favor the persistence of some deleterious mutations, and the results of a McDonald-Kreitman test [12] indicate that many of the amino acid polymorphisms in humans are moderately deleterious.

Another factor is positive selection in one region of the world driving an allele to high frequency, but that allele is pathogenic in other regions of the world. A classic example is the *HBB-S* allele of the gene encoding beta-globin. This allele encodes a mutant beta-globin that in combination with alpha-globin and heme constitutes HbS. This is the hemoglobin variant that causes red blood cells to form a sickled, inflexible morphology when deoxygenated, and thus leads to sickle cell disease. However, the *HBB-S* allele in heterozygotes reduces the susceptibility of humans to malaria, and thus it is a protective allele in regions of the world in which malaria is endemic. In fact, haplotype analysis has shown that the *HBB-S* allele has arisen independently multiple times in recent human history [2, 63]. This indicates a strong positive selection in the presence of the malarial parasite. Unfortunately, the negative consequence is that people who are homozygous for the *HBB-S* allele are highly prone to sickle cell disease.

A third factor is that some disease-associated variants were protective in the more distant past but are now detrimental for most contemporary human lifestyles. In the "thrifty genotype" hypothesis [59], the limited caloric intake and need for high activity levels in ancestral humans would have favored a thrifty genotype that made efficient use of food. However, many contemporary humans live in an environment with an excess of available food. Being "too thrifty" with energy metabolism could lead to problems such as diabetes. Disease-associated variants that were advantageous in the past should match the amino acid at that position in ancestor, and some of these will still be seen in related species. Indeed, human disease-related variants match with the amino acid in the corresponding position of chimpanzee [14] and rhesus macaque [70] in about 16 and 200 cases, respectively. Further studies of these candidates are needed, but the results suggest that retention of an ancestral state is also contributing to human disease alleles.

19.5 Evolution in Regions That Do Not Code for Proteins or mRNA

Despite the importance of protein-coding regions to genome function, these sequences account for about one-third of the sequences that have been under selection for a common function in eutherian mammals. Accounting for the remaining selection in noncoding regions is a major on-going effort in genomics and genetics. Two functional categories are the focus of much attention: genes that do not code for proteins, such as microRNA (miRNA) genes, and gene regulatory regions. An equally important question is to what phylogenetic depth functional noncoding regions are conserved. These issues will be examined in this section.

19.5.1 Ultraconserved Elements

The level of constraint on genomic sequences spans a wide range, and it likely that different functions are subject to distinctive levels of constraint. The most intense constraint is revealed in the human DNA segments called *ultraconserved elements*, or UCEs [4]. These are the 481 human DNA segments that are identical to mouse DNA for at least 200 nucleotides. Sequences that code for proteins have frequent mismatches between human and mouse at synonymous sites, so these UCEs are under stronger purifying selection than most exons. This pattern of conservation indicates that all nucleotides in the identical segment are critical for some function. The UCEs are broadly conserved in vertebrates, and they show the slowest rate of divergence over the period of vertebrate evolution of any known elements in the genome (Table 19.2, Fig. 19.5).

Determining the roles for the UCEs is currently a matter of intense interest. Only a small fraction (23%) overlaps with mRNA for known protein-coding genes. Thus, the majority is associated with some noncoding

function. About half of those tested serve as tissue-specific enhancers in transgenic mouse embryos [64]. A small number are related to each other, and examination of these has revealed a family of sequences derived from an ancient transposable element that have been recruited for activity as a distal enhancer for one gene and part of an exon for another [3]. Another subset of very slowly changing regions (across most eutherians) was examined for rapid change along the human lineage since divergence from chimpanzee. These *human accelerated regions* include a gene that encodes an RNA that may function in cortical development [66]. A full explanation of the stringent constraint on each nucleotide within the UCEs remains elusive. Not only is the intensity of constraint beyond that seen for almost all protein-coding regions, but even RNAs with considerable secondary structure rarely show this resistance to substitution.

