Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution

International Chicken Genome Sequencing Consortium*

**Lists of participants and affiliations appear at the end of the paper*

We present here a draft genome sequence of the red jungle fowl, *Gallus gallus*. Because the chicken is a modern descendant of the dinosaurs and the first non-mammalian amniote to have its genome sequenced, the draft sequence of its genome—composed of approximately one billion base pairs of sequence and an estimated 20,000–23,000 genes—provides a new perspective on vertebrate genome evolution, while also improving the annotation of mammalian genomes. For example, the evolutionary distance between chicken and human provides high specificity in detecting functional elements, both non-coding and coding. Notably, many conserved non-coding sequences are far from genes and cannot be assigned to defined functional classes. In coding regions the evolutionary dynamics of protein domains and orthologous groups illustrate processes that distinguish the lineages leading to birds and mammals. The distinctive properties of avian microchromosomes, together with the inferred patterns of conserved synteny, provide additional insights into vertebrate chromosome architecture.

Genome sequence comparison is a modern extension of the long-standing use of other species as models to illuminate aspects of human biology and medicine. Large-scale genome analyses also highlight the evolutionary dynamics of selective and mutational processes at different chronological scales[1–4]. We present here results obtained from an extensive analysis of a draft sequence of the chicken genome, which has evolved separately from mammalian genomes for ~310 million years (Myr)[4,5] (Fig. 1). This genome is the first to be sequenced at this particular evolutionary distance from humans, and, as shown previously[6–8], it provides an excellent signal-to-noise ratio for the detection of functional elements. Our analysis of the 6.6 × coverage draft sequence of the chicken genome resulted in the following main observations.

- The nearly threefold difference in size between the chicken and mammalian genomes reflects a substantial reduction in interspersed repeat content, pseudogenes and segmental duplications within the chicken genome.
- Chicken–human aligned segments tend to occur in long blocks of conserved synteny. We find a relatively low rate of chromosome translocations in both lineages from the last common ancestor, whereas intrachromosomal rearrangements (for example, inversions) are more common.
- Syntenic relationships for certain classes of non-coding RNA (ncRNA) genes differ from those of protein-coding genes. This observation implies a novel mode of evolution for some ncRNA genes.
- Expansion and contraction of multigene families seem to have been major factors in the independent evolution of mammals and birds.
- The sizes of chicken chromosomes, which span a range of nearly two orders of magnitude, correlate negatively with recombination rate, G+C and CpG content, and gene density but positively with repeat density.
- Synonymous substitution rates are elevated for genes in both chicken microchromosomes and in subtelomeric regions of macrochromosomes.
- There is a paucity of retroposed pseudogenes in the chicken genome, in contrast to mammalian genomes, greatly simplifying the classification of chicken gene content. This is explained by the

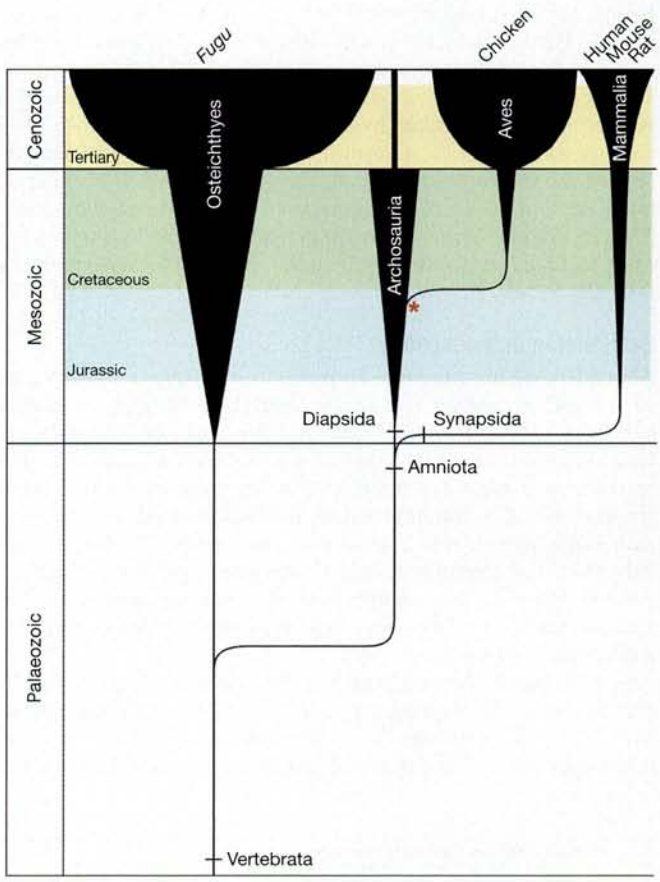


Figure 1 Basal vertebrate evolution showing extant species whose genomes have been sequenced. The horizontal axis represents estimated relative species diversity. The Archosauria include the Aves, their Mesozoic dinosaur predecessors, and Crocodilia; the Lepidosauria (lizards, snakes and tuataras) are not indicated. Archaeopteryx (indicated by an asterisk) is considered to be the first known bird and lived approximately 150 Myr ago. See also ref. 159.

high specificity of the reverse transcriptase from the predominant interspersed repeat element in the chicken genome: the CR1 long interspersed nucleotide element (LINE).

- Unlike all other vertebrate genomes studied so far, no short interspersed nucleotide elements (SINEs) have been active in the chicken genome for the last ~50 Myr.
- Alignment of the chicken and human genomes identifies at least 70 megabases (Mb) of sequence that is highly likely to be functional in both species.
- Many of the chicken–human aligned, non-coding sequences occur far from genes, frequently in clusters that seem to be under selection for functions that are not yet understood.

Perspectives on the domestic chicken

The chicken (*Gallus gallus*) is an important model organism that bridges the evolutionary gap between mammals and other vertebrates and serves as the main laboratory model for the ~9,600 extant avian species. The chicken also represents the first agricultural animal to have its genome sequenced. Modern birds (Ornithurae) evolved from therapod dinosaurs[9,10] in the middle of the Mesozoic era (Fig. 1). Chickens were domesticated in Asia at least by 5400 BC, perhaps as early as 8000 BC[11–13]. Darwin[14] suggested that the red jungle fowl was the nearest ancestor to the domestic chicken, a view later confirmed by mitochondrial DNA analysis[15].

Genetic analysis of the chicken dates back to the start of the twentieth century[16,17], and hundreds of well-characterized mutant stocks and inbred lines have been developed[18]. The chicken embryo has been an especially useful vertebrate system for developmental biologists[19] owing to experimental advantages of *in ovo* embryogenesis. Furthermore, the chicken has been used in seminal studies in virology, oncogenesis and immunology[20–22]. The chicken genetic linkage map, initiated early in the last century[23], now includes 2,172 genetic loci with a total length near 4,000 cM[24,25]. Most avian karyotypes contain chromosomes of markedly different lengths, termed the macro- and microchromosomes, and thus bird karyotypes are quite distinctive as compared with those of mammals[26]. The chicken karyotype ($2n = 78$) is made up of 38 autosomes and one pair of sex chromosomes, with the female as the heterogametic sex (ZW female, ZZ male).

Sequencing and assembly

All sequencing libraries were prepared from DNA of a single female of the inbred line of red jungle fowl (UCD 001) to minimize heterozygosity and provide sequence for both the Z and W sex chromosomes, albeit at 50% of the autosomal coverage. The assembly was generated from ~6.6 × coverage in whole-genome shotgun reads, a combination of plasmid, fosmid and bacterial artificial chromosome (BAC)-end read pairs (Supplementary Table S1). The assembly (Table 1) was generated using PCAP[27], a parallel algorithm that exploits both read-pairing constraint information and base quality values (see Supplementary Information for a description of the methods).

A BAC-based physical map for the chicken was developed in parallel with the sequence assembly[28]. Along with the genetic map[25,29–31], this provides the main scaffolding for the assembly into larger ordered and oriented groupings ('ultracontigs') as well as the mechanism for chromosomal assignment (see Methods). After integrating data from the physical map with the assembly, several additional steps were taken to improve the initial assembly of chicken chromosome sequences. This included using expressed sequence tag (EST) and messenger RNA data to aid the ordering and orientation of sequence, and using map and sequence data to aid in localization of centromeres and telomeres (see Methods). The resulting assembly consists of 574 segments made up of 84 ultracontigs (ordered and oriented by their relationship to the physical map) and 490 'supercontigs' (ordered and oriented by read-pairing data, but not linked to the physical map) anchored to chicken chromosomes. Of the 1.05 gigabases (Gb) of assembled sequence, 933 Mb were localized to specific chromosomes, 907 Mb of which were ordered and oriented along those chromosomes.

