Molecular evolution meets the genomics revolution

Kenneth H. Wolfe[1] & Wen-Hsiung Li[2]

doi:10.1038/ng1088

Changes in technology in the past decade have had such an impact on the way that molecular evolution research is done that it is difficult now to imagine working in a world without genomics or the Internet. In 1992, GenBank was less than a hundredth of its current size and was updated every three months on a huge spool of tape. Homology searches took 30 minutes and rarely found a hit. Now it is difficult to find sequences with only a few homologs to use as examples for teaching bioinformatics. For molecular evolution researchers, the genomics revolution has showered us with raw data and the information revolution has given us the wherewithal to analyze it. In broad terms, the most significant outcome from these changes has been our newfound ability to examine the evolution of genomes as a whole, enabling us to infer genome-wide evolutionary patterns and to identify subsets of genes whose evolution has been in some way atypical.

Molecular evolution research has always been opportunistic. Many scientists working in the field, ourselves included, do little or no work at the bench and instead rely on the public DNA sequence databases to provide the grist for our research mill. This practice dates back to the earliest evolutionary analyses on the first mRNA sequences[1–3]. Consequently, many discoveries in molecular evolution have been facilitated by advances in genomics technology. Frequently, data that were not originally collected for evolutionary purposes have subsequently yielded important evolutionary insights (Fig. 1). The flip side of this opportunism is that there have been few glimpses of a 'big picture' in molecular evolution research, despite the growing data sets. Fundamental questions, such as the relative roles of neutral evolution versus darwinian selection, have not been addressed systematically but rather in a piecemeal manner, as permitted by the available data.

In this review we summarize some areas of molecular evolution research in which genomics has had a strong impact in the past decade. We consider five disparate areas of particular interest: the origins of new genes, the prevalence of positive natural selection, the asymmetry of mutation patterns, regional variation in mutation rates, and the evolution of genome organization. We have tried to include examples from a broad range of organisms. If there is an overall theme to our review, it is that genomics, bioinformatics and molecular evolution are becoming more and more intertwined: evolutionary considerations are becoming central to the interpretation of genomics data, progress in molecular evolution research depends on genomics data, and nobody can handle the data without bioinformatics.

Where do new genes come from?

Because the number of genes in an organism's genome is linked (loosely) to its biological complexity, the process by which new genes are formed has fascinated geneticists for a long time[4]. Three mechanisms of gene formation are imaginable: duplication of pre-existing genes, creation of mosaic genes from parts of other genes, and *de novo* invention of genes from DNA that was previously non-coding. Examples of all three are known, as discussed below.

Gene duplication

Complete gene duplication is the most familiar of the gene formation mechanisms and probably accounts for most new genes. The relative conservation of intron/exon structure within gene families in most eukaryotes suggests that successful gene duplications occur more readily through DNA-mediated events than through the reverse transcription of mRNA intermediates, although the latter process does occur[5,6]. Lynch and Conery[7] used genome sequences from several eukaryotes to estimate the rate at which gene duplication occurs. They found the rate to be relatively uniform across species and of the order of 0.01 duplications per gene per million years. Their study emphasized the short half-life of duplicate genes, which was estimated to be only 3–8 million years. Eukaryotic genomes can be therefore viewed as proving grounds in which duplicate genes are continually generated, tested and often discarded.

Duplicated sequences either degenerate into pseudogenes or turn into new genes, and there has been much discussion about what factors govern the fate of a newly duplicated sequence. If a new gene is an exact copy of another gene, the only way that it

[1]*Department of Genetics, Smurfit Institute, University of Dublin, Trinity College, Dublin 2, Ireland.* [2]*Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA. Correspondence should be addressed to K.H.W. (e-mail: khwolfe@tcd.ie).*

can confer an immediate selective advantage is through selection that favors increased amounts of its protein or mRNA product, such as may occur for ribosomal protein genes. As two duplicate genes diverge, subfunctionalization can occur in which the two genes accumulate different degenerative mutations such that each ends up with a subset of the original gene's functions, making both of them essential[8]. Occasionally, a duplicate gene may gain mutations that confer a new function and thus a selective advantage for its persistence in the genome.

Perhaps the most dramatic way of increasing the number of genes in an organism is to double the whole genetic content through polyploidization[9]. Of the eukaryotes whose genomes have been sequenced, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* show evidence of having gone through relatively recent polyploid stages. The presence of many large series of duplicated genes on different human chromosomes[10,11], and the one-to-many relationship between some regions of the human genome and the *Amphioxus* genome[12], indicate that at a minimum the genome of an ancestor of vertebrates underwent duplications of large tracts of chromosomes. The subsequent evolution of newly formed polyploid species is poorly understood, but studies of polyploid plants created in the laboratory have shown that their genomes can undergo marked and very rapid rearrangements, resulting in an almost immediate loss of many gene copies and the silencing of other loci by methylation[13–15].

Mosaic genes

A more innovative way to create a gene is by the 'Lego approach'. There are many recent examples of genes that have been assembled from duplicated parts of other genes. Genome projects have been particularly useful for identifying the sources of the various pieces of DNA involved. Among the most spectacular examples of gene assembly are genes that transferred from the mitochondrial genome to the nuclear genome during recent plant evolution[16,17]. For these transfers to be successful, the protein encoded by the gene must be imported back into the mitochondrion, usually by means of an amino-terminal transit peptide. Often, the newly transferred gene has acquired DNA encoding a transit peptide from another gene, either by duplication of the relevant exons[18] or by alternative splicing with exon sharing[19]. Similarly, chimeric genes formed during recent evolution have been identified in the human[20,21] and *Drosophila melanogaster*[22,23] genomes.

In mammals, the transduction of L1 elements that flank gene-coding DNA has the potential to create chimeric genes by exon shuffling[6,24], although no examples of genes formed in this way have been found[25]. A gene can also turn into two by fission, as illustrated by a gene encoding nitric oxide synthase in a snail[26]; a recent DNA inversion inside this gene broke it into two separate smaller genes encoding parts of the original protein.

De novo gene formation

The formation of genes from noncoding DNA seems to be a rare phenomenon, but a few examples, such as the *morpheus* gene family in primates[27], have been reported. *Morpheus* is a very rapidly evolving transcript derived from a repeat sequence that is present in multiple copies on human chromosome 16. A repetitive sequence element was also involved in the genesis of another human gene, *LQK1* (ref. 28). The antifreeze glycoprotein (*AFGP*) gene of the Antarctic fish *Dissostichus mawsoni*[29] was formed by the duplication of a pancreatic trypsinogen gene, followed by the deletion of all exons except the first and the last, with replacement of the central portion of the gene by a highly repetitive sequence encoding $(Thr\text{-}Ala\text{-}Ala)_n$ oligomers. Notably, convergent evolution at the molecular level during the cooling of the Arctic and Antarctic Oceans during past few million years has resulted in almost identical sequences for the antifreeze peptides in the fish in these oceans[30].

Lateral gene transfer

Another source of genes is lateral gene transfer between species. This is very evident among bacteria for which genome sequences from several, closely related species or strains are available, such as the *Escherichia coli* and *Salmonella typhi* group[31–33]. The *E. coli* strains K12 and O157:H7 share in common a 'backbone' genome totaling 4.1 Mb of DNA, but substantial strain-specific 'islands' of DNA contribute a further 0.5 and 1.3 Mb, respectively, to the two strains[34].

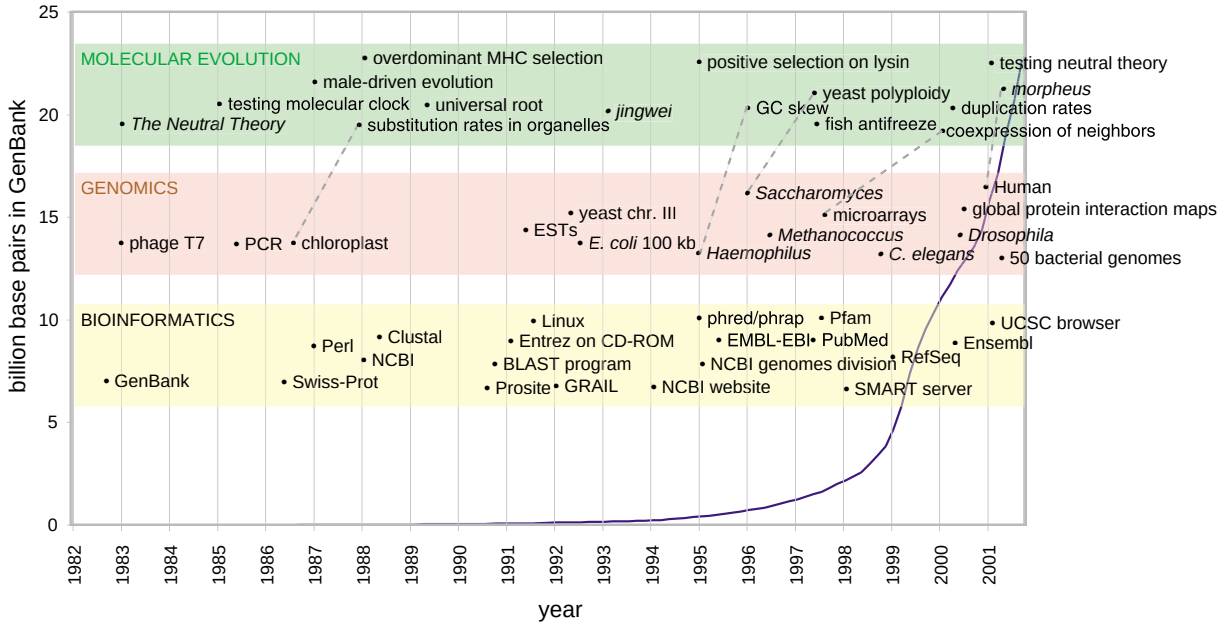

Fig. 1 Timeline of developments in bioinformatics, genomics and molecular evolution, charted against the accumulation of DNA sequence information in GenBank, which was established in 1982. Links between genomics data and subsequent molecular evolution advances are indicated by broken lines.