Another enigmatic aspect to UCEs is their restriction to vertebrates. Protein sequences, which evolve faster than UCEs in vertebrates, frequently show significant similarity between vertebrate and invertebrates species. Sometimes the similarity extends from vertebrates to eubacteria. In contrast, no homolog to a UCE sequence has been observed outside vertebrates. Worms (and possibly other invertebrates) have analogous highly constrained noncoding sequences, but they differ in sequence from the vertebrate UCEs [81]. Thus, this stringent constraint on noncoding sequences may have evolved in parallel in vertebrates and invertebrates. Finding the sources of the UCEs and explaining how they could be under such intense constraint are important goals for future work. Answers to these questions may reveal aspects of genome function that have yet to be imagined. The fact that the roles and origins of the most stringently constrained sequences in vertebrates are still unknown illustrates how much still needs to be accomplished in comparative genomics.

19.5.2 Evolution Within Noncoding Genes

Many genes do not code for protein, and these must account for some of the noncoding DNA that is under constraint. However, some of the better-known noncoding genes do not help explain the fraction under constraint, but for technical reasons. Consider the genes for RNAs utilized in the mechanics of protein synthesis, such as ribosomal RNAs (rRNA) and transfer RNAs (tRNAs). The rRNA genes are clustered in highly duplicated regions on the short arms of chromosomes 13, 14, 15, 21, and 22. These regions are not included in the assemblies of the human genome, and thus they do not contribute to the minimal estimate of 5% of the genome under constraint in mammals. The tRNA genes are small and contribute little to the selected fraction. Other RNAs, such as snRNAs involved in splicing and processing of precursors to mRNA, also tend to be encoded on small genes. Multiple copies of sequences related to the snRNA genes are present in the human genome, some of which may no longer be active. The contribution of snRNA genes to the fraction of the human genome under constraint needs further study.

The miRNAs do not code for protein, but they negatively regulate mRNA function or abundance. Hybridization of an miRNA to its mRNA target to generate a duplex with some mismatches leads to inhibition of translation of the mRNA. Hybridization of an miRNA to its target to generate a perfect duplex leads to degradation of the target mRNA (see Chap. XX).

The known miRNA genes are constrained, with many conserved from humans to chickens. However, the full set of miRNA genes is not known, and information is limited about the structure and conservation of genes encoding the precursors to miRNAs. Thus, the miRNAs clearly are important contributors to the fraction of the genome under purifying selection, and they could account for substantially more of the constraint that is currently known.

Members of another class of RNA that apparently does not code for protein are detected by hybridization of copies of cytoplasmic RNA to high-density tiling arrays of nonrepetitive human genomic DNA. These results show transcription of protein-coding genes as expected, but about half the transcribed regions are not associated with known genes [34]. These unannotated transcripts, referred to as *transfrags*, are often of low abundance and are expressed in a limited set of tissues. The contribution of transfrags to constrained sequences in human is a matter of current study (e.g., [67, 79]).