Table 1 **Whole-genome assembly statistics**

Genome feature	>1 kb number	N50 length (kb)	N50 number	Largest (kb)
Contigs	98,612	36	7,486	442
Supercontigs	32,767	7,067	37	33,505

Statistics presented are for the whole-genome assembly before integration of physical mapping data. Contigs are contiguous sequences not interrupted by gaps, and supercontigs are ordered and oriented contigs including estimated gap sizes. The N50 statistic is defined as the largest length L such that 50% of all nucleotides are contained in contigs of size at least L. A total of 10,743,700 reads were included in the final assembly. Only 4.39% of the total sequencing reads presented to the assembler were not used in the final assembly.

Assessment of the coverage and quality of the genome assembly

We estimated the coverage of the assembly using both finished BACs and available mRNA sequences (Methods). In a set of 38 finished autosomal BAC sequences from the same red jungle fowl female (covering 6 Mb of sequence), 98% of finished bases could be aligned with the draft whole-genome shotgun assembly, with an overall substitution rate of 0.02% and no deletions or insertions. Similarly, of a set of 23,212 chicken mRNAs and 485,000 ESTs[32], 97% and 96% respectively are at least partially found in the assembly. Of these, 10% are only partially found or are fragmentary. This lack of contiguity contributes in part to the 5–10% of genes estimated to be absent from the Ensembl chicken gene set (see below). Representation of the (G+C)-rich extremes of the genome may be less complete. In one small region of incompletely sequenced BACs (3.6 Mb) orthologous to human chromosome 19 (HSA19) (I. Ovcharenko *et al.*, unpublished data), where the average G+C content was 52% (with some regions exceeding 60%), coverage fell to 82%. Furthermore, we examined a set of 400 genes that were represented in chicken mRNA or ESTs and had single orthologues in five diverse species (human, mouse, rat, *Takifugu rubripes* (*Fugu*) and *Drosophila*) but were predicted to be absent from, or at least partially truncated in, the chicken Ensembl gene set (see below). Over 70% were in fact partially found in the assembly. Of the 400, the largest fraction missing (21%) were HSA19 orthologues from a region known to be unusually rare in chicken clone libraries (L. Gordon *et al.*, unpublished data). The missing genes have a higher G+C content than average and many, including some HSA19 orthologues, are associated with intronic simple sequence repeats (see Methods).

Comparisons to 6 Mb of finished red jungle fowl BAC clone sequence revealed alignment with 311 chicken contigs from 62 supercontigs, which were used to assess possible ordering errors (see Methods). No orientation problems were detected, and only two order discordances (misordered sequence contigs within a supercontig) were discovered. This would extrapolate to a total of ~400 kilobases (kb) of misordered contigs in the current assembly. In addition, eight cases were found in which a contig was incorrectly inserted into a supercontig, equivalent to ~1 Mb of incorrect insertions in the full genome.

Recent duplications are especially difficult to place within whole genome assemblies. Relaxed assembly may collapse duplicated segments into one, and stringent assembly may break sequences into duplicates because of sequencing errors. In the chicken, 'all-versus-all' comparison shows that 11% (~123 Mb) of the genome sequence is in pairwise alignments larger than 1 kb with more than 90% sequence identity. The bulk (91% of the 11%), however, are highly similar (>98%) and might represent false duplication. In a direct test (see Methods), only 22% of duplications with near-perfect sequence identity (>98%) and 26% (32.3 out of 122.7 Mb) of the full set were confirmed.

Because the assembly process incorporated genetic markers (Supplementary Table S2), the genetic map does not provide

independent assessment. However, recent placement of 142 additional genetic markers, mapped after the assembly, suggests that less than 0.75% (~6 Mb) of the sequence has been assigned to a wrong chromosome. Thus, the assembly correctly places the vast majority of the chicken genome in long contiguous stretches. It is well ordered and oriented and faithfully represents older segmental duplications (at the cost of a modest false increase in the most recent duplications). The draft provides an excellent substrate for initial global analysis, recognizing that the elucidation of the full sequence will be critical to allow final, definitive conclusions.

Gene content of the chicken genome

The genome sequence of an organism encodes both ncRNAs and proteins. Extensive analysis of the genome sequences of human[1], mouse[2] and rat[3] has provided our current best assessment of mammalian gene content and has illuminated much about the evolution of genes. The chicken genome provides new perspectives on both the structure and content of mammalian genes, as well as yielding insight into avian gene content and evolution of ncRNA genes.

Non-coding RNA genes

A total of 571 ncRNA genes, from over 20 distinct gene families, were identified within the chicken genome assembly (Table 2) using bioinformatic approaches[33,34] (see Methods). Predicted ncRNA pseudogenes are greatly reduced in number relative to their human ncRNA counterparts. The chicken ncRNA predictions therefore represent a set that is mainly functional. If ncRNA genes maintain their placement with respect to neighbouring genes, chicken ncRNA gene locations could be used to identify which mammalian copies are likely to be functional and which are probable pseudogenes. However, few chicken and human ncRNA genes are paired in regions of conserved synteny (Table 2), relative to the high level of shared gene order observed for protein-coding genes (see below). Those classes of ncRNAs that are most often syntenic are microRNAs (miRNAs) and small nucleolar RNAs (snoRNAs), which are often found in the introns of protein-coding genes (or, rarely, of specialized 'host' genes[35]). Most ncRNA genes thus seem to have been translocated to distant genomic sites during vertebrate evolution, without accumulating large numbers of pseudogenes, as would be expected were this process to occur via retrotransposition. This is also in contrast to duplication of genes via unequal crossing over, which results in tandem copies. These insights will require considerably more analysis for a definitive resolution, but it seems that these ncRNAs may not use the same duplication and/or translocation mechanisms as protein-coding genes.

Development of a protein-coding gene set

An evidence-based system (Ensembl[36]) and two comparative gene prediction methods (Twinscan[37] and SGP-2 (ref. 38)) together predicted a common set of 106,749 protein-coding exons, with 85,929 additional exons predicted by one or two methods (Supplementary Table S3). Particular attention was paid to the identification of selenoproteins, which are usually mispredicted in annotated genomes because of their usage of the TGA codon, usually a stop codon, to code for the amino acid selenocysteine (see Methods). Of the human genes predicted using chicken as the "informant", only 311 genes predicted by SGP-2 are absent from previously identified sets (namely, Vega[40], Ensembl[41], RefSeq[42], MGC[43] and H-Invitational[44]) and have homologous chicken predictions that possess orthologous intron positions. These data, and those of another study (E. Eyras *et al.*, unpublished data), suggest that most of the protein-coding genes conserved among vertebrates are represented in existing complementary DNA sets.

We tested the sensitivity and specificity of the chicken gene predictions. Sensitivity was assessed by comparing predicted exons to those of chicken cDNAs[32] representing long open reading frame (ORF)-containing protein-coding genes (Table 3). All three methods correctly predicted about 80% of cDNA-based exons with >80% coverage. An independent SAGE-based analysis (ref. 166, and M. B. Wahl *et al.*, unpublished data) provided a similar, although marginally lower, estimate. Specificity was assessed by testing random exon pairs from the prediction sets using polymerase chain reaction with reverse transcription (RT–PCR) (E. Eyras *et al.*, unpublished data, and ref. 44). Briefly, Ensembl predictions

Table 2 **Families of ncRNA genes in the chicken genome**

RNA type	Number in chicken	Number in human	Chicken in synteny*	Conserved syntenyt	Function
tRNA	280	496‡	158	52 (33%)	Protein synthesis
5S rRNA	12	301	4	0 (0%)	Protein synthesis
5.8S rRNA	3	9	1	0 (0%)	Protein synthesis
18S rRNA	0	0‡	–	–	Protein synthesis
28S rRNA	0	0‡	–	–	Protein synthesis
U1	18	146			Spliceosome
U2	6	88			Spliceosome
U4	4	119			Spliceosome
U5	9	36			Spliceosome
U6	15	821	45	9 (20%)	Spliceosome
U4atac	1§	1‡			Minor spliceosome
U6atac	4§	5‡			Minor spliceosome
U11	1§	1‡			Minor spliceosome
U12	1	2			Minor spliceosome
miRNA	121	191	87	77 (89%)	Translation repression
snoRNA	83§	245‡	63	50 (79%)	rRNA/snRNA processing
RNase P	1	1			tRNA 5′-end processing
U7	1	184			Histone mRNA 3′-end processing
SRP	3	12			Protein secretion
7SK	4	166	12	7 (58%)	Translational regulation (?)
Y	2	739			Ro RNP component
Telomerase RNA	1	1			Telomerase
BIC	1	1	–	–	Unknown

ncRNA genes were predicted as described in the Methods, except where indicated. Some human gene counts include significant numbers of pseudogenes.
*The number of chicken predictions located in conserved blocks that have defined syntenic regions in human (grouped into classes).
†The proportion of chicken predictions that have a syntenic human prediction.
‡Human genome ncRNA predictions from T. Jones and S. R. Eddy (ftp://ftp.genetics.wustl.edu/pub/eddy/annotation/human-hg16/). This human set contains 7,196 ncRNAs, 6,124 of which are putative pseudogenes.
§Chicken ncRNA genes identified by homology.