Whether lateral gene transfer is as prevalent in eukaryotes as it is in bacteria remains to be seen. For example, it is unclear at present whether the 'orphan' genes (those without homologs in other species) found in the genomes of some yeast species are derived from *de novo* gene formation from lateral transfer from unidentified donor species, or are simply the result of evolving very fast[35].

Positive selection and the neutral theory

Much effort has been directed at detecting the presence of positive selection during the evolution of a gene, owing to the abundance of DNA sequence data and the development of detection methodology[36–40]. In addition, the increasing amount of DNA sequence and polymorphism data has stimulated re-examination of the neutral theory of molecular evolution.

In the search for examples of positive selection, much attention has been paid to genes involved in defense against pathogens (Table 1). One of the first discoveries was that the antigenic regions of major histocompatibility complex (MHC) proteins and immunoglobulins are under overdominant selection[41–43]. Diversity-enhancing selection has been proposed for colicins in *E. coli*[44]; colicins are toxin proteins produced by and active against *E. coli* and related bacteria. Evidence has been found for directional positive selection during the early evolution of eosinophil cationic protein (ECP). This protein was derived by duplication of the ribonuclease gene encoding eosinophil-derived neurotoxin (EDN) in the common ancestor of Old World primates, but it acquired a different function by becoming a potent toxin to pathogenic bacteria and parasites[45]. Positive selection has also occurred in EDN: substitutions at two interacting sites in this toxin increased its ribonucleolytic activity by 13-fold and, together with other substitutions, also increased its antiviral potency[46]. Evidence for positive selection has also been provided for other antipathogen proteins such as glycophorin A, RH50 and interleukin-2 (Table 1).

In pathogens, the evolution of proteins involved in evading the defensive systems of hosts has often been driven by positive selection. For example, the circumsporozoite protein is a cell-surface protein of the sporozoite of malaria parasites (*Plasmodium* spp.) and evidence of positive selection has been found for its immunogenic regions[47]. Other well-known examples are the merozoite surface antigen-1 gene of *Plasmodium falciparum*[48] and the envelope gene of human immunodeficiency viruses[49,50]. Many other examples are listed in Table 1.

Table 1 • Genes or proteins in which positive darwinian selection has been detected

Gene or protein	Organisms	References
Defensive systems or immunity		
MHC genes	primates, rodents	41,43
immunoglobulin V_H genes	primates, rodents	42
colicin genes	*E. coli*	44
type I interferon genes	mammals	163
neomycin resistance protein	*E. coli*	164
neurotoxin	snake	164
α_1-proteinase inhibitor genes	rodents	165
defensin genes	rodents	166
Rh blood group and RH50 genes	primates, rodents	167,168
Fv1	*Mus*	169
ECP	Old World primates	45
transferrin gene	salmonid fishes	170
ribonucleases	primates, rodents	46,171
class I chitinase gene	*Arabis*, *A. thaliana*	172
glycophorin A	human, primates	168,173
interleukin-2	mammals	174
Evading defensive systems or immunity		
circumsporozoite protein	*P. falciparum*	47
merozoite surface antigen-1	*P. falciparum*	48
CSP, TRAP, MSA-2 and PF83	*P. falciparum*	164,175
porin protein 1 gene	*Neisseria*	176
E gene	phages G4, ϕX174, S13	164
envelope gene	equine infectious anemia virus	164
glycoprotein *gH* gene	pseudorabies virus	164
invasion plasmid antigen genes	*Shigella*	164
msp 1α	Rickettsia anaplasma marginale	164
outer membrane protein	*Chlamydia*	164
σ1 protein gene	Reovirus	164
virulence determinant gene	*Yersinia*	164
S and *HE* glycoprotein genes	murine coronavirus	177
hemagglutinin gene	human influenza A virus	178
δ-antigen coding region	hepatitis D virus	179
nef gene	HIV	180
envelope gene	HIV	49,50
capsid genes	foot and mouth disease virus	181
Male reproduction		
Acp26Aa	*D. melanogaster*	54–56,182
lysin	teguline gastropods	51,183,184
bindin	sea urchins	52,53
Sry gene	primates	185
18-kDa fertilization protein	Abalone (*Haliotis*)	186
S-RNase gene	Rosaceae	187
androgen-binding protein	rodents	188
protamine 1	human, chimpanzee	168,189
protamine 2	human, chimpanzee	168,189
TMAP	teguline gastropods	190
acrosin-trypsin inhibitor	human	168
PSP94	human	168

Table 1 • (continued)

Gene or protein	Organisms	References
Female reproduction		
egg-laying hormone genes	*Aplysia californica*	164
zona pellucida ZP2	mammals	184
zona pellucida ZP3	mammals	184
oviductal glycoprotein	mammals	184
chorionic gonadotropin	primates	58
Miscellaneous		
Adh	*D. melanogaster*	36
G6PD	*D. melanogaster*	191
jingwei	*D. melanogaster*	22
phospholipase A2 gene	Crotalinae snakes	192
ATP synthase F_o subunit gene	*E. coli*	164
CDC6	*S. cerevisiae*	164
prostatein peptide C3 gene	rat	164
interleukin-3 gene	primates	193
interleukin-4 gene	rodents	193
Growth hormone gene	primates, Artiodactyla	194,195
lysozyme	primates	37,59
Pem homeodomain	mice, rats	196
κ-casein gene	bovids	197
COX4 gene	primates	198
hemoglobin β-chain gene	Antarctic fishes	199
Ods homeobox gene	*D. melanogaster*	200
conotoxins	predatory snails	201
COX7A isoform genes	primates	202
BRCA1	human, chimpanzee	203
Mth	*D. melanogaster*	204
morpheus genes	human, great apes	27
dopamine receptor D4	human	205

Much effort has been focused on genes that are involved directly in reproduction. In free-spawning marine invertebrates, the evolution of species-specific fertilization is important for reproductive isolation, and the biochemistry and evolution of many proteins that mediate fertilization have been studied extensively. In the abalone, the sperm protein lysin creates a hole in the egg vitelline envelope by binding to its egg receptor, and the evolution of the species specificity of lysin is promoted by positive selection[51]. The sea urchin gamete-recognition protein bindin has evolved similarly through positive selection[52,53]. In other organisms, male-specific proteins, such as the male ejaculatory protein Acp26Aa in *Drosophila*[54–56], are often targets of positive selection (Table 1). A broader study of expressed sequence tags (ESTs) from 176 male reproductive protein genes in *Drosophila* has shown that about 11% of ESTs are subject to positive selection[57].

Although positive selection is a recurrent theme in male reproductive proteins, only a few female reproductive proteins, such as chorionic gonadotropin, have been found to be driven by positive selection (Table 1). Chorionic gonadotropin is an essential signal in establishing pregnancy in higher primates but has not been found in other mammals, indicating that it is a new reproductive protein in higher primates. The β-subunit of this female reproductive hormone arose by duplication from the luteinizing hormone β-subunit in the common ancestor of higher primates, and its carboxy-terminal portion has undergone several periods of positive selection in New World monkeys and hominoids[58].

Positive selection has also been found in genes that confer an advantage for the organism to adapt to a different environment or physiological requirement. Lysozyme has apparently undergone adaptive evolution in langur monkeys[37,59], which are unique among primates because they have a foregut in which bacteria ferment leaves, followed by a true stomach that expresses high quantities of lysozyme to digest bacteria. Similarly, adaptive evolution of a duplicated pancreatic ribonuclease gene has occurred in a langur monkey to help digest bacteria[46].

Each of the above-mentioned studies examined whether a protein has experienced positive selection in the course of its evolution. A more general issue that has been controversial since the proposal of the neutral mutation hypothesis in 1968 is the proportion of amino acid substitutions in protein evolution that is driven by positive selection[60]. This proportion has been estimated recently from DNA polymorphism and divergence data to be about 35–45% in *Drosophila* and human[61–63]. These estimates are considerably higher than those proposed by the neutral theory of molecular evolution[64]. Not surprisingly, the proportion is higher for genes that have evolved fast and lower for those that have evolved slowly[63]. Because these estimates were based on limited data, however, this issue should be re-examined when more data become available.