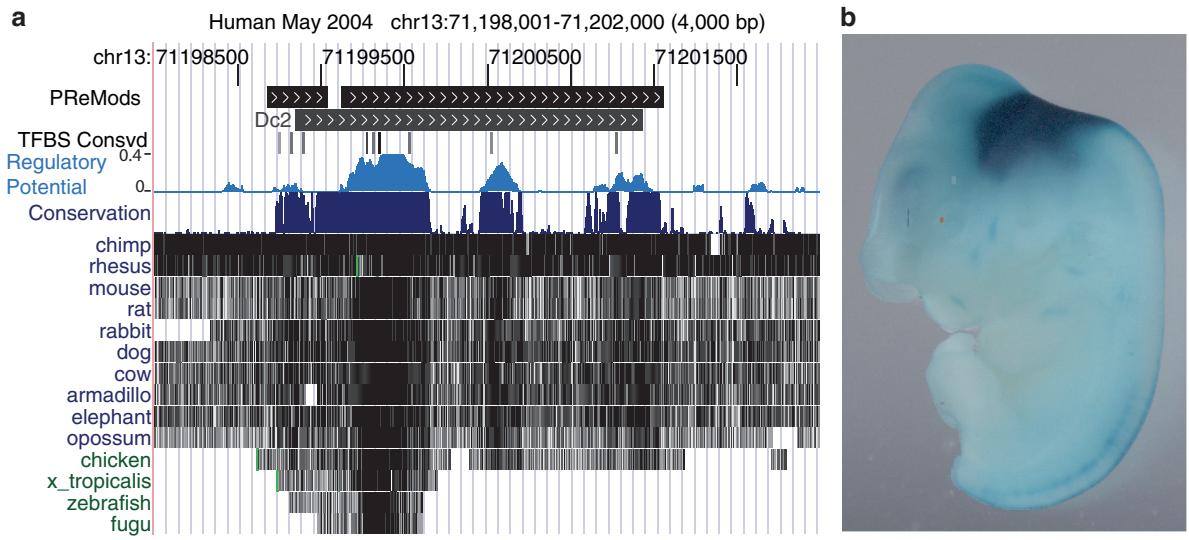

Fig. 19.16 An enhancer of the *DACH1* gene predicted by comparative genomics. This human gene is homologous to the *Drosophila* gene *dachshund*, and it is needed for development of the central nervous system and other organs. Within one of the very large introns of *DACH1* are some deeply conserved DNA segments. (**a**) Several features of the Dc2 region, including its conservation from humans to fish, high regulatory potential [78], and prediction as a regulatory module by the PReMod pipeline [6]. Examples of conserved matches to transcription factor binding site motifs are also shown. (**b**) This DNA segment is sufficient to enhance expression of a beta-galactosidase reporter gene in the hindbrain of a transgenic mouse embryo. The blue stain is a marker for beta-galactosidase activity. The Dc2 region was shown to be an enhancer by Nobrega et al. [62]; the image is from the Enhancer Browser (Table 19.3)

19.5.3 Evolution and Function in Gene Regulatory Sequences

DNA sequences needed to regulate the level, developmental timing, and tissue-specificity of gene expression include promoters that designate the correct start site for transcription, enhancers that increase the level of expression, silencers that decrease the level of expression, and insulators that separate genes and regulatory regions from the effects of neighboring regulatory regions. Many but not all of these regulatory regions are conserved among mammals [26, 64]. Some of the DNA sequences that regulate genes encoding developmental regulatory proteins are conserved from mammals to fish, indicative of strong constraint [62, 88]. One example is shown in Fig. 19.16. However, other regulatory regions show more rapid evolution, e.g., replacing one motif for binding a transcription factor with a similar sequence in another location [19, 50] or being present in only one lineage. Despite numerous studies of the extent of conservation of regulatory regions in individual loci, no clear consensus had emerged on the dominant pattern of conservation.

A major limitation to previous studies has been the small number of regulatory regions that have been identified experimentally. Establishing the role of a segment of DNA in regulation requires multiple experiments, and traditionally these were done in a highly directed manner that did not lend itself to high throughput. Now it is possible to enrich DNA for sites occupied by transcription factors (by chromatin immunoprecipitation or ChIP) and then hybridize this enriched DNA to high-density tiling arrays of genomic DNA (DNA chips). This ChIP-chip experiment [69] reveals sites bound by transcription factors in a high-throughput manner. Experiments by the ENCODE Project Consortium [79] evaluating sites occupied by several transcription factors have yielded a large set (over 1,000) of putative transcriptional regulatory regions in about 1% of the human genome. This large set of DNA intervals implicated in transcriptional regulation was identified by experiments that are agnostic to interspecies sequence conservation, and thus it is