Table 3 **Sensitivity of gene prediction**

Feature	Ensembl	Twinscan	SGP-2
Exact exon (%)	61	53	60
80% coverage exon (%)	85	77	85
Total exons	179,084	195,665	203,834

Sensitivity of gene predictions as measured by comparison to ORF-containing cDNAs. Numbers are the percentage of coding exons from the cDNA-based models found by the three prediction systems. The sensitivity numbers are quoted at two levels: exact exon prediction and >80% coverage of the cDNA exon.

have a false positive rate of ~4%. When an exon pair is predicted by any two of the three methods (predominantly joint Twinscan plus SGP-2 exons) ~50% are confirmed, suggesting that some genes are missing from the Ensembl set, but we cannot reliably distinguish these from a similarly large number of Twinscan plus SGP-2 false positives. Using our estimates of specificity and sensitivity, we predict a total of between 20,000 and 23,000 protein-coding genes in chicken, with 80–90% of these found in the present Ensembl set (see Methods). This estimate overlaps the lower bounds in the corresponding ranges for mammalian genomes determined by similar calculations (for example, see refs 2, 3, 45).

Evolutionary conservation of gene components

Alignments of chicken and human orthologous protein-coding genes demonstrate the expected pattern of sequence conservation, with highest identity in protein-coding exons and minimal identity in introns (Fig. 2). These alignments allowed us to examine sequence conservation at different sites within genes.

Alignments of coding regions often did not extend to the previously annotated human protein start codons. Rather, we observed a fourfold increase in the frequency of methionine at the first position of the alignment (Fig. 3), suggesting that these internal ATG codons could be the true start sites for at least some of ~2,000 human genes. For these proteins, the overall distribution of amino acids upstream of the end of the alignment in human was markedly different from that downstream and was more consistent with a codon distribution derived from non-coding nucleotide sequence. Using this comparative signal and other features, such as the Kozak sequence[46], we can potentially improve the annotation of mammalian protein-coding start sites.

Sequence conservation around mammalian splice sites can be predicted by divergence at unselected (non-consensus) base pairs at the neutral rate, coupled with purifying selection on sites matching the splice site consensus[47]. Given the high level of neutral site divergence that has occurred between mammalian and chicken orthologous sequences (see neutral evolutionary rate, below), one would expect that orthologous mammalian–chicken splice sites should show a level of conservation no different from that of any unrelated pair of splice sites. However, in contrast to analogous comparisons within the mammalian lineage, there is a detectable signal in orthologous splice site comparisons beyond the consensus derived from comparing non-orthologous splice sites (Supplementary Fig. S1). This suggests that either some subtle classes of splice site sequences are conserved beyond the generic consensus that can only be observed at the bird–mammal evolutionary distance, or that there is a significant but weak conservation in mammalian introns that is not detectable in mammalian–bird alignments[48].

To explore the role of conserved non-coding sequence segments that are probably regulators of protein-coding genes, we examined the frequency of non-coding alignments of at least 100 base pairs (bp) in, respectively, the 5′ flanking region, 5′ untranslated region (UTR), at least one intron, 3′ UTR, or 3′ flanking region (see Methods) within human–chicken orthologue pairs in relation to gene function (as determined by gene ontology (GO) category, Table 4). Some GO categories (for example, development and transcriptional regulation) showed enrichment for conservation in all five regions, suggesting that conserved regulatory signals exist within all of these locations. However, other categories showed more specific patterns. As one example, introns of ion channel genes are particularly enriched for conserved sequences, in agreement with reports that such introns contain RNA-editing targets[49,50].

Pseudogenes and retroposed copies in the chicken genome

Only 51 duplicates of protein-coding genes probably formed by retroposition (that is, exhibiting loss of introns)[51] were identified in the chicken genome, in contrast to the more than 15,000 cases observed in mammalian genomes[3,52]. In mammals, the ancient LINE1 (L1) transposable element is responsible for the origin of most if not all retroposed (pseudo) genes[53]. Although birds host their own LINE-like elements (chicken repeat 1 (CR1); see below)[54], the reverse transcriptase encoded by these elements is unlikely to copy polyadenylated mRNAs[55], probably explaining the paucity of processed pseudogenes in chicken. Within the set of 51 (Supplementary Table S4), 36 clearly represent pseudogenes, because their former coding regions are disabled by alterations (including frameshifts and premature stop codons) that preclude protein function. Among the remaining 15 elements, eight show strong evidence for selective constraint (Supplementary Table S4) and therefore may represent functional retroposed genes. We found no

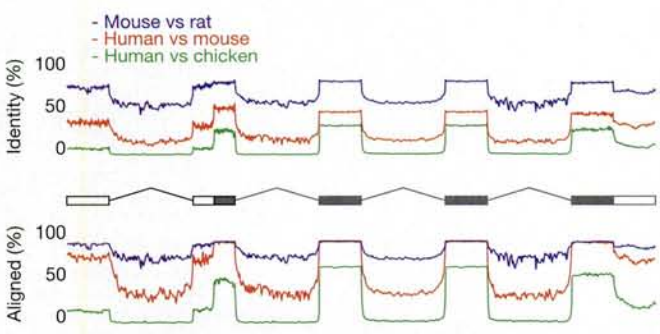


Figure 2 An idealized protein-coding gene structure showing average percentage alignment and average percentage identity (including gaps and unaligned regions) over 10,000 orthologous gene structures in either human–chicken, human–mouse or mouse–rat alignments (as aligned by BLASTZ[156]). The reference structure was taken from human or mouse, and only those with cDNA-based definitions of the structure were used. The central figure shows an idealized gene structure, with the grey exons representing coding sequence and white boxes representing 3′ and 5′ untranslated regions.

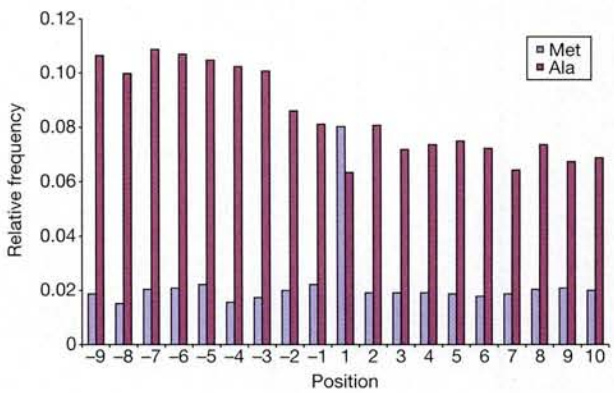


Figure 3 Histogram of amino acid distributions centred on the start of human–chicken alignments where the alignment is >30 amino acids from the putative translation start in human and less than 100 amino acids in length, using the human protein sequence. Alanine is shown as an example of non-methionine amino acids: many amino acids show significant changes before compared with after the alignment.

Table 4 **Conservation near human genes**

5′ flank	5′ UTR	Intron	3′ UTR	3′ flank	GO
0.99	0.95	9.7×10^{-13}	0.020	0.12	Adenyl nucleotide binding
0.99	0.96	0.0	1.7×10^{-2}	0.13	ATP binding
0.13	0.1	5.1×10^{-8}	8.4×10^{-5}	2.5×10^{-2}	Calmodulin binding
6.5×10^{-9}	1.1×10^{-6}	9.3×10^{-8}	1.1×10^{-6}	1.5×10^{-7}	Development
5.8×10^{-2}	9.9×10^{-4}	3.8×10^{-11}	8.2×10^{-4}	1.6×10^{-4}	Ion channel activity
1.0	1.0	1.1×10^{-10}	7.4×10^{-2}	0.42	Motor activity
6.0×10^{-5}	8.9×10^{-4}	1.6×10^{-5}	2.4×10^{-2}	0.10	Muscle development
2.7×10^{-3}	1.0×10^{-3}	1.7×10^{-5}	1.9×10^{-8}	6.5×10^{-9}	Neurogenesis
0.97	0.67	1.2×10^{-12}	1.8×10^{-5}	0.13	Protein metabolism
0.85	0.91	1.2×10^{-12}	6.3×10^{-4}	0.29	Protein-tyrosine kinase activity
1.4×10^{-6}	1.1×10^{-4}	5.7×10^{-4}	9.8×10^{-13}	2.3×10^{-10}	Regulation of transcription
1.2×10^{-9}	6.3×10^{-2}	1.0	1.0	0.20	Rhodopsin-like receptor activity
8.0×10^{-2}	8.5×10^{-2}	3.0×10^{-5}	1.1×10^{-4}	5.9×10^{-3}	Steroid hormone receptor activity
1.0	1.0	1.2×10^{-3}	8.7×10^{-10}	2.3×10^{-2}	Ubiquitin-protein ligase activity

Enrichment for conservation with chicken in several classes of gene-proximal regions for some GO categories of human RefSeq genes. *P*-values (that is, probabilities that enrichments as strong or stronger than the observed ones may occur by chance alone; see Methods) are shown.

clear bias towards either particular gene families or chromosomal locations for the retrocopies (Supplementary Table S4).