Strand asymmetry in DNA mutation

The two strands of DNA differ with respect to replication and transcription. During replication, the leading strand is synthesized continuously, whereas the lagging strand is synthesized discontinuously, and transcription overexposes the nontranscribed strand to DNA damage. Both processes are therefore asymmetric and might bias the occurrence of mutations between the two strands. Indeed, this possibility has been supported by experimental studies[65,66] and by statistical analyses of genomic sequence data (reviewed in refs. 67,68). The latter studies have been especially useful for understanding the prevalence and causes of strand asymmetry in DNA mutation.

Two commonly used measures for strand asymmetry are the GC skew, $(G-C)/(G+C)$, and the TA skew, $(T-A)/(T+A)$, where G, C, T and A denote the frequencies of the four nucleotides in the strand under study[69]. These two skews detect deviations from G = C and T = A, which are the expected frequencies on each strand when there is no bias in mutation and selection between the two strands. An early analysis of the genomes of *E. coli*, *Bacillus subtilis* and *Haemophilus influenzae* showed that the GC skew is stronger than the TA skew, but both skews switch sign at the origin of replication and are stronger in intergenic regions and in third codon positions, which suggests that mutational bias is largely responsible for the asymmetry[69]. In general these observations hold for eubacteria (Fig. 2; refs. 67,68,70).

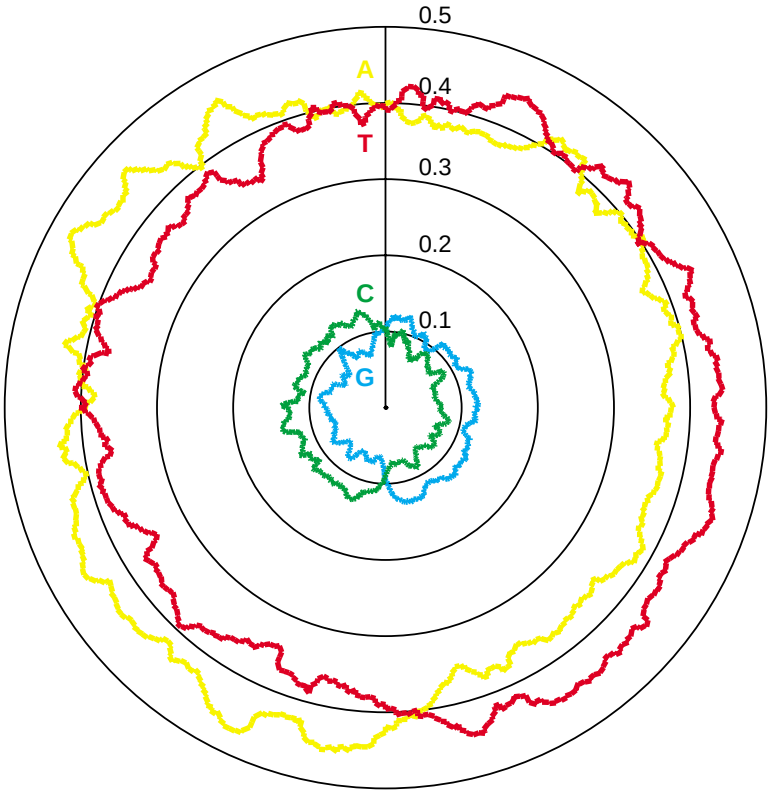

Fig. 2 Variation in base composition around the genome of *Campylobacter jejuni*. The radar plot shows the frequency of the four nucleotides at synonymous (fourfold degenerate) codon positions, calculated as a moving average from synonymous sites within a window of 40 kb of genomic sequence. The origin of replication is at the top. The leading strand is relatively rich in T and G. Sequence data are from ref. 214.

Various theories have been proposed to explain strand bias on the basis of the asymmetry of the replication bubble. For example, different replication error rates between the two strands, different processivities of the leading and lagging strands, and different repair efficiencies between the two strands have been proposed, but none has found much support. By contrast, the cytosine deamination theory[68] has received much attention. Because the leading strand is in a single-stranded state to act as a template for synthesizing the lagging strand, it is exposed for longer periods to DNA damage, especially cytosine deamination, which increases C to T mutations. This largely explains the strong GC skew, although there may be other factors involved in strand asymmetries[71].

The deamination theory can also explain the strong compositional asymmetry in mitochondrial genomes, in which the skew is clearly high at synonymous codon positions[72–75]. The replication of mitochondrial DNA is highly asymmetrical: the daughter H strand displaces the parental strand so that the parental H strand remains single-stranded and exposed to damage until paired with the newly synthesized L strand.

Deamination also seems to form the basis of strand asymmetries in transcription-induced mutations in eubacteria[76]. During transcription, cytosine deamination is less frequent on the template strand than on the nontranscribed strand, because the former is shielded by the RNA polymerase and the nascent mRNA[77]. In combination with a much higher number of genes on the leading strand (see below), transcription-induced mutations can contribute to large-scale compositional asymmetries between the leading and lagging strands in bacterial genomes (Fig. 2).

As yet, however, there is no evidence of asymmetric directional mutation pressure in eukaryotes[78,79], with the exception of subtelomeric sequences in yeast[80]; this is probably due to the presence of multiple replication origins in eukaryotes, many of which may often change locations. In Archaea, little evidence of strand asymmetry was found in early studies[81,82], but GC skews and a single origin of replication have been identified recently in three *Pyrococcus* species[83].

The presence of asymmetric mutational pressure has many evolutionary implications. First, it may complicate the estimation of evolutionary distances because traditional methods assume strand symmetry. Second, it may be an important source of variation in codon usage and amino acid usage[84,85]. Third, it may have been responsible for the higher number of genes located on the leading strand in many bacterial genomes[82,84]. Last, genes on the two strands may evolve at different rates, and those that have switched their orientation relative to the direction of replication may show accelerated rates of nucleotide and amino acid substitution[71,86,87].

Effects of genomic location on mutation rates

Many studies have focused on the extent of variation in the mutation rate among regions of the mammalian genome and the possible causes of this variation. The possibility of a higher mutation rate in males than in females was first proposed by Haldane[88]. Such a difference should lead to a higher mutation rate in Y-linked sequences than in X-linked and autosomal sequences, and Miyata *et al.*[89] developed a method for estimating the male-to-female ratio (α) of mutation rates from the substitution rates in homologous Y-linked and X-linked (or autosomal) sequences. Applications of this method to noncoding sequences gave estimates of $\alpha = 5–6$ in Old World primates, $\alpha \approx 4$ in cats, and $\alpha \approx 2$ in murid rodents (Table 2), indicating that α increases with increasing generation time.

In addition, it has been estimated that the values in murid rodents and Old World primates are similar to the male-to-female ratios of the numbers of germ cell divisions in these organisms[90]. These observations have been taken both as evidence for the view that mutations occur mainly during DNA replication in the germ line and as support for the generation-time effect hypothesis[90], which postulates that the molecular clock runs faster in short-living animals than in long-living ones.

This issue is by no means resolved. When the rate of silent-site evolution of X-linked genes was compared with that of autosomal genes, α was estimated to be infinity—in other words, beyond the maximum value expected from sex differences. It was therefore proposed that the high α values estimated from comparisons of X-linked and Y-linked sequences were due to a reduced mutation rate in the X chromosome rather than to an increased mutation rate in the Y chromosome; that is, there is very weak or no male-driven evolution[91]. But this view is not supported by the finding of a higher rate of male mutation in birds, although male birds are homogametic, which is opposite to what is found in mammals[92]. In addition, a recent study comparing the substitution rates in homologous autosomal and Y-linked sequences has supported strong male-driven evolution in higher primates (Table 2)[93].

Table 2 • Ratio of substitution rates on different chromosomes and male-to-female ratio of mutation rate in different organisms

Taxa	Gene pair	Rate ratio (m)[a]	α (95% CI)[b]	References
primates	*AMELY/AMELX*	Y/X = 2.16	5.14 (2.42–16.6)	206
primates	*ZFY/ZFX*	Y/X = 2.27	6.26 (2.63–32.4)	207
primates	*SMCY/SMCX*	Y/X = 2.03	4.20 (2.20–10.0)	208
primates	noncoding	Y/A = 1.68	5.25 (2.44–∞)	93
cats	*ZFY/ZFX*	Y/X = 2.06	4.38 (3.76–5.14)	209
rodents	*ZFY/ZFX*	Y/X = 1.42	1.80 (1.0–3.2)	210
rodents	*Ube1Y/Ube1X*	Y/X = 1.50	2.0 (1.0–3.9)	211
birds	*CHD1Z/CHD1W*	Z/W = 4.65	6.5 (2.8–10.2)	92
birds	*CHD1Z/CHD1W*	Z/W = 3.06	4.1 (3.1–5.1)	212
birds	*ATP5A1Z/ATP5A1W*	Z/W = 0.66, 0.52, 0.274	1.8; 2.3; 5.0	213

[a]Ratio of substitution rates on different chromosomes. [b]Estimated male-to-female ratio of mutation rate; 95% confidence intervals are given in parentheses.