an ideal set in which to determine the phylogenetic depth of conservation [42]. As shown in Fig. 19.5 and Table 19.2, about two out of three of these putative transcriptional regulatory regions are conserved from humans to other placental mammals (but no further), and about one out of three are conserved to marsupials. Less than 10% are conserved from humans to birds. An equal fraction, about 3%, is found at the two extremes of conservation, viz., found only in primates or conserved from humans to fish. Thus, the bulk of the regulatory regions are conserved in placental mammals, and we expect that comparisons among these species will continue to be effective at finding and better understanding these regulatory regions. However, a particular phylogenetic depth of conservation is not a consistent property of gene regulatory sequences. Rather, the depth of conservation is a property that varies among the regulatory sequences. Ongoing studies may reveal whether particular functions of regulatory regions or their targets correlate with the depth of conservation.

Although it is not a property shared by all putative regulatory regions, many do have a significant signal for purifying selection. A small majority (about 55%) overlap at least in part with DNA segments that are in the 5% of the human genome that is under strong selection [79]. However, only about 10% of the nucleotides in the putative regulatory regions are under strong constraint, suggesting that small subregions of enhancers and promoters, e.g., binding sites for particular transcription factors, are under purifying selection. Thus, the putative regulatory sequences identified in the ENCODE project contribute only a small amount to the 5% under strong constraint [79].

19.5.4 Prediction and Tests of Gene Regulatory Sequences

Effective use of comparative genomics to find gene regulatory sequences is challenging for at least two reasons. The variation in phylogenetic depth of conservation is a major complication; some human regulatory regions will be observed only in alignments of primates, whereas others align with species as distant as fish. Although the large majority of regulatory regions are conserved in multiple placental mammals, even some apparently neutral DNA aligns reliably at this phylogenetic distance. Thus, the ability to align at this distance is not a property that identifies regulatory regions with good specificity.

Most efforts to detect candidate gene regulatory regions from aligned sequences also use some form of pattern information. For example, the known regulatory regions are clusters of binding sites for transcription factors. The binding sites are short (about 6–8 bp) and many allow degeneracy (e.g., either purine or either pyrimidine works equally well at some sites). Therefore, the binding site motifs themselves do not confer strong specificity. However, in combination with clustering and conservation, this set of criteria has good power to detect novel regulatory regions [6]. A set of about 200,000 regions, called *PReMods*, has been identified as predicted regulatory regions in the human genome using this approach.

The motifs for binding sites in regulatory regions are not known completely. These currently unknown motifs can be incorporated into the prediction of regulatory regions by using machine-learning procedures to find distinctive patterns of alignment columns that are common in a training set of alignments in known regulatory regions, but are less abundant in a set of alignments from likely neutral DNA. The statistical models describing these distinctive patterns are then used to score any alignment for its *regulatory potential*. One implementation of this approach has generated a set of about 250,000 regions of human DNA with a high regulatory potential [78]. Many of these overlap with the PReMods discovered as conserved clusters of transcription factor-binding motifs. Regions with high regulatory potential and a conserved binding site for an erythroid transcription factor are validated at a good rate as enhancers in erythroid cells [84].

In summary, several methods based on comparative genomics can be used with some success to predict gene regulatory sequences, but none achieves the level of reliability desired. Deep conservation of noncoding sequences, e.g., from human to chicken or human to fish, can be used without additional information about patterns such as binding site motifs. However, this approach will miss the majority of gene regulatory regions. For noncoding sequences conserved among placental mammals, clustering of pattern information should be incorporated. The pattern information can either be based on prior knowledge

(such as binding motifs) or learned from training sets. Currently, in vivo occupancy of DNA segments by transcription factors is being determined comprehensively by ChIP-chip and related methods. Integration of this information with the comparative genomics should add considerable power to the identification of regulatory regions [21].