Interspersed repeat content of the chicken genome

Interspersed repeats are predominantly copies of transposable elements in various stages of decay. Less than 9% of the chicken genome could be classified as such, and only 11% was annotated when satellites or any lower-copy-number repetitive DNA segments were included (Table 5; see also Methods). This is markedly lower than the 40–50% interspersed repeat density observed in mammalian genomes[1,2], and leaves, with coding regions comprising another 4%, over 85% (~900 Mb) of the current assembly unexplained. This large amount of unexplained genetic matter—which could primarily constitute ancient transposable elements that have mutated beyond recognition—and the high divergence (age) of many recognized repeats (Supplementary Fig. S2) suggest that the low interspersed repeat density in chicken is due to low (recent) transposable element activity rather than to a high deletion rate.

Most of the interspersed repeats are CR1 LINE copies

A single type of non-LTR (long terminal repeat) retrotransposon or LINE, CR1 (refs 54, 56, 57), comprises over 80% of all interspersed repeats in the chicken genome (200,000 copies). CR1 resembles the mammalian L1 element in having a (G+C)-rich internal promoter region, followed by two ORFs. A full-length CR1 element is 4.5 kb, but all but 0.6% of the CR1 copies are truncated from their 5′ end. As sequences near the 5′ end are needed for retrotransposition, the success of CR1 in the bird genome seems remarkable. Comparison of the length distribution of CR1 and L1 copies (Supplementary Fig. S3) suggests a higher efficiency of the L1 reverse transcriptase or higher stability of the L1 transcript[55]. CR1-like elements, unlike L1 elements, do not create target site duplications, probably explaining the absence of copies with 5′ inversions so common for L1 (Supplementary Fig. S3). It is unclear whether CR1 is currently active in the chicken. We found only one full-length element with intact ORFs in the chicken genome: a copy at chromosome 6 (GGA6; 661,970–666,111) that is >2% divergent from the CR1-F consensus or any other element. However, long, minimally divergent CR1 copies are often interrupted by sequence gaps in the present assembly, and a still-active CR1 source gene could have been collapsed and thus missed.

We reconstructed 11 complete CR1 source genes from the copies in the chicken assembly, although the abundance of 3′ end fragments allowed further division into 22 subfamilies. Figure 4 shows the phylogenetic relationship of the 11 source genes based on the alignment of their ORF2 product. The evolution of CR1 in the bird genome seems to differ from that of L1 in mammals, in that several widely divergent elements have been active in parallel, whereas in mammals a single lineage of L1 has been dominant[58,59]. The most recently active CR1 elements in chicken (CR1-F and CR1-B) are less than 70% identical over their ORF2 coding region, whereas the human and mouse L1 ORF2 products are 78% identical. This high divergence between subfamilies probably led to earlier estimates of only 30,000 (ref. 57) and 100,000 (ref. 55) copies, and the current estimate of 200,000 copies is also probably low. The location of the turtle CR1 element in the ORF2-based phylogenetic tree suggests that the main branches of chicken CR1 elements may predate the turtle–bird speciation. Otherwise the tree follows species phylogeny, suggesting that CR1 elements are ancient, vertically transmitted inhabitants of vertebrate genomes.

The most consequential difference between CR1 and L1 elements is in their 3′ end structures: the L1 3′ UTR has evolved with seemingly little constraint on the primary sequence except the polyadenylated tail[59], whereas the CR1 3′ UTR is remarkably conserved between all derived subfamilies[55] and ends with a $(ATTCTRTG)_n$ microsatellite in all chicken CR1 subfamilies, as well as in the turtle CR1 and the ancient L3 element. The CR1 reverse transcriptase presumably has high substrate specificity, whereas the L1 proteins are known to be highly promiscuous. All polyadenylated transcripts in mammals are potential substrates for the L1 reverse transcriptase, and mammalian genomes are littered with processed pseudogenes and SINEs that have been retroposed by the L1 machinery. As noted above, processed pseudogenes are rare in the chicken genome.

Table 5 **Composition of interspersed repeats in the chicken genome**

Repeat type	Copy number	Density				
		Overall (%)	Macro (%)	Micro (%)	Z (%)	Unassigned (%)
CR1	205,000	6.4	7.4	3.2	10.4	8.0
MIRs/LINE2	10,000	0.1	0.1	0.1	0.1	0.1
LTR elements	12,000	1.3	1.3	0.5	1.8	3.3
DNA transposons	13,000	0.8	1.0	0.3	1.5	0.8
Simple repeats	12,000	0.7	0.6	0.4	0.7	1.5
Satellites	2,000	0.1	<0.1	<0.1	<0.1	0.9
Total	254,000	9.4	10.2	4.5	14.5	14.5

'Macro' represents the five largest chromosomes; 'Micro' represents chromosomes smaller than and including chromosome 12; 'Z' is the male sex chromosome; and 'Unassigned' are sequences unassigned to a chromosome.

Missing SINEs

SINEs are small, non-autonomous retroposons derived from structural RNAs; they contain an internal polymerase III promoter and, generally, a 3′ end derived from a LINE-like element in the same genome, which when transcribed is recognized by the LINE machinery for transposition. In all vertebrate and most other animal genomes studied thus far, at least one recently or currently active SINE family has been found, mostly derived from a transfer RNA and often constituting the most numerous interspersed repeat. It is therefore remarkable that the abundant chicken CR1 element

does not seem to be associated with a single SINE, especially as the closely related CR1 element in turtles[60] and L3 in ancestral mammals are or were paired with SINEs. There are about 10,000 faint matches in the chicken genome to MIR and MIR3 (the SINEs associated with L2 and L3, respectively), which originated before the mammal–bird speciation, but unlike in mammals, no new SINEs seem to have replaced these ancient elements upon their extinction.

Retrovirus-like elements

As in mammalian genomes, all retrotransposons with LTRs in the chicken genome belong to the vertebrate-specific class of retroviruses, whereas no copies of gypsy or copia retrotransposons were found. We reconstructed 14 internal and 41 LTR sequences for endogenous retroviruses (ERVs) or their non-autonomous companions, with representatives of all three recognized subclasses (class I to III). The assembly contains one of the two copies of the avian sarcoma/leukosis virus provirus known to exist in the sequenced genome (see Methods). The class II chicken GGERVK10 and class III GGERVL endogenous retroviruses may still be active, as their copies in the chicken genome are less than 3% diverged. We reconstructed several subfamilies of the most abundant ERV, GGERVL, which is most closely related to the mammalian ERV-L[61]. Considering the long descent of ERVL subfamilies in mammals[62], it is tempting to propose that GGERVL and its mammalian relatives have been endogenous in both lineages since the mammalian–bird split. However, all GGERVL subfamilies are young (between 0% and 13% divergence), and the nature of retroviruses suggests that independent introduction in the germ line is a valid possibility as well.