It has been proposed that variation in mutation rate also occurs among autosomal regions[94,95]. More explicitly, the regional mutation pressure hypothesis postulates that the rate and pattern of mutation varies among genomic regions[95]. This hypothesis has been supported by the observations that silent sites in adjacent genes evolve at more similar rates than do non-adjacent genes[96], and that the G+C content of a repetitive element tends to become similar to the G+C content of the region into which it was inserted[97]. The finding of local similarity in mutation rate has been contested by Kumar and Subramanian[98], who claim that when genes whose G+C content is not at equilibrium are excluded from the comparison, local similarity in mutation rate is no longer observed. It is not clear, however, whether this can explain the observation of significant variations in rate among autosomes (for example, see refs. 99,100). Additional support for a regional variation in mutation rate comes from the observation that the synonymous rate in a mammalian gene is correlated positively with the G+C content at the third codon positions of the gene[101,102]. This correlation should lead to uneven mutation rates among genomic regions because the G+C content varies among regions of eukaryotic genomes[103].

Recombination is another factor that might cause regional variation in mutation rate because it has been proposed to be mutagenic and its rate varies along the genome. In yeast, recombination involves double-strand breaks (DSBs), the repair of which is error-prone such that recombination increases the chance of mutation[104]. In mammals, recombination, although not known to involve DSBs, also seems to be mutagenic, as implied by the 170-fold increase in silent substitution achieved when the last three exons of *Fxy* became part of the pseudoautosomal region (PAR) in the *Mus musculus domesticus* lineage[105]; PAR has a much higher recombination rate as compared with regions unique to the X chromosome.

In addition, a strong correlation between recombination rate and G+C content has been observed in many organisms, including yeast[106,107], *D. melanogaster*[108] and mammals[109,110]. In mammals, the direct observation of mismatch corrections in simian cells identified a GC-biased mismatch correction mechanism during the recombination process[109]. Thus, recombination might underlie a positive correlation between G+C content and mutation rate and might be an important factor for the variation in mutation rate and pattern among regions.

Evolution of genome structure and organization

Complete genome sequences provide us with information about the position of every gene on a chromosome, and comparative genomics allows us to study how gene locations evolve. In bacteria, genes with related functions are often located close together on the chromosome because they are co-transcribed as operons. In the nematode *Caenorhabditis elegans*, about 15% of the genes are co-transcribed with their neighbors, but only a few of the operons seem to contain genes that are obviously functionally related[111,112].

Although most other eukaryotes lack operons, we are familiar with the idea that some parts of the genome contain gene clusters with functional themes, such as the MHC and the *Hox* gene clusters. A spectacular example is the discovery by Wang *et al.*[113] that half of the genes expressed specifically in human spermatogonia are encoded on the X or Y chromosomes. Chromosomal clustering of functionally related genes has been found recently in both *C. elegans*[114] and *D. melanogaster*[115].

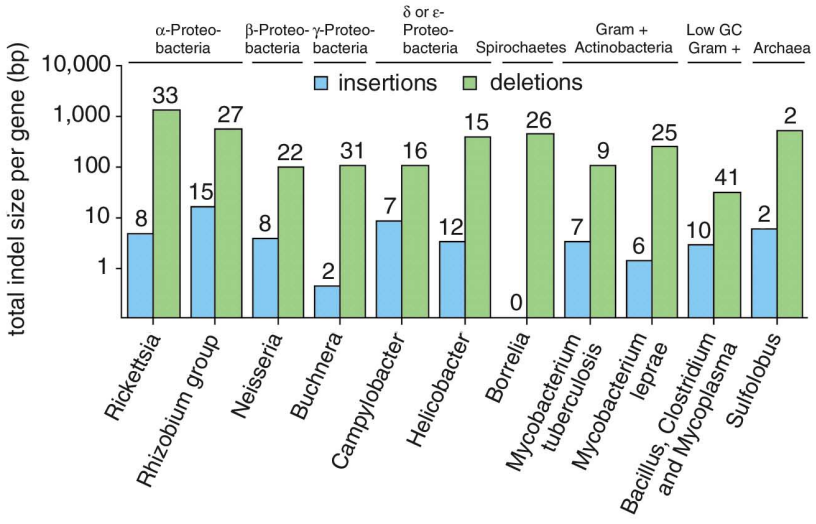

Fig. 3 Frequency of deletions and insertions in bacterial genomes. Frequencies are based on the comparative analyses of pseudogenes and their functional counterparts in a closely related species, generally from the same genus, with at least one functional gene in a bacterial outgroup. Columns indicate the average total size of deletions and insertions per pseudogene (in bp). Numbers at the tops of columns indicate the numbers of each type of event. Figure used, with permission, from ref. 142.

Pioneering studies have also shown that, across the genome, adjacent genes are co-regulated more often than is expected by chance. This has been shown for the yeast genome using transcription data from microarrays[116–118], and for the human genome using tissue distribution of mRNAs[119]. These preliminary results suggest that the 'beads on a string' model of how genes are ordered on chromosomes is inadequate, and that there may be some adaptive significance to where genes are located.

Comparison of genome sequences between closely related species, such as human and mouse, often shows extensive conservation of gene order[120,121]. At increasing evolutionary distance, this conservation breaks down by processes including local rearrangements, such as inversions of single genes, and breakpoints corresponding to interchromosomal rearrangements[122–124]. If there are significant clusters of functionally related genes in most eukaryotic genomes, they should become apparent as units of conserved linkage that are resistant to evolutionary rearrangement; however, this has not as yet been tested.

Comparative genomics can have practical applications—for example, in groups of species where there are great differences in genome size. The maize genome is roughly 12 times larger than the rice genome, but the two are very similar in terms of gene order. The difference in size is due to vastly increased numbers of transposable elements in the maize genome, which inflate intergenic distances and, to a lesser extent, intron sizes. The maize genome is still expanding and is estimated to have doubled in size in the past 3 million years[125]. It is not known what factors, if any, govern genome size. Petrov and colleagues[126,127] have shown that the rate at which DNA deletions accumulate varies widely among different species of insect, and that the species with lower deletion rates have larger genomes.

Genomes can shrink as well as expand. Extreme DNA deletion pressures may explain how several genomes that are intracellular residents have become so compact. The most familiar of these are the mitochondrial genomes of animals, which have almost no intergenic DNA, although other examples have been found in the past few years. The nucleomorph genomes of cryptomonad[128] and chlorachniophyte[129,130] algae are descendants of algal nuclear genomes that became residents inside other eukaryotic cells in two independent endosymbiosis events. The microsporidian *Encephalitozoon cuniculi*[131,132] is an obligate intracellular parasite of human cells. Highly convergent genomic evolution is seen in these three genomes. All three have very short intergenic spacers, tiny introns and shortened proteins and have also lost many genes that were present in their free-living relatives. In all three genomes, a single ribosomal DNA unit is located beside the telomeres on every chromosome.

Prokaryotic genomes vary in size from 0.6 to 13 Mb (ref. 133). This variation, although much smaller than that in eukaryotic genomes, is more than 20-fold. It was proposed that the larger genomes of such organisms as *E. coli* have evolved from smaller ones by successive cycles of genome duplication[134]; however, this hypothesis has received no support. For example, sequence data from the *E. coli* genome show no evidence of genome duplication[135]. In addition, phylogenetic analyses suggest that the increases in genome size occurred independently in different lineages[136] and that bacteria with the smallest genomes are not primitive but derived from bacteria with larger genomes[137].

The current view is that genome size increases through horizontal gene transfer[138,139], duplication of genes or operons[140,141] and duplicative transposition of transposable elements and genes, but how these processes can lead to a large increase in genome size is not well understood. It seems that in bacteria that encounter various habitats and substrates, the genome size can increase through the addition of ecologically relevant genes. For example, the genome of *Streptomyces coelicolor*, which is the largest genome that has been fully sequenced for a bacterium (8.7 Mb), includes many genes that are not found in related mycobacteria (such as those for toxin biosynthesis), enabling it to exploit many different nutrient sources and live in a highly competitive soil environment[141]. The growth of this genome seems to be through the successive addition of genes and DNA fragments by lateral transfer and gene duplication, and the decisive factor is the presence of selection for more diverse metabolic abilities[141].

Unlike in eukaryotes, the genome size variation in bacteria almost directly translates into variation in gene number. Indeed, among the completely sequenced bacterial genomes, a tenfold variation in genome size is reflected by a similar variation in gene number[142,143]. The correspondence between genome size and gene number reflects the compactness of bacterial genomes; that is, there is little nonfunctional DNA in a bacterial genome. This streamlining was thought to confer the advantage of rapid DNA replication[137,144], but cell doubling times show no relationship with genome size[142]. The much higher frequencies of deletions as compared with insertions found in pseudogenes in symbiont and parasitic bacterial genomes (Fig. 3) have been taken as evidence that the compactness of bacterial genomes is largely due to deletional bias[142,145,146].

Deletional bias has been also suggested to be the main cause of gene loss in symbiont and parasitic bacteria[142,143]. In other words, genes are lost in large deletions or inactivated and eroded when selection is not strong enough to maintain them. Indeed, many of the discarded genes encode products (such as tRNAs and components of the DNA recombination and repair pathways) that would seem to be just as useful in parasitic genomes as in other organisms[143,147]. Many such losses might have occurred when the effective population size of a lineage was diminished owing to restricted habitats (hosts) or to bottlenecks at the time of infection. But although the independently derived small genomes approach similar sizes and numbers of genes, they comprise mostly different genes[148].