19.6 Resources for Comparative Genomics

The large amount and wide variety of data on comparative genomics of mammals and other species can be daunting to those who wish to use them. Also, as discussed throughout this chapter, the level of conservation of functional regions tends to vary from region to region. Detailed information needs to be readily accessible for individual regions and for classes of features across a genome. These needs are accommodated by genome browsers and data marts. Computational tools for further analysis of the data are also available, and one workspace for such tools will be described here.

19.6.1 Genome Browsers and Data Marts

Genome browsers show tracks of user-specified information for a designated locus in a genome. The major browsers for mammalian genomes are the UCSC Genome Browser [45], Ensembl [31], and MapView at NCBI [86] (Table 19.3). Comparative genomics tracks showing results of whole-genome alignments are available at the UCSC Genome Browser and Ensembl. As illustrated in Fig. 19.4, the regions of the human genome aligning with a comparison species can be seen as nets and chains. Inferences about severity of constraint are captured on the "Conservation" track (similar to that in Fig. 19.9), based on phastCons [74].

Often it is desirable to collect and analyze all members of a feature set across a genome or large genomic intervals. This requires the ability to query on the databases of features that underly the browsers. Two such "data marts" are the UCSC Table Browser [35] and BioMart at Ensembl [36]. Both provide interactive query pages to provide access to the data.

19.6.2 Genome Analysis Workspaces

Once the data have been obtained, users frequently need to analyze them further. Different data sets may need to be combined or compared. The level of constraint or regulatory potential may be needed. Estimates of evolutionary rates may be desired. Different tasks will require distinct sets of tools. Considerable progress can be made by acquiring the necessary computer programs and executing them on the user's computer system. However, this leaves it to the user to find or write the needed tools.

An alternative is to connect versatile data acquisition with integrated suites of computational tools in a common workspace such as Galaxy [7] (Table 19.3). This resource allows users to import data from various sources, such as the UCSC Table Browser, BioMart, or files from the user's computer. Once imported, a wide variety of operations can be performed on the data sets, such as edits, subtractions, unions, and intersections. Summary statistics can be computed and distributions can be plotted. Various evolutionary genetic analyzes can be performed.

Table 19.3 Data resources and analysis workspaces for comparative genomics

Name	Description	URL
UCSC Genome Browser	Sequences, comparative genomics, annotations	http://genome.ucsc.edu
Ensembl	Sequences, comparative genomics, annotations	http://www.ensembl.org/
NCBI MapViewer	Gene, EST and other maps of chromosomes	http://www.ncbi.nlm.nih.gov/mapview/
UCSC Table Browser	Query for genomic features	http://genome.ucsc.edu/cgi-bin/hgTables
BioMart	Query for features of genes	http://www.ensembl.org/biomart/martview/
VISTA Enhancer Browser	Data on conserved noncoding regions tested as developmental enhancers	http://enhancer.lbl.gov/
Galaxy	Interactive workspace for analysis of genome sequences, alignments and annotation	http://main.g2.bx.psu.edu/