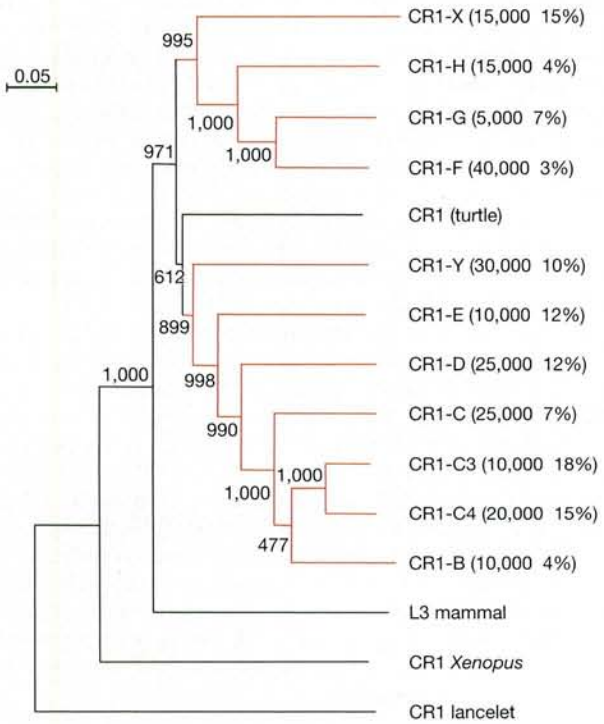


Figure 4 Neighbour-joining tree showing the phylogenetic relationship of the 11 major chicken CR1 subfamilies, CR1 in the *Platemys spixii* turtle (GI:2317255), the *Branchiostoma floridae* lancelet (GI:17529693), *Xenopus laevis* and the ancient mammalian L3 (ref. 160), based on the multiple alignment of the ORF2 products (we derived consensus sequences for the chicken and *Xenopus* CR1s and L3 with complete ORF2 products). Chicken lineages are indicated in red. Bootstrap values are indicated, as well as (in parentheses) the observed copy number of each subfamily and the average substitution level of the copies compared to the consensus. The nomenclature for chicken CR1 is an extension from ref. 161, which defined the subfamilies CR1-A to CR1-F on the basis of the 3′-end fragments.

DNA transposons

Only two ancient DNA transposon families were found in the chicken genome: the unrelated activator-like Charlie12_GG and mariner-like GGMAR elements. Thus, like that of mammals[1,63], the bird germ line may be protected from infiltration by DNA transposons. Both Charlie12_GG and GGMAR left copies that are now ~16% diverged from the consensus, and appear to have been active contemporaneously. This is apparent from a hybrid interspersed repeat consisting of a GGMAR copy within a Charlie12_GG element, from which we infer that the mariner-like element must have transposed into the activator-like element when the latter was still active.

Protein content evolution of chicken and mammalian genomes

Conservation of vertebrate domain and protein content

About 60% of chicken protein-coding genes have a single human orthologue (Fig. 5); for the remainder, orthology relationships are more complex or are not detectable. Chicken and human 1:1 orthologue pairs exhibit lower sequence conservation (median amino acid identity of 75.3%) than rodent and human 1:1 orthologue pairs[3] (~88%), as expected (Fig. 6a). Orthologous sequences involved in cytoplasmic and nuclear functions are more conserved than those implicated in reproduction, host defence and adaptation to the environment (Fig. 6b). Sequence conservation of expressed chicken genes, in comparison to their human orthologues, is non-uniformly distributed among chicken tissues. Sequences expressed in the chicken brain, as indicated by EST data, are more conserved than testis-expressed sequences (Fig. 6c). Moreover, genes with ESTs from few (<4) tissue types are significantly ($P < 0.001$) less

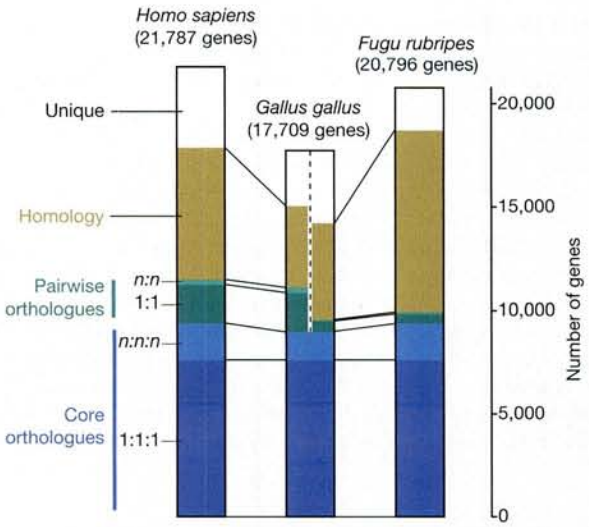


Figure 5 Chicken genes classified according to their predicted evolutionary relationships with genes of two other model vertebrates (*Fugu* and human). Forty-three per cent of the chicken genes are present in 1:1:1 orthology relationships for the three species. Also present in three species are *n:n:n* (many:many:many) orthologues; putative gene duplication events have resulted in multiple genes in at least one of the species. Pairwise orthologues are assigned when orthology is not detectable in the third species. Between *Fugu* and chicken, pairwise orthologues are rare (as expected), and might be indicative of gene loss in the lineage leading to humans. For a substantial number of genes, clear orthology relations cannot be described at all, but some similarity to genes in the other species remains detectable ('Homology', *E*-value cutoff is 10^{-6} in Smith–Waterman searches at the protein level). See Methods for details of orthology assignment.

conserved than sequences with ESTs from many (>6) tissues (median identities of 68.3% and 76.0%, respectively). The latter finding mirrors previous results for mammalian genes[64–66].

Genes that are conserved between human and chicken are often also conserved in fish: 72% (7,606) of chicken–human 1:1 orthologues also possess a single orthologue in the *T. rubripes* genome[67]. These genes represent a conserved core that is likely to be present in most vertebrates (Fig. 5). Their sequences are more conserved than other chicken–human 1:1 orthologues (Fig. 6a), indicating that they have been subjected to a greater degree of purifying selection. Chicken, human and pufferfish genes also encode a similar set of protein domains. Of 1,085 Pfam and SMART domain families in InterPro[68–70] that are encoded in at least three chicken genes, only two families are unrepresented among human genes, and a further 21 families are absent from *Fugu* genes.

Using a manual approach combining synteny information and highly sensitive search parameters, we were able to identify chicken orthologues of several immune-related genes previously thought to be specific to mammals. Owing to high sequence divergence (Fig. 6b), these were not predicted during the automated gene build process. These include the antimicrobial protein cathelicidin, colony-stimulating factor, leukaemia inhibitory factor, interleukin-3 (IL-3), -4, -7, -9, -13 and -26, and three secretoglobins. *IL-26* was previously known only in humans (it is a pseudogene in mouse and rat[71]), thus chicken represents the only available model organism with which to investigate *IL-26* function.

Gene and family differences

We next sought to identify those protein and domain families that are over- or under-represented in chicken, compared with mammals (Table 6 and Fig. 7). However, this type of analysis can be complicated by artefacts inevitably present in a draft genome project such as those introduced by gaps in sequencing coverage, assembly, or gene prediction. In chicken, although many are present in partial, fragmented form within the genome assembly, we estimate that roughly 5–10% of genes are substantially truncated