Future developments

Will the next decade of molecular evolutionary genomics be as exciting as the past one? We think so. The next decade will certainly see an explosion of comparative genome sequencing. As the cost of DNA sequencing falls and the capacity of sequencing centers grows, it will become feasible to investigate the complete genomes of sets of related species. Such a study has been already begun with yeast species, for which the fully sequenced genome of *S. cerevisiae* has provided a reference point for a survey of 13 other yeast species that have been sequenced at low coverage[149], and plans are afoot to sequence completely the genomes of more than a dozen other fungi[150].

The combination of several related sequences and genome-wide transcription data should allow the evolution of regulatory elements to be studied in unprecedented detail. An ambitious project already underway aims to sequence an homologous multi-megabase region from 11 vertebrates[151]. These projects, particularly those that generate vast amounts of low-coverage sequence, will cause a bioinformatics headache in terms of making the data and annotations readily accessible and searchable by the whole community, but they will provide raw materials for understanding the evolution of eukaryotic genomes.

An area that is at last becoming tractable is the divergence of gene expression between duplicate genes, a subject of interest to both geneticists and evolutionists[4,8,152,153]. In the past, studies of expression divergence usually have been limited to a few gene families, thereby providing no general picture of the pace of expression divergence between duplicate genes in a genome.

Fortunately, a broad picture is now achievable, owing to the advent of microarray gene expression technology and the complete sequences of many genomes.

Wagner[154] examined whether expression divergence increases with the protein distance between duplicate genes using microarray data from yeast and concluded that expression divergence and protein sequence divergence are decoupled. But this result does not imply that expression divergence is decoupled from evolutionary time, because protein distance may not be a good proxy of divergence time. Although a protein may evolve at a roughly constant rate among evolutionary lineages, the rate of amino acid substitution varies tremendously among proteins[155,156]; therefore, a single substitution rate cannot be used to date the divergence times of different protein pairs.

By comparison, the rate of synonymous substitution is more uniform among genes[155,156], and a study of the relationship between expression divergence and synonymous distance has indicated that expression divergence increases rapidly with evolutionary time[157]. Because only yeast data have been considered so far, the issue of expression divergence between duplicate genes remains open. Not only do we need to study other species, especially multicellular organisms, to reach a general conclusion, but we also need to develop statistical methods for quantifying gene expression divergence.

Another exciting area is the evolution of cellular networks, such as the protein-protein interaction network[158]. Initial studies show that the rate of evolution of a protein is correlated with the number of partners with which it interacts[159]. Genome-wide studies[159–161] on whether the rate of molecular evolution in a gene is correlated with the phenotypic effect of mutations in the gene are starting to address the old issue of whether protein dispensability affects the rate of protein evolution[162].

More generally, we feel that the molecular evolution community is still struggling to gain a sense of how a whole genome evolves. The study of genomic evolution is still in a 'gold-rush' phase and, rather like the dot.com industry, a period of retrenchment and consolidation may be necessary before we can recognize the truly significant shifts that have taken place. At present, it is not easy to tell which facets of a genome have been shaped by selective pressures (the size of its gene families? its repetitive DNA content? its gene order?) and which are neutral phenomena. It is still difficult to design experiments that can explore adequately the molecular mechanisms underlying evolutionary change.

We are hopeful that further technological advances will lead to a democratization of genomics, whereby the sorts of experiments that are now only feasible for high-priority organisms will become accessible to smaller laboratories and for organisms of more specialized interest, so that 'big' evolutionary questions can be asked in appropriate taxa. The recent choice of the honeybee as a target for genome sequencing[150] is a step in this direction. But there are even bigger pictures that are scarcely being glimpsed at the moment. If we ever think that we are close to understanding how a genome works, or that one mammalian genome is pretty much the same as another, a visit to a zoo will quickly humble us.

Acknowledgments

We thank S. Yi and K. Makova for help, and L. Hurst for comments. This work was supported by grants from the National Institutes of Health (to W.-H.L.) and from Science Foundation Ireland (to K.H.W.)

1. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275–276 (1977).
2. Fitch, W.M. Estimating the total number of nucleotide substitutions since the common ancestor of a pair of genes: comparison of several methods and three β hemoglobin messenger RNAs. *J. Mol. Evol.* **16**, 153–209 (1980).
3. Li, W.-H., Gojobori, T. & Nei, M. Pseudogenes as a paradigm of neutral evolution. *Nature* **292**, 237–239 (1981).
4. Ohno, S. *Evolution by Gene Duplication* (George Allen and Unwin, London, 1970).
5. Betran, E., Wang, W., Jin, L. & Long, M. Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene. *Mol. Biol. Evol.* **19**, 654–663 (2002).
6. Long, M. Evolution of novel genes. *Curr. Opin. Genet. Dev.* **11**, 673–680 (2001).
7. Lynch, M. & Conery, J.S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
8. Force, A. *et al*. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
9. Wolfe, K.H. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* **2**, 333–341 (2001).
10. McLysaght, A., Hokamp, K. & Wolfe, K.H. Extensive genomic duplication during early chordate evolution. *Nat. Genet.* **31**, 200–204 (2002).
11. Gu, X., Wang, Y. & Gu, J. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat. Genet.* **31**, 205–209 (2002).
12. Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P. & Inoko, H. Evidence of *en bloc* duplication in vertebrate genomes. *Nat. Genet.* **31**, 100–105 (2002).
13. Kashkush, K., Feldman, M. & Levy, A.A. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* **160**, 1651–1659 (2002).
14. Comai, L. *et al*. Phenotypic instability and rapid gene silencing in newly formed *Arabidopsis* allotetraploids. *Plant Cell* **12**, 1551–1568 (2000).
15. Lee, H.S. & Chen, Z.J. Protein-coding genes are epigenetically regulated in *Arabidopsis* polyploids. *Proc. Natl. Acad. Sci. USA* **98**, 6753–6758 (2001).
16. Adams, K.L., Daley, D.O., Qiu, Y.-L., Whelan, J. & Palmer, J.D. Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. *Nature* **408**, 354–357 (2000).
17. Adams, K.L., Qiu, Y.L., Stoutemyer, M. & Palmer, J.D. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. *Proc. Natl. Acad. Sci. USA* **99**, 9905–9912 (2002).
18. Kadowaki, K., Kubo, N., Ozawa, K. & Hirai, A. Targeting presequence acquisition after mitochondrial gene transfer to the nucleus occurs by duplication of existing targeting signals. *EMBO J.* **15**, 6652–6661 (1996).
19. Kubo, N., Harada, K., Hirai, A. & Kadowaki, K. A single nuclear transcript encoding mitochondrial *RPS14* and *SDHB* of rice is processed by alternative splicing: common use of the same mitochondrial targeting signal for different proteins. *Proc. Natl. Acad. Sci. USA* **96**, 9207–9211 (1999).
20. Courseaux, A. & Nahon, J.L. Birth of two chimeric genes in the Hominidae lineage. *Science* **291**, 1293–1297 (2001).
21. Eichler, E.E. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17**, 661–669 (2001).
22. Long, M. & Langley, C.H. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* **260**, 91–95 (1993).
23. Wang, W., Brunet, F.G., Nevo, E. & Long, M. Origin of *sphinx*, a young chimeric RNA gene in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **99**, 4448–4453 (2002).
24. Moran, J.V., DeBerardinis, R.J. & Kazazian, H.H. Jr. Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534 (1999).
25. Pickeral, O.K., Makalowski, W., Boguski, M.S. & Boeke, J.D. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* **10**, 411–415 (2000).
26. Korneev, S. & O'Shea, M. Evolution of nitric oxide synthase regulatory genes by DNA inversion. *Mol. Biol. Evol.* **19**, 1228–1233 (2002).
27. Johnson, M.E. *et al*. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).
28. Lipovich, L., Hughes, A.L., King, M.C., Abkowitz, J.L. & Quigley, J.G. Genomic structure and evolutionary context of the human feline leukemia virus subgroup C receptor (hFLVCR) gene: evidence for block duplications and *de novo* gene formation within duplicons of the hFLVCR locus. *Gene* **286**, 203–213 (2002).
29. Chen, L., DeVries, A.L. & Cheng, C.H. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc. Natl. Acad. Sci. USA* **94**, 3811–3816 (1997).
30. Chen, L., DeVries, A.L. & Cheng, C.H. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc. Natl. Acad. Sci. USA* **94**, 3817–3822 (1997).
31. Parkhill, J. *et al*. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413**, 848–852 (2001).
32. McClelland, M. *et al*. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature* **413**, 852–856 (2001).
33. Porwollik, S., Wong, R.M. & McClelland, M. Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis. *Proc. Natl. Acad. Sci. USA* **99**, 8956–8961 (2002).
34. Perna, N.T. *et al*. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529–533 (2001).
35. Malpertuy, A. *et al*. Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes-specific genes. *FEBS Lett.* **487**, 113–121 (2000).
36. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
37. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568–573 (1998).
38. Suzuki, Y. & Gojobori, T. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **16**, 1315–1328 (1999).
39. Suzuki, Y. & Nei, M. Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **18**, 2179–2185 (2001).
40. Yang, Z. & Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19**, 908–917 (2002).
41. Hughes, A.L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170 (1988).
42. Tanaka, T. & Nei, M. Positive darwinian selection observed at the variable-region genes of immunoglobulins. *Mol. Biol. Evol.* **6**, 447–459 (1989).