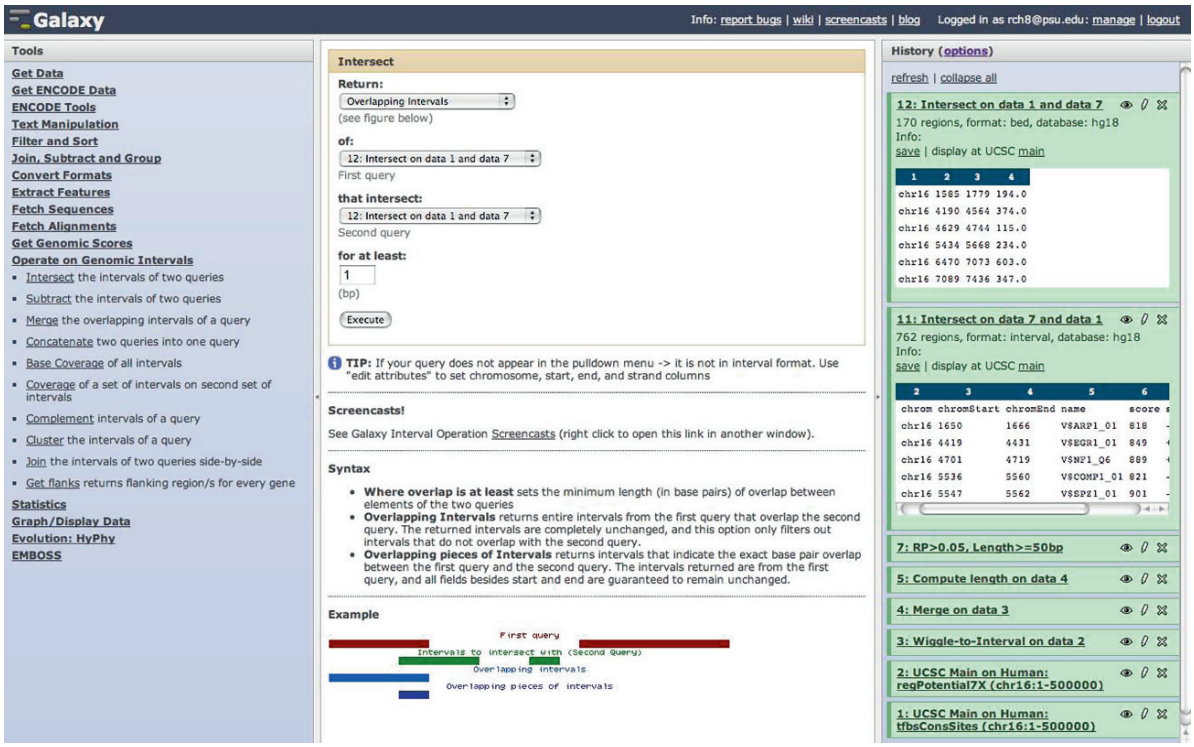

Fig. 19.17 Using Galaxy to find predicted regulatory regions. The user interface for Galaxy has three panels. Tools for obtaining and analyzing data are selected from the left panel, and the user selects input data and other parameters in the central panel. A history of previous results is maintained on the right panel. In this example, candidates for gene regulatory modules in a 500 kb region of human Chromosome 16 are obtained by queries to the UCSC Table Browser to obtain conserved matches to transcription factor binding motifs (query 1) and regions of of high regulatory potential (score >=0.05 in query 2; these results were converted to intervals, merged and filtered for length >=50 bp to obtain the results in query 7). Intersections reveal conserved motifs that are in regions of high regulatory potential (query 11) and *vice versa* (query 12)

Precomputed scores such as phastCons and regulatory potential can be aggregated on specified intervals. The interface at Galaxy for a series of operations that can predict gene regulatory regions is shown in Fig. 19.17.

19.7 Concluding Remarks

Comparative genomics brings considerable power but daunting challenges to the study of human genetics. No aspect of comparative genomics has been perfected; even the commonly used methods of aligning sequences and predicting protein-coding genes have room for improvement. However, considerable insight and functionality can be gleaned from the predictions and comparisons that are currently available. Real biological variation, for example, in the rate of evolutionary change at different loci or the phylogenetic depth of conservation of a feature class, means that no single threshold for a conservation-based score will be adequate to find all the features of interest. However, as the variation is better understood and as functional correlates of the variation are established, then the potential power of comparative genomics will be better harnessed. Current data can be readily accessed and evaluated. Additional types of data, such as genome-wide ChIP-chip results, coupled with tools for better integration of disparate data types, should lead to considerable future progress in the functional annotation of the human genome.