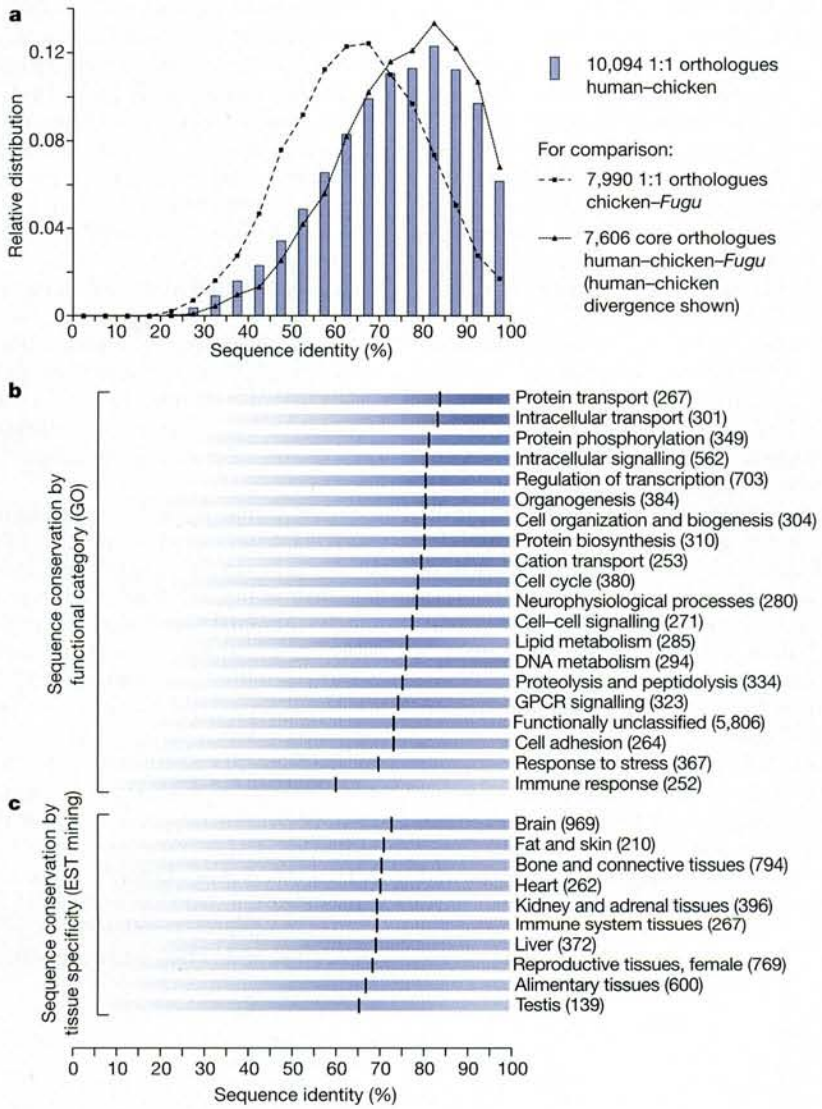


Figure 6 Sequence identity of orthologues. **a**, The percentage amino acid identity distribution of 1:1 orthologues between human and chicken, between chicken and *Fugu*, and between human–chicken orthologues that are also present in 1:1 relationships with *Fugu*. **b**, The percentage amino acid identity distribution of human–chicken 1:1 orthologues broken down by functional categories according to the GO subtree 'biological process'. Genes assigned to multiple categories were counted more than once. Vertical bars indicate the medians of the distributions. **c**, The percentage amino acid identity distribution of chicken–human 1:1 orthologues broken down by tissue category. Vertical bars indicate the medians of the distributions. Female reproductive tissues include uterovaginal, ovary and oviduct; immune system tissues include spleen, thymus, caecal tonsil and bursa of Fabricius; and alimentary tissues include gizzard, stomach, and large and small intestines.

in or missing from the Ensembl gene set (see Methods). Nonetheless, selecting our approaches with this in mind there are many insights to be gained, particularly in those areas not affected by these issues such as gene losses in human and gene innovations and expansions in chicken. Here, we document gene innovations, losses and expansions in chickens and mammals, and the evolution of function in orthologues.

Gene innovations in chicken. Expansion of and innovation within gene families in different lineages are often correlated with divergent phenotypes. For example, scales, claws and feathers of birds are formed using an avian-specific family of keratins, whereas hair fibre formation in mammals involves a distinct keratin family, which has greatly expanded within this lineage (Fig. 7). Of the ~150 avian keratins identified in the chicken genome, 30 are found in tandem arrays on GGA27. We also considered proteins that are specific to the eggshell, such as ovocleidin 116, for which homologues were not previously known outside of birds. However, with further analysis, we predict that ovocleidin 116 has mammalian orthologues, namely primate matrix extracellular phosphoglycoprotein (Blastp, $E = 2 \times 10^{-4}$), indicating that, despite low sequence identity (35% over 80 amino acids), both avian and mammalian genes perform similar roles in calcification. Gene innovation may also occur through domain accretion[1]. These events are extremely rare among mammalian genes[72], and we succeeded in detecting only a single instance within birds for a gene (ENSGALG00000000805) that encodes both scavenger receptor cysteine-rich (SRCR) and fibrinogen-related domains. A single exon, encoding the SRCR domain, appears to have been inserted into this gene (Supplementary Fig. S4) relatively recently, because comparison to its paralogue (ENSGALG00000000732) shows considerably fewer synonymous substitutions (K_S value of 0.034) than do the majority of chicken–human orthologue pairs (median K_S value of 1.66).

Genes absent from chicken. Genes encoding vomeronasal receptors, casein milk proteins, salivary-associated proteins (statherin and histatins) and enamel proteins seem to be absent from the chicken: from within both the EST sets and the genome. This is unlikely to result from imperfections in the chicken genome assembly because it preserves orthologues of closely linked (syntenic) mammalian genes. These absences might therefore mirror the evolution of the vomeronasal organ and mammary glands in mammals, and the loss of teeth in birds. The presence in fish of enamel-associated genes and their absence in chicken, together with absence of chicken casein and salivary-associated genes, which all cluster together in mammalian genomes, is consistent with a previous suggestion[73] that these have all descended, by gene duplication and rapid sequence diversification, from a common ancestor, possibly an enamel-associated gene.

Loss of an entire domain family in the protein repertoire is unusual. Nonetheless, the SCAN domain family seems to be unrepresented in the chicken gene set and genome. SCAN domains are dimerization motifs that are found, often with zinc finger domains, in more than 60 human proteins[74,75]. In *Fugu* and zebrafish, a total of three SCAN domains are found, not associated with zinc fingers but instead within large retrovirus-like proteins. This phyletic distribution indicates a possible acquisition of new function in the synapsid or mammalian lineages, whereas in the avian lineage the domain might have remained associated with retroviral sequences that died out subsequently.

Overall, we find that among the predicted chicken genes there is a notable under-representation of genes that are widely conserved and were presumably present in the mammal–bird common ancestor. We estimated gene innovation and loss by considering patterns of the presence or absence of orthologous genes for nine metazoan genomes, and by reconstructing the most parsimonious gene contents of ancestral root species, using the plant *Arabidopsis thaliana* as an outgroup (Fig. 8; see also Methods). Despite uncertainties stemming from the incompleteness of current genome sequences, all metazoan lineages seem to be gaining orthologous core genes, with the notable exception of the ancestor of the Diptera[76,77]. Among the vertebrates, the chicken seems to be the only species thus far sequenced to have lost more of those genes than it has gained over an extended period of time. At present, it is difficult to estimate to what extent this observation is influenced by gaps in the assembly. Indeed, 57% of the apparent losses that are otherwise conserved as single copy genes in mammals and insects are represented in chicken EST libraries (reciprocal best match, bitscore ≥ 200). However, this coverage is lower than the average for this gene class (88%), suggesting that not all of the absences can be attributed to assembly gaps. There are a number of potential explanations for these data. The most straightforward is that some of these core genes were deleted within the avian lineage. An alternative is that they have experienced accelerated evolution. Targeted investigations will be required in order to resolve whether birds have lost unusually large numbers of genes that are otherwise conserved among metazoa.

Expansion of domains and gene families in chicken. Three domain families are significantly over-represented in chicken with respect to human (Fig. 7). The most notable case is an apparent 40% expansion of SRCR domains. Many of these domain sequences are similar to those of human *DMBT1*, a proposed regulator of mucosal homeostasis[78]. This suggests a link between the abundance of SRCR domains in chicken and its adaptive requirements in mucosal defence.

To gain a more in-depth view of family expansions in chicken we identified, from the set of automatically derived orthologues, nine chicken gene families containing at least fourfold more representatives than in human (Fig. 7). Most of these families seem to have roles in immunity and host defence. Family sizes were further refined by re-predicting genes and pseudogenes from chicken and human assemblies using Exonerate (G. Slater and E. Birney, unpublished software) and Genewise[79] (Table 7). The most marked expansion involves at least 218 non-identical chicken genes that are predicted to be orthologous to one of two olfactory receptor genes (*OR5U1* and *OR5BF1*). In humans, *OR5U1* and *OR5BF1* genes lie within olfactory receptor gene clusters that each are positioned next to paralogous major histocompatibility complex (MHC) class I gene clusters[80]. These olfactory receptors have been tentatively proposed to be involved in odorant-mediated detection of MHC diversity[81]. Duplication of and/or gene conversion within *OR5U1*-like and *OR5BF1*-like genes in chicken appear to have been

Table 6 **Changes to the genome content of chicken and human**

Species	Probable innovation	Change in number	Loss
Changes to domains			
Birds	Avian-specific feather keratin domain	SRCR domain: over-represented in chicken	SCAN dimerization domain
Mammals	Mammalian hair keratin domains	Intermediate filaments: over-represented in mammals	Avidin domain (egg-white)
Changes to genes			
Birds	Gene with novel domain combination: fibrinogen-related and SRCR domains	Olfactory receptor type 5U1/5BF1: expansion in chicken	Enamelin and amelogenin
Mammals	Caseins (milk proteins)	Mammalian taste receptors: expansion in mammals	DNA photolyase (nocturnal lifestyle)

Interpretation of presence and absence patterns as well as family size variation between chicken and human. For details on the method of detection see Methods.