43. Hughes, A.L., Ota, T. & Nei, M. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Mol. Biol. Evol.* **7**, 515–524 (1990).
44. Riley, M.A. Positive selection for colicin diversity in bacteria. *Mol. Biol. Evol.* **10**, 1048–1059 (1993).
45. Zhang, J., Rosenberg, H.F. & Nei, M. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. USA* **95**, 3708–3713 (1998).
46. Zhang, J., Zhang, Y.P. & Rosenberg, H.F. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat. Genet.* **30**, 411–415 (2002).
47. Hughes, A.L. Circumsporozoite protein genes of malaria parasites (*Plasmodium* spp.): evidence for positive selection on immunogenic regions. *Genetics* **127**, 345–353 (1991).
48. Hughes, A.L. Positive selection and interallelic recombination at the merozoite surface antigen-1 (MSA-1) locus of *Plasmodium falciparum*. *Mol. Biol. Evol.* **9**, 381–393 (1992).
49. Bonhoeffer, S., Holmes, E.C. & Nowak, M.A. Causes of HIV diversity. *Nature* **376**, 125 (1995).
50. Yamaguchi-Kabata, Y. & Gojobori, T. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J. Virol.* **74**, 4335–4350 (2000).
51. Lee, Y.H., Ota, T. & Vacquier, V.D. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. BioI. Evol.* **12**, 231–238 (1995).
52. Metz, E.C. & Palumbi, S.R. Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. *Mol. Biol. Evol.* **13**, 397–406 (1996).
53. Palumbi, S.R. All males are not created equal: fertility differences depend on gamete recognition polymorphisms in sea urchins. *Proc. Natl. Acad. Sci. USA* **96**, 12632–12637 (1999).
54. Aguade, M., Miyashita, N. & Langley, C.H. Polymorphism and divergence in the *Mst26A* male accessory gland gene region in *Drosophila*. *Genetics* **132**, 755–770 (1992).
55. Tsaur, S.C. & Wu, C.I. Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila*. *Mol. Biol. Evol.* **14**, 544–549 (1997).
56. Tsaur, S.C., Ting, C.T. & Wu, C.I. Positive selection driving the evolution of a gene of male reproduction, *Acp26Aa*, of *Drosophila*: II. Divergence versus polymorphism. *Mol. Biol. Evol.* **15**, 1040–1046 (1998).
57. Swanson, W.J., Clark, A.G., Waldrip-Dail, H.M., Wolfner, M.F. & Aquadro, C.F. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **98**, 7375–7379 (2001).
58. Maston, G.A. & Ruvolo, M. Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. *Mol. Biol. Evol.* **19**, 320–335 (2002).
59. Messier, W. & Stewart, C.B. Episodic adaptive evolution of primate lysozymes. *Nature* **385**, 151–154 (1997).
60. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
61. Smith, N.G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022–1024 (2002).
62. Fay, J.C., Wyckoff, G.J. & Wu, C.I. Positive and negative selection on the human genome. *Genetics* **158**, 1227–1234 (2001).
63. Fay, J.C., Wyckoff, G.J. & Wu, C.I. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**, 1024–1026 (2002).
64. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
65. Roberts, J.D., Izuta, S., Thomas, D.C. & Kunkel, T.A. Mispair-, site-, and strand-specific error rates during simian virus 40 origin-dependent replication *in vitro* with excess deoxythymidine triphosphate. *J. Biol. Chem.* **269**, 1711–1717 (1994).
66. Fijalkowska, I.J., Jonczyk, P., Tkaczyk, M.M., Bialoskorska, M. & Schaaper, R.M. Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome. *Proc. Natl. Acad. Sci. USA* **95**, 10020–10025 (1998).
67. Francino, M.P. & Ochman, H. Strand asymmetries in DNA evolution. *Trends Genet.* **13**, 240–245 (1997).
68. Frank, A.C. & Lobry, J.R. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238**, 65–77 (1999).
69. Lobry, J.R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13**, 660–665 (1996).
70. Tillier, E.R. & Collins, R.A. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.* **50**, 249–257 (2000).
71. Rocha, E.P. & Danchin, A. Ongoing evolution of strand composition in bacterial genomes. *Mol. Biol. Evol.* **18**, 1789–1799 (2001).
72. Jermiin, L.S., Graur, D. & Crozier, R.H. Evidence from analyses of intergenic regions for strand-specific directional mutation pressure in metazoan mitochondrial DNA. *Mol. Biol. Evol.* **12**, 558–563 (1995).
73. Perna, N.T. & Kocher, T.D. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *J. Mol. Evol.* **41**, 353–358 (1995).
74. Tanaka, M. & Ozawa, T. Strand asymmetry in human mitochondrial DNA mutations. *Genomics* **22**, 327–335 (1994).
75. Reyes, A., Gissi, C., Pesole, G. & Saccone, C. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol. Biol. Evol.* **15**, 957–966 (1998).
76. Francino, M.P. & Ochman, H. Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. *Mol. Biol. Evol.* **18**, 1147–1150 (2001).
77. Beletskii, A. & Bhagwat, A.S. Correlation between transcription and C to T mutations in the non-transcribed DNA strand. *Biol. Chem.* **379**, 549–551 (1998).
78. Karlin, S., Campbell, A.M. & Mrazek, J. Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* **32**, 185–225 (1998).
79. Francino, M.P. & Ochman, H. Strand symmetry around the β-globin origin of replication in primates. *Mol. Biol. Evol.* **17**, 416–422 (2000).
80. Gierlik, A., Kowalczuk, M., Mackiewicz, P., Dudek, M.R. & Cebrat, S. Is there replication-associated mutational pressure in the *Saccharomyces cerevisiae* genome? *J. Theor. Biol.* **202**, 305–314 (2000).
81. Mrázek, J. & Karlin, S. Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. USA* **95**, 3720–3725 (1998).
82. McLean, M.J., Wolfe, K.H. & Devine, K.M. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.* **47**, 691–696 (1998).
83. Myllykallio, H. *et al*. Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science* **288**, 2212–2215 (2000).
84. McInerney, J.O. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc. Natl. Acad. Sci. USA* **95**, 10698–10703 (1998).
85. Lafay, B. *et al*. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutation biases. *Nucleic Acids Res.* **27**, 1642–1649 (1999).
86. Tillier, E.R. & Collins, R.A. Replication orientation affects the rate and direction of bacterial gene evolution. *J. Mol. Evol.* **51**, 459–463 (2000).
87. Szczepanik, D. *et al*. Evolution rates of genes on leading and lagging DNA strands. *J. Mol. Evol.* **52**, 426–433 (2001).
88. Haldane, J.B.S. The rate of spontaneous mutation of a human gene. *J. Genet.* **31**, 317–326 (1935).
89. Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K. & Yasunaga, T. Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 863–867 (1987).
90. Li, W.H., Ellsworth, D.L., Krushkal, J., Chang, B.H. & Hewett-Emmett, D. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phyl. Evol.* **5**, 182–187 (1996).
91. McVean, G.T. & Hurst, L.D. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature* **386**, 388–392 (1997).
92. Ellegren, H. & Fridolfsson, A.K. Male-driven evolution of DNA sequences in birds. *Nature Genet.* **17**, 182–184 (1997).
93. Makova, K.D. & Li, W.H. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624–626 (2002).
94. Filipski, J. Why the rate of silent codon substitutions is variable within a vertebrate's genome. *J. Theor. Biol.* **134**, 159–164 (1988).
95. Wolfe, K.H., Sharp, P.M. & Li, W.H. Mutation rates differ among regions of the mammalian genome. *Nature* **337**, 283–285 (1989).
96. Matassi, G., Sharp, P.M. & Gautier, C. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**, 786–791 (1999).
97. Gu, Z., Wang, H., Nekrutenko, A. & Li, W.H. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. *Gene* **259**, 81–88 (2000).
98. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci. USA* **99**, 803–808 (2002).
99. Lercher, M.J., Williams, E.J. & Hurst, L.D. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol. Biol. Evol.* **18**, 2032–2039 (2001).
100. Ebersberger, I., Metzler, D., Schwarz, C. & Paabo, S. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70**, 1490–1497 (2002).
101. Bielawski, J.P., Dunn, K.A. & Yang, Z. Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. *Genetics* **156**, 1299–1308 (2000).
102. Smith, N.G. & Hurst, L.D. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153**, 1395–1402 (1999).
103. Nekrutenko, A. & Li, W.H. Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res.* **10**, 1986–1995 (2000).
104. Strathern, J.N., Shafer, B.K. & McGill, C.B. DNA synthesis errors associated with double-strand-break repair. *Genetics* **140**, 965–972 (1995).
105. Perry, J. & Ashworth, A. Evolutionary rate of a gene affected by chromosomal position. *Curr. Biol.* **9**, 987–989 (1999).
106. Gerton, J.L. *et al*. Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* **97**, 11383–11390 (2000).
107. Birdsell, J.A. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol. Biol. Evol.* **19**, 1181–1197 (2002).
108. Takano-Shimizu, T. Local changes in GC/AT substitution biases and in crossover frequencies on *Drosophila* chromosomes. *Mol. Biol. Evol.* **18**, 606–619 (2001).
109. Brown, T.C. & Jiricny, J. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* **54**, 705–711 (1988).
110. Fullerton, S.M., Bernardo Carvalho, A. & Clark, A.G. Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* **18**, 1139–1142 (2001).
111. Blumenthal, T. *et al*. A global analysis of *Caenorhabditis elegans* operons. *Nature* **417**, 851–854 (2002).
112. von Mering, C. & Bork, P. Teamed up for transcription. *Nature* **417**, 797–798 (2002).
113. Wang, P.J., McCarrey, J.R., Yang, F. & Page, D.C. An abundance of X-linked genes expressed in spermatogonia. *Nat. Genet.* **27**, 422–426 (2001).
114. Roy, P.J., Stuart, J.M., Lund, J. & Kim, S.K. Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature* **418**, 975–979 (2002).
115. Spellman, P.T. & Rubin, G.M. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J. Biol.* **1**, 5 (2002).
116. Cohen, B.A., Mitra, R.D., Hughes, J.D. & Church, G.M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* **26**, 183–186 (2000).
117. Kruglyak, S. & Tang, H. Regulation of adjacent yeast genes. *Trends Genet.* **16**, 109–111 (2000).
118. Mannila, H., Patrikainen, A., Seppanen, J.K. & Kere, J. Long-range control of expression in yeast. *Bioinformatics* **18**, 482–483 (2002).
119. Lercher, M.J., Urrutia, A.O. & Hurst, L.D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* **31**, 180–183 (2002).
120. Dehal, P. *et al*. Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293**, 104–111 (2001).
121. Mural, R.J. *et al*. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**, 1661–1671 (2002).
122. Seoighe, C. *et al*. Prevalence of small inversions in yeast gene order evolution. *Proc. Natl. Acad. Sci. USA* **97**, 14433–14437 (2000).