References

1. 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
2. Antonarakis SE, Boehm CD, Serjeant GR, Theisen CE, Dover GJ, Kazazian HH Jr (1984) Origin of the beta S-globin gene in blacks: the contribution of recurrent mutation or gene conversion or both. Proc Natl Acad Sci USA 81:853–856
3. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441:87–90
4. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. Science 304:1321–1325
5. Biswas S, Akey JM (2006) Genomic insights into positive selection. Trends Genet 22:437–446
6. Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D et al (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res 16:656–668
7. Blankenberg D, Taylor J, Schenck I, He J, Zhang Y, Ghent M, Veeraraghavan N, Albert I, Miller W, Makova K et al (2007) A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. Genome Res in press
8. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391–1394
9. Bray N, Pachter L (2003) MAVID multiple alignment server. Nucleic Acids Res 31:3525–3526
10. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Comparative Sequencing Program NISC, Green ED, Sidow A, Batzoglou S (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13:721–731
11. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94
12. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD et al (2005) Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157
13. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA (2005) Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res 15:1553–1565
14. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87
15. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B et al (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963
16. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15:901–913
17. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M (2004) The Ensembl automatic gene annotation system. Genome Res 14:942–950
18. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, DC, pp 345–352
19. Dermitzakis E, Clark A (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. Mol Biol Evol 19:1114–1121
20. Dermitzakis ET, Reymond A, Antonarakis SE (2005) Conserved non-genic sequences -an unexpected feature of mammalian genomes. Nat Rev Genet 6:151–157
21. Elnitski L, Jin VX, Farnham PJ, Jones SJ (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res 16:1455–1464
22. Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Paabo S (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869–872
23. Evans PD, Gilbert SL, Mekel-Bobrov N, Vallender EJ, Anderson JR, Vaez-Azizi LM, Tishkoff SA, Hudson RR, Lahn BT (2005) Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. Science 309:1717–1720
24. Fishel R, Lescoe MK, Rao MR, Copeland NG, Jenkins NA, Garber J, Kane M, Kolodner R (1993) The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. Cell 75:1027–1038
25. Weinstock GRA, Metzker ML Muzny DM Sodergren EJ Scherer GM, Scott S, Steffen G, Worley KC Burch PE D et al (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493–521
26. Hardison RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet 16:369–372
27. Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D et al (2003) Covariation in frequencies of substitution, deletion, transposition and recombination during eutherian evolution. Genome Res 13:13–26
28. Henderson J, Salzberg S, Fasman KH (1997) Finding genes in DNA with a Hidden Markov Model. J Comput Biol 4:127–141
29. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919
30. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MA, Delany ME et al (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716
31. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T et al (2007) Ensembl 2007. Nucleic Acids Res 35:D610–D617
32. International Hapmap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
33. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