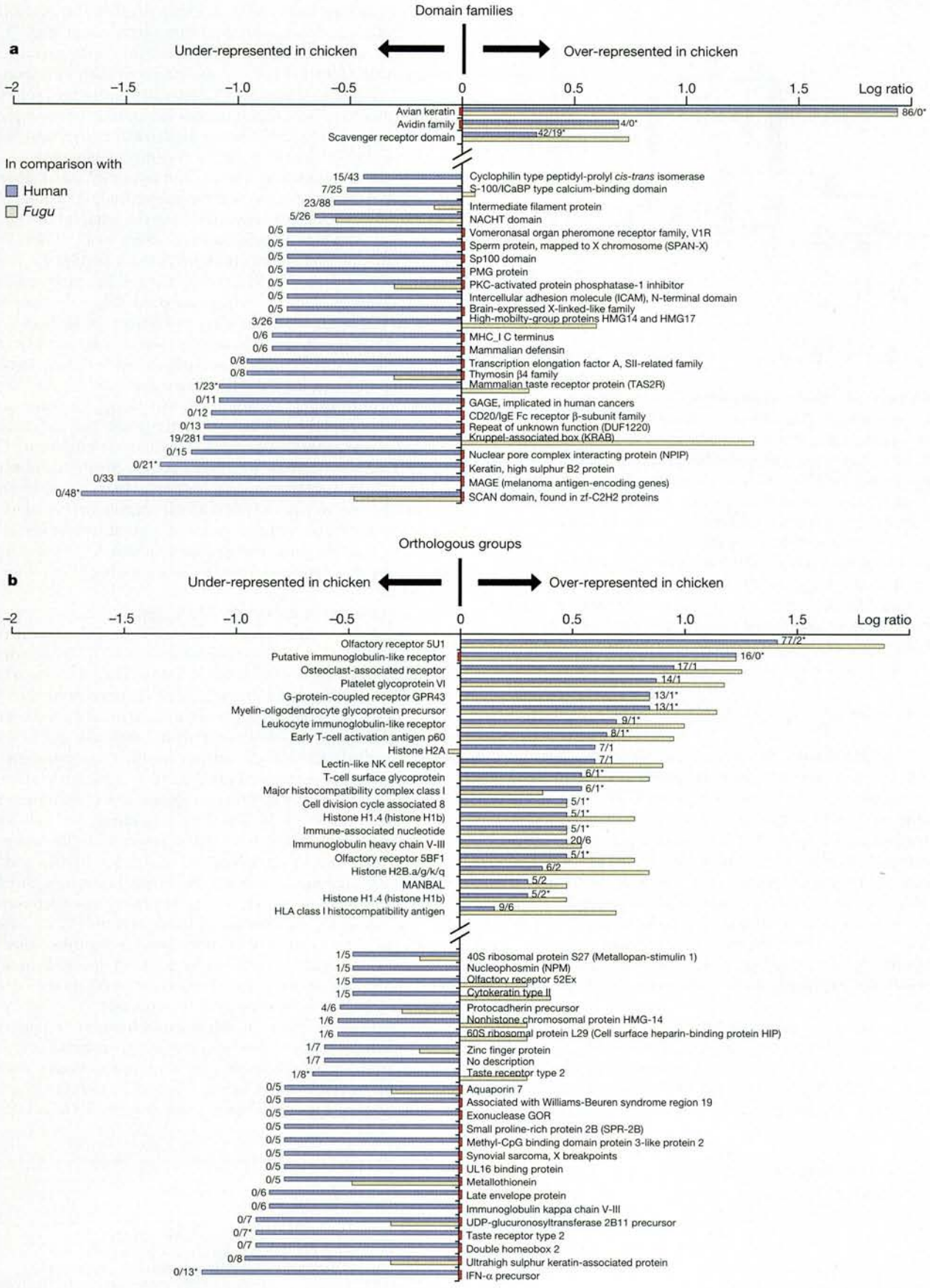


Figure 7 Loss, innovation, expansions and contractions of protein families: domain counts and orthologous relations. All at-least-twofold over- and under-representations (separated by a solidus) are shown for both members of domain families (**a**) and 'many to many' orthology relations (**b**). Ranking of families and groups has been done with respect to the human genome; *Fugu* data are also shown for comparison. An asterisk indicates that manual analyses refined what were otherwise automatic counts. Families not subjected to twofold variations are not shown.

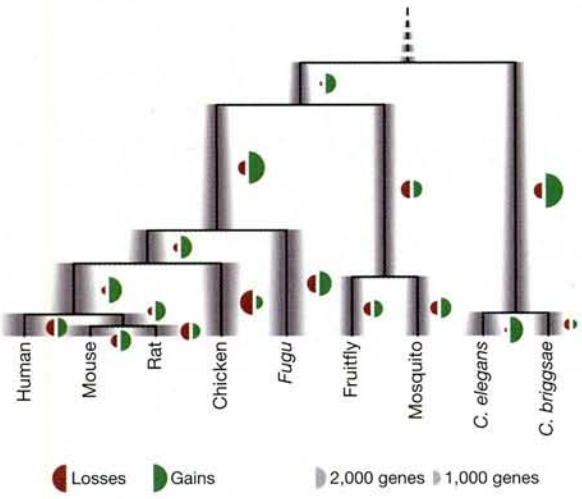


Figure 8 Approximate history of animal genomes. The gene content of ancestral animals—as estimated by using automatically delineated orthologous groups (see Methods)—assuming parsimony as well as a lack of horizontal gene transfer among animals. The inferred number of lost (and gained) genes on each lineage is shown as a half-circle, the area of which is proportional to the number. The shadings under the trees give a rough indication of the number of genes per genome. Wherever necessary, the *Arabidopsis* genome was used as an outgroup to infer the direction of changes. Ancestral estimates close to the root are likely to be underestimates because of unrecognized orthology relations and putative ancestral genes failing to survive in any of the extant genomes. The gene count of present-day genomes was approximated by considering only genes with orthology support; any remaining genes were considered only if they had substantial similarity support within the genome (to avoid spurious gene predictions, pseudogenes and/or fragments; see Methods).

both extensive and relatively recent in this lineage (Supplementary Fig. S5).

The *OR5U1/OR5BF1*-like genes contribute the majority of at least 283 olfactory receptor genes in the current chicken assembly, a similar number to that found for humans[82]. The large size of the olfactory receptor family seems to run counter to the textbook view (for example, see ref. 83) that birds have a poor sense of smell[84]. Individual chicken researchers, however, had already suspected that chickens are not particularly anosmic (for example, see ref. 84). On five chicken chromosomes, these olfactory receptor genes are present in subtelomeric clusters, which, by analogy with human olfactory receptor clusters[85], may be associated with their rapid evolution[86]. As with these genes in mammals, chicken olfactory receptor genes are interspersed with many (~100) homologous pseudogenes.

Gene innovations in mammals. The largest mammal-specific gene family is the high/ultrahigh sulphur hair keratins (also named keratin-associated proteins). Thirty-seven copies have been reported on HSA17q12-21 (ref. 87) embedded in the larger type I keratin cluster, which is present also in other mammals[88]. Another large orthologous family with multiple members in human, but none in chicken, is a subgroup of the α-interferons (Fig. 7). The expansion and diversification of interferons into various subfamilies is thought to be a mammalian innovation in response to different pathogen challenges[89,90]. Some highly diverged interferon homologues have been found outside mammals (for example, see ref. 89), and the distant relatives in chicken and *Fugu* appear to have duplicated independently from the mammalian radiation.

Gene losses in mammals. The chicken genome enables increased precision in dating gene losses in the lineage leading to humans: we can identify losses that happened within the synapsid lineage, but before the mammalian radiation (an interval of ~110–310 Myr ago). Only one loss affects an entire domain family, the avidins, which are represented in the chicken but not in sequenced mammalian genomes (Fig. 7). Avidins are present in oviparous vertebrates as minor egg-white proteins, and closely similar homologues in zebrafish, sea urchin and bacterial genomes indicate a loss of this domain family in mammals. Mammals also seem to have lost genes encoding vitellogenin I and II, which are yolk storage proteins providing nutrients to the early embryo in the egg. These losses were probably concordant with reduction in egg size and with internalization of the embryo during mammalian development. Other mammalian losses appear to reflect a presumed episode of nocturnal lifestyle in early mammalian history[91]. Apart from the known loss of CPD-photolyase, which is a DNA-repair enzyme dependent on light energy[92], we find that mammals appear to have lost at least one pigmentation gene, indigoidine synthase A, whose bacterial homologue is an enzyme involved in generating blue colour pigments[93].