123. Coghlan, A. & Wolfe, K.H. Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. *Genome Res.* **12**, 857–867 (2002).
124. Gilley, J. & Fried, M. Extensive gene order differences within regions of conserved synteny between the *Fugu* and human genomes: implications for chromosomal evolution and the cloning of disease genes. *Hum. Mol. Genet.* **8**, 1313–1320 (1999).
125. SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. & Bennetzen, J.L. The paleontology of intergene retrotransposons of maize. *Nature Genet.* **20**, 43–45 (1998).
126. Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L. & Shaw, K.L. Evidence for DNA loss as a determinant of genome size. *Science* **287**, 1060–1062 (2000).
127. Bensasson, D., Petrov, D.A., Zhang, D.X., Hartl, D.L. & Hewitt, G.M. Genomic gigantism: DNA loss is slow in mountain grasshoppers. *Mol. Biol. Evol.* **18**, 246–253 (2001).
128. Douglas, S. *et al*. The highly reduced genome of an enslaved algal nucleus. *Nature* **410**, 1091–1096 (2001).
129. Gilson, P.R. & McFadden, G.I. The miniaturized nuclear genome of eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed, and the smallest known spliceosomal introns. *Proc. Natl. Acad. Sci. USA* **93**, 7737–7742 (1996).
130. Gilson, P.R. & McFadden, G.I. Jam packed genomes—a preliminary, comparative analysis of nucleomorphs. *Genetica* **115**, 13–28 (2002).
131. Peyret, P. *et al*. Sequence and analysis of chromosome I of the amitochondriate intracellular parasite *Encephalitozoon cuniculi* (Microspora). *Genome Res.* **11**, 198–207 (2001).
132. Katinka, M.D. *et al*. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414**, 450–453 (2001).
133. Graur, D. & Li, W.-H. *Fundamentals of Molecular Evolution* 443 (Sinauer, Sunderland, MA, 1999).
134. Zipkas, D. & Riley, M. Proposal concerning mechanism of evolution of the genome of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **72**, 1354–1358 (1975).
135. Riley, M. & Labedan, B. Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module. *J. Mol. Biol.* **268**, 857–868 (1997).
136. Herdman, M. The evolution of bacterial genomes. In *The Evolution of Genome Size* (ed. Cavalier-Smith, T.) 37–68 (John Wiley and Sons, Chichester 1985).
137. Andersson, S.G. & Kurland, C.G. Reductive evolution of resident genomes. *Trends Microbiol.* **6**, 263–268 (1998).
138. Nelson, K.E. *et al*. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329 (1999).
139. Ochman, H., Lawrence, J.G. & Groisman, E.A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
140. Jordan, I.K., Makarova, K.S., Spouge, J.L., Wolf, Y.I. & Koonin, E.V. Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.* **11**, 555–565 (2001).
141. Bentley, S.D. *et al*. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147 (2002).
142. Mira, A., Ochman, H. & Moran, N.A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**, 589–596 (2001).
143. Moran, N.A. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108**, 583–586 (2002).
144. Maniloff, J. The minimal cell genome: 'on being the right size'. *Proc. Natl. Acad. Sci. USA* **93**, 10004–10006 (1996).
145. Andersson, J.O. & Andersson, S.G. Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. *Mol. Biol. Evol.* **18**, 829–839 (2001).
146. Clark, M.A., Baumann, L., Thao, M.L., Moran, N.A. & Baumann, P. Degenerative minimalism in the genome of a psyllid endosymbiont. *J. Bacteriol.* **183**, 1853–1861 (2001).
147. Andersson, S.G. *et al*. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140 (1998).
148. Koonin, E.V. How many genes can make a cell: the minimal-gene-set concept. *Annu. Rev. Genomics Hum. Genet.* **1**, 99–116 (2000).
149. Souciet, J. *et al*. Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. *FEBS Lett.* **487**, 3–12 (2000).
150. Pennisi, E. Chimps and fungi make genome 'top six'. *Science* **296**, 1589–1591 (2002).
151. Thomas, J.W. & Touchman, J.W. Vertebrate genome sequencing: building a backbone for comparative genomics. *Trends Genet.* **18**, 104–108 (2002).
152. Markert, C.L. Cellular differentiation—an expression of differential gene function. In *Congenital Malformations* 163–174 (International Medical Congress, New York, 1964).
153. Ferris, S.D. & Whitt, G.S. Evolution of the differential regulation of duplicate genes after polyploidization. *J. Mol. Evol.* **12**, 267–317 (1979).
154. Wagner, A. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proc. Natl. Acad. Sci. USA* **97**, 6579–6584 (2000).
155. Li, W.-H. *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).
156. Makalowski, W. & Boguski, M.S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. USA* **95**, 9407–9412 (1998).
157. Gu, Z., Nicolae, D., Lu, H.H.-S. & Li, W.-H. Rapid divergence in expression between duplicate genes inferred from microarray gene expression data. *Trends Genet.* **18**, 609–613 (2002).
158. Wagner, A. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* **18**, 1283–1292 (2001).
159. Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. & Feldman, M.W. Evolutionary rate in the protein interaction network. *Science* **296**, 750–752 (2002).
160. Hirsh, A.E. & Fraser, H.B. Protein dispensability and rate of evolution. *Nature* **411**, 1046–1049 (2001).
161. Papp, B., Pal, C. & Hurst, L.D. Gene dispensibility does not determine the rate of evolution. *Nature* (in the press).
162. Wilson, A.C., Carlson, S.S. & White, T., J. Biochemical evolution. *Annu. Rev. Biochem.* **46**, 573–639 (1977).
163. Hughes, A.L. The evolution of the type I interferon gene family in mammals. *J. Mol. Evol.* **41**, 539–548 (1995).
164. Endo, T., Ikeo, K. & Gojobori, T. Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* **13**, 685–690 (1996).
165. Goodwin, R.L., Baumann, H. & Berger, F.G. Patterns of divergence during evolution of α1-proteinase inhibitors in mammals. *Mol. Biol. Evol.* **13**, 346–358 (1996).
166. Hughes, A.L. & Yeager, M. Coordinated amino acid changes in the evolution of mammalian defensins. *J. Mol. Evol.* **44**, 675–682 (1997).
167. Kitano, T., Sumiyama, K., Shiroishi, T. & Saitou, N. Conserved evolution of the Rh50 gene compared to its homologous Rh blood group gene. *Biochem. Biophys. Res. Commun.* **249**, 78–85 (1998).
168. Wyckoff, G.J., Wang, W. & Wu, C.I. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403**, 304–309 (2000).
169. Qi, C.F. *et al*. Molecular phylogeny of Fv1. *Mamm. Genome* **9**, 1049–1055 (1998).
170. Ford, M.J., Thornton, P.J. & Park, L.K. Natural selection promotes divergence of transferrin among salmonid species. *Mol. Ecol.* **8**, 1055–1061 (1999).
171. Singhania, N.A. *et al*. Rapid evolution of the ribonuclease A superfamily: adaptive expansion of independent gene clusters in rats and mice. *J. Mol. Evol.* **49**, 721–728 (1999).
172. Bishop, J.G., Dean, A.M. & Mitchell-Olds, T. Rapid evolution in plant chitinases: molecular targets of selection in plant–pathogen coevolution. *Proc. Natl. Acad. Sci. USA* **97**, 5322–5327 (2000).
173. Baum, J., Ward, R.H. & Conway, D.J. Natural selection on the erythrocyte surface. *Mol. Biol. Evol.* **19**, 223–229 (2002).
174. Zhang, J. & Nei, M. Positive selection in the evolution of mammalian interleukin-2 genes. *Mol. Biol. Evol.* **17**, 1413–1416 (2000).
175. Hughes, M.K. & Hughes, A.L. Natural selection on *Plasmodium* surface proteins. *Mol. Biochem. Parasitol.* **71**, 99–113 (1995).
176. Smith, N.H., Maynard Smith, J. & Spratt, B.G. Sequence evolution of the *porB* gene of *Neisseria gonorrhoeae* and *Neisseria meningitidis*: evidence of positive Darwinian selection. *Mol. Biol. Evol.* **12**, 363–370 (1995).
177. Baric, R.S., Yount, B., Hensley, L., Peel, S.A. & Chen, W. Episodic evolution mediates interspecies transfer of a murine coronavirus. *J. Virol.* **71**, 1946–1955 (1997).
178. Fitch, W.M., Bush, R.M., Bender, C.A. & Cox, N.J. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl. Acad. Sci. USA* **94**, 7712–7718 (1997).
179. Wu, J.C. *et al*. Recombination of hepatitis D virus RNA sequences and its implications. *Mol. Biol. Evol.* **16**, 1622–1632 (1999).
180. Zanotto, P.M., Kallas, E.G., de Souza, R.F. & Holmes, E.C. Genealogical evidence for positive selection in the *nef* gene of HIV-1. *Genetics* **153**, 1077–1089 (1999).
181. Haydon, D.T., Bastos, A.D., Knowles, N.J. & Samuel, A.R. Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates. *Genetics* **157**, 7–15 (2001).
182. Aguade, M. Positive selection drives the evolution of the Acp29AB accessory gland protein in *Drosophila*. *Genetics* **152**, 543–551 (1999).
183. Hellberg, M.E. & Vacquier, V.D. Rapid evolution of fertilization selectivity and lysin cDNA sequences in teguline gastropods. *Mol. Biol. Evol.* **16**, 839–848 (1999).
184. Swanson, W.J., Aquadro, C.F. & Vacquier, V.D. Polymorphism in abalone fertilization proteins is consistent with the neutral evolution of the egg's receptor for lysin (VERL) and positive darwinian selection of sperm lysin. *Mol. Biol. Evol.* **18**, 376–383 (2001).
185. Pamilo, P. & O'Neill, R.J. Evolution of the *Sry* genes. *Mol. Biol. Evol.* **14**, 49–55 (1997).
186. Vacquier, V.D., Swanson, W.J. & Lee, Y.H. Positive Darwinian selection on two homologous fertilization proteins: what is the selective pressure driving their divergence? *J. Mol. Evol.* **44**, S15–22 (1997).
187. Ishimizu, T. *et al*. Identification of regions in which positive selection may operate in S-RNase of Rosaceae: implication for S-allele-specific recognition sites in S-RNase. *FEBS Lett.* **440**, 337–342 (1998).
188. Karn, R.C. & Nachman, M.W. Reduced nucleotide variability at an androgen-binding protein locus (*Abpa*) in house mice: evidence for positive natural selection. *Mol. Biol. Evol.* **16**, 1192–1197 (1999).
189. Rooney, A.P. & Zhang, J. Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive Darwinian selection? *Mol. Biol. Evol.* **16**, 706–710 (1999).
190. Hellberg, M.E., Moy, G.W. & Vacquier, V.D. Positive selection and propeptide repeats promote rapid interspecific divergence of a gastropod sperm protein. *Mol. Biol. Evol.* **17**, 458–466 (2000).
191. Eanes, W.F., Kirchner, M. & Yoon, J. Evidence for adaptive evolution of the *G6pd* gene in the *Drosophila melanogaster* and *Drosophila simulans* lineages. *Proc. Natl. Acad. Sci. USA* **90**, 7475–7479 (1993).
192. Nakashima, K. *et al*. Accelerated evolution in the protein-coding regions is universal in crotalinae snake venom gland phospholipase A2 isozyme genes. *Proc. Natl. Acad. Sci. USA* **92**, 5605–5609 (1995).
193. Shields, D.C., Harmon, D.L. & Whitehead, A.S. Evolution of hemopoietic ligands and their receptors. Influence of positive selection on correlated replacements throughout ligand and receptor proteins. *J. Immunol.* **156**, 1062–1070 (1996).
194. Wallis, M. The molecular evolution of vertebrate growth hormones: a pattern of near-stasis interrupted by sustained bursts of rapid change. *J. Mol. Evol.* **43**, 93–100 (1996).
195. Liu, J.C., Makova, K.D., Adkins, R.M., Gibson, S. & Li, W.H. Episodic evolution of growth hormone in primates and emergence of the species specificity of human growth hormone receptor. *Mol. Biol. Evol.* **18**, 945–953 (2001).
196. Sutton, K.A. & Wilkinson, M.F. Rapid evolution of a homeodomain: evidence for positive selection. *J. Mol. Evol.* **45**, 579–588 (1997).
197. Ward, T.J., Honeycutt, R.L. & Derr, J.N. Nucleotide sequence evolution at the κ-casein locus: evidence for positive selection within the family Bovidae. *Genetics* **147**, 1863–1872 (1997).
198. Wu, W., Goodman, M., Lomax, M.I. & Grossman, L.I. Molecular evolution of cytochrome *c* oxidase subunit IV: evidence for positive selection in simian primates. *J. Mol. Evol.* **44**, 477–491 (1997).
199. Bargelloni, L., Marcato, S. & Patarnello, T. Antarctic fish hemoglobins: evidence for adaptive evolution at subzero temperature. *Proc. Natl. Acad. Sci. USA* **95**, 8670–8675 (1998).
200. Ting, C.T., Tsaur, S.C., Wu, M.L. & Wu, C.I. A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* **282**, 1501–1504 (1998).