19

34. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G et al (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res 14:331–342
35. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ (2004) The UCSC table browser data retrieval tool. Nucleic Acids Res 32:D493–D496
36. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E (2003) EnsMart – a generic system for fast and flexible access to biological data. Genome Res 14:160–169
37. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res 16:980–989
38. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12:996–1006
39. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci USA 100:11484–11489
40. Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626
41. Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267:275–276
42. King DC, Taylor J, Cheng Y, Martin J, ENCODE Transcriptional Regulation Group, ENCODE Multispecies Alignment Group, Chiaromonte F, Miller W, Hardison RC (2007) Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. Genome Res 17:799–816
43. Kondrashov AS (2002) Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. Hum Mutat 21:12–27
44. Kreitman M (2000) Methods to detect selection in populations with applications to the human. Annu Rev Genomics Hum Genet 1:539–559
45. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A et al (2007) The UCSC genome browser database: update 2007. Nucleic Acids Res 35:D668–D673
46. Kumar S, Hedges SB (1998) A molecular timescale for vertebrate evolution. Nature 392:917–920
47. Leach FS, Nicolaides NC, Papadopoulos N, Liu B, Jen J, Parsons R, Peltomaki P, Sistonen P, Aaltonen LA, Nystrom-Lahti M et al (1993) Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. Cell 75:1215–1225
48. Li WH, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239
49. Wade Lindblad-Toh K, Mikkelsen CM, Karlsson TS, Jaffe EK, Kamal DB, Clamp M, Chang M, Kulbokas EJ 3 rd, Zody MC et al (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819
50. Ludwig MZ, Bergman C, Patel NH, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403:564–567
51. Maclean JA 2nd, Chen MA, Wayne CM, Bruce SR, Rao M, Meistrich ML, Macleod C, Wilkinson MF (2005) Rhox: a new homeobox gene cluster. Cell 120:369–382
52. Margulies EH, Blanchette M, Comparative sequencing program NISC, Haussler D, Green ED (2003) Identification and characterization of multi-species conserved sequences. Genome Res 13:2507–2518
53. Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, Birney E, Keefe D, Schwartz AS, Hou M et al (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res 17:760–774
54. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16:1046–1047
55. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 354:114–116
56. Mekel-Bobrov N, Gilbert SL, Evans PD, Vallender EJ, Anderson JR, Hudson RR, Tishkoff SA, Lahn BT (2005) Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens. Science 309:1720–1722
57. Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156:297–304
58. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453
59. Neel JV (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 14:353–362
60. Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York
61. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ et al (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3:e170
62. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. Science 302:413
63. Pagnier J, Mears JG, Dunda-Belkhodja O, Schaefer-Rego KE, Beldjord C, Nagel RL, Labie D (1984) Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. Proc Natl Acad Sci USA 81:1771–1773
64. Pennacchio LA, Rubin EM (2001) Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2:100–109
65. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD et al (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444:499–502
66. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A et al (2006) An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443:167–172
67. Ponjavic J, Ponting CP, Lunter G (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res 17:556–565
68. Pruitt KD, Maglott DR (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29:137–140
69. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E et al (2000) Genome-wide location and function of DNA binding proteins. Science 290:2306–2309
70. Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234

71. Schwartz S, Elnitski L, Li M, Weirauch M, Riemer C, Smit A, NISC Comparative Sequencing Program, Green ED, Hardison RC, Miller W (2003) MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. Nucleic Acids Res 31:3518–3524
72. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W (2003) Human-mouse alignments with Blastz. Genome Res 13:103–105
73. Siepel A, Haussler D (2004) Computational identification of evolutionarily conserved exons. In: Proceeding of the 8th annual international conference on research in computational biology (RECOMB '04), pp. 177–186
74. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050
75. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197
76. Spurr NK, Gough A, Goodfellow PJ, Goodfellow PN, Lee MG, Nurse P (1988) Evolutionary conservation of the human homologue of the yeast cell cycle control gene cdc2 and assignment of Cd2 to chromosome 10. Hum Genet 78:333–337
77. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595
78. Taylor J, Tyekucheva S, King DC, Hardison RC, Miller W, Chiaromonte F (2006) ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements. Genome Res 16:1596–1604
79. The ENCODE Project Consortium (2007) The ENCODE pilot project: identification and analysis of functional elements in 1% of the human genome. Nature 447:799–816
80. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680
81. Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G (2007) Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biol 8:R15
82. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:e72
83. Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci USA 103:135–140
84. Wang H, Zhang Y, Cheng Y, Zhou Y, King DC, Taylor J, Chiaromonte F, Kasturi J, Petrykowska H, Gibb B et al (2006) Experimental validation of predicted mammalian erythroid cis-regulatory modules. Genome Res 16:1480–1492
85. Waterston RH, LindbladToh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562
86. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S et al (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35:D5–D12
87. Wiehe T, Gebauer-Jung S, Mitchell-Olds T, Guigo R (2001) SGP-1: prediction and validation of homologous genes based on sequence alignments. Genome Res 11:1574–1583
88. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K et al (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3:e7
89. Wu JQ, Shteynberg D, Arumugam M, Gibbs RA, Brent MR (2004) Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing. Genome Res 14:665–671