201. Duda, T.F., Jr. & Palumbi, S.R. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. *Proc. Natl. Acad. Sci. USA* **96**, 6820–6823 (1999).
202. Schmidt, T.R., Goodman, M. & Grossman, L.I. Molecular evolution of the *COX7A* gene family in primates. *Mol. Biol. Evol.* **16**, 619–626 (1999).
203. Huttley, G.A. *et al*. Adaptive evolution of the tumour suppressor *BRCA1* in humans and chimpanzees. Australian Breast Cancer Family Study. *Nat. Genet.* **25**, 410–413 (2000).
204. Schmidt, P.S., Duvernell, D.D. & Eanes, W.F. Adaptive evolution of a candidate gene for aging in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **97**, 10861–10865 (2000).
205. Ding, Y.C. *et al*. Evidence of positive selection acting at the human dopamine receptor D4 gene locus. *Proc. Natl. Acad. Sci. USA* **99**, 309–314 (2002).
206. Huang, W., Chang, B.H., Gu, X., Hewett-Emmett, D. & Li, W.H. Sex differences in mutation rate in higher primates estimated from *AMG* intron sequences. *J. Mol. Evol.* **44**, 463–465 (1997).
207. Shimmin, L.C., Chang, B.H. & Li, W.H. Male-driven evolution of DNA sequences. *Nature* **362**, 745–747 (1993).
208. Chang, B.H.-J., Hewett-Emmett, D. & Li, W.-H. Male-to-female ratios of mutation rate in higher primates estimated from intron sequences. *Zool. Studies* **35**, 36–48 (1996).
209. Pecon Slattery, J. & O'Brien, S.J. Patterns of Y and X chromosome DNA sequence divergence during the Felidae radiation. *Genetics* **148**, 1245–1255 (1998).
210. Chang, B.H.-J., Shimmin, L.C., Shyue, S.K., Hewett-Emmett, D. & Li, W.H. Weak male-driven molecular evolution in rodents. *Proc. Natl. Acad. Sci. USA* **91**, 827–831 (1994).
211. Chang, B.H.-J. & Li, W.H. Estimating the intensity of male-driven evolution in rodents by using X-linked and Y-linked *Ube1* genes and pseudogenes. *J. Mol. Evol.* **40**, 70–77 (1995).
212. Kahn, N.W. & Quinn, T.W. Male-driven evolution among Eoaves? A test of the replicative division hypothesis in a heterogametic female (ZW) system. *J. Mol. Evol.* **49**, 750–759 (1999).
213. Carmichael, A.N., Fridolfsson, A.K., Halverson, J. & Ellegren, H. Male-biased mutation rates revealed from Z and W chromosome-linked ATP synthase α-subunit (*ATP5A1*) sequences in birds. *J. Mol. Evol.* **50**, 443–447 (2000).
214. Parkhill, J. *et al*. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**, 665–668 (2000).