Gene expansions in mammals. The largest mammalian expansion observed at the domain level involves the Kruppel-associated box (KRAB) domain, whose presence coincides with C2H2 zinc fingers in transcription factors. Whereas the KRAB domain has been found in more than 400 human genes[94], the chicken seems to have fewer than 140. The complete absence of this domain in *Fugu* underlines the marked expansions of this family in mammals and to a lesser degree in birds. Orthology analysis identifies an apparent gene expansion of G-protein-coupled taste receptors in mammals (Fig. 7a, b). These fast-evolving receptors were previously studied only in mammals[95,96]. In the chicken genome, we find similar numbers of type I receptors (those responsible for sweet and umami taste; that is, certain amino acids such as Glu and Asp, and related compounds) but only three type II receptors, compared with around 30 in mammals. Type II taste receptors are thought to be responsible for the sensing of bitter tastants[97,98], an adaptive avoidance system that may be more variable than the other taste systems. This might indicate that birds have a limited capacity for bitter taste, or that they have recruited other G-protein-coupled receptor subtypes for sensing bitter compounds.

Gene duplication events in chicken and human. To understand gene repertoire differences between avian and mammalian lineages, we quantified the number of gene duplication events since the divergence of chicken and human lineages by examining species-specific duplicate pairs (paralogues that reciprocally have their best

Table 7 **Details of family expansions in chicken**

Number of chicken genes	Number of chicken pseudogenes	Chicken:human gene ratio	Annotation
202	16	202:1	Olfactory receptors 5U1/5BF1
25	6	25:1	Immunoglobulin-like receptor CHIRs
14	3	14:1	Major histocompatibility complex, class I-like sequence
13	0	13:1	Myelin-oligodendrocyte glycoprotein; immunoglobulin V-region-like B-G antigen
26	1	26:3	G-protein-coupled receptor 43; free fatty acid activated receptor 2 in leukocytes
8	0	8:1	Early activation antigen CD69
6	0	6:1	T-cell surface glycoprotein CD8 α-chain
5	0	5:1	Immune-associated nucleotide 6-like; guanosine triphosphatase

In-depth analysis of gene families that have been expanded at least fourfold, as revealed by automatic orthology analysis.

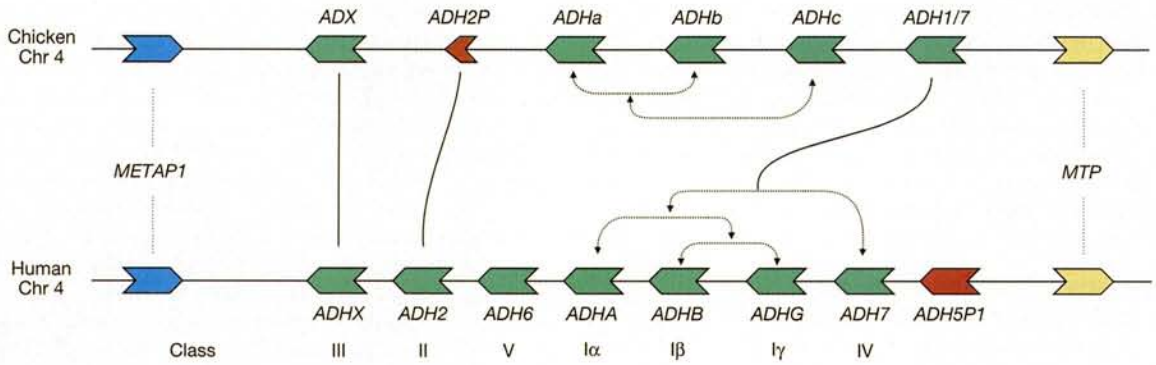


Figure 9 Evolution within the orthologous alcohol dehydrogenase (ADH) gene clusters of similar size in chicken and human. Architecture of the ADH clusters in human (HSA4q23, between positions 100.3 and 101 Mb) and chicken (GGA4, between 60.4 and 60.5 Mb). Both ADH regions are flanked by orthologous genes: *METAP1* (upstream, blue) and *MTP* (downstream, yellow). Gene names and classes used here for the eight human ADH genes are in accordance with a recent nomenclature proposed for this protein family[99]. Using the human ADHX protein as homologous template, we have identified up to six ADH gene copies in the orthologous chicken region: five are complete and one is fragmented and likely to be an *ADH2* pseudogene representing a niche loss. Only chicken proteins encoded by genes *ADHa* and *ADH1/7* had been previously identified[162,163]. Known genes or complete predictions with no truncations are displayed in green, whereas incomplete or truncated copies (pseudogenes) are in red. The probable correspondences between the chicken and human genes (shown as lines connecting both clusters) are revealed by the neighbourhood-joining tree of the proteins shown in Supplementary Fig. S6.

hit in the same genome). Paralogues exceeding 95% pairwise identity were excluded from both data sets to avoid chicken genome assembly artefacts. In chicken, 13% (221 out of 1,712) of duplicate pairs have their closest hit in the chicken genome. In contrast, 27% (647 out of 2,426) of paralogues with <95% identity from the human genome are human specific, a pattern that is significantly different compared with that seen in the chicken (Fisher's exact test, $P < 10^{-15}$). Thus, despite similar numbers of older duplications in both genomes (1,779 in humans versus 1,491 in chicken), there have been significantly fewer paralogues specific to the chicken lineage.

Evolution of function in orthologous gene clusters

Five alcohol dehydrogenase (ADH) genes plus one pseudogene fragment are present on GGA4, whereas seven genes (plus one pseudogene) are found in the orthologous segment on HSA4p23 (ref. 99) (Fig. 9; see also Supplementary Fig. S6). The *ADH2* gene appears only recently to have been silenced in chicken because it is still functional in the ostrich[100]. Two recent duplication events have given rise to three chicken ADH genes that might compensate for some of the lost *ADH2* functionality. These duplications, and others in human and other vertebrate lineages[99] (Supplementary Fig. S6), might indicate changes in expression patterns or even reveal episodes of adaptation to nutritional and detoxification requirements involving changes to limited gene repertoires.

Change of function in orthologues

For the past 40 years, avian species have been described as lacking a functional urea cycle because of the absence in the chicken liver of the enzymatic activity that initiates the cycle: carbamoyl phosphate synthetase 1 (*CPS1*)[101]. This observation has been linked to a difference in excretion of nitrogenous waste: whereas mammals excrete urea, birds excrete uric acid as a possible stratagem in reducing the build-up of soluble urea during development in the egg. Contrary to these expectations we have identified a full-length, apparently functional, chicken *CPS1* gene on GGA7 in the orthologous location to the human *CPS1* gene (Supplementary Fig. S7a). Characterization of the expression profile of this candidate gene revealed expression in brain and muscle, and in immune tissues (Supplementary Fig. S7b). These data, and the presence of all the other enzymes of the mammalian urea cycle in the chicken genome (data not shown), indicate that this cycle is intact in birds but might perform a function distinct from the generation of urea for excretion.

Exploring chicken genome architecture

The assembly of the chicken genome, with its distinctive karyotype and smaller size, provides an opportunity to explore important issues in genome structure and evolution. Cytogenetic, molecular and mapping data have suggested previously that microchromo-

Table 8 **Comparison of properties of mammalian (A+T)/(G+C)-rich regions with chicken macro- and microchromosomes**

Characteristic	(A+T)-rich	(G+C)-rich	Macrochromosomes	Intermediate/microchromosomes
Ultrastructure				
Cytogenetic band type	G-band	R-band	G-band	R-band
DNA sequences				
Gene density	Low	High	Low	High
Intron size	Longer than (G+C)-rich	Shorter than (A+T)-rich	Longer than intermediate/microchromosomes	Shorter than macrochromosomes
G+C content	Lower than (G+C)-rich	Higher than (A+T)-rich	Lower than intermediate/microchromosomes	Higher than macrochromosomes
CpG island density	Less than (G+C)-rich	More than (A+T)-rich	Less than intermediate/microchromosomes	More than macrochromosomes
Long interspersed repeats	More than (G+C)-rich	Less than (A+T)-rich	More than intermediate/microchromosomes	Less than macrochromosomes
Short interspersed repeats	Less than (G+C)-rich	More than (A+T)-rich	Almost none	Almost none
Mutation				
Synonymous rate (K_S)	–	–	Lower than intermediate/microchromosomes	Higher than macrochromosomes
Non-synonymous rate (K_A)	–	–	Same	Same
K_A/K_S ratio	–	–	Higher than intermediate/microchromosomes	Lower than macrochromosomes
GC3	–	–	Lower than intermediate/microchromosomes	Higher than macrochromosomes
Function				
DNA replication	Later than (G+C)-rich	Earlier than (A+T)-rich	–	Mostly early
Meiotic recombination rate	Lower than (G+C)-rich	Higher than (A+T)-rich	Lower than intermediate/microchromosomes	Higher than macrochromosomes
Methyl-C content	Lower than (G+C)-rich	Higher than (A+T)-rich	Lower than intermediate/microchromosomes	Higher than macrochromosomes

somes are (G+C)-rich, CpG-rich and gene-rich[102–104], exhibit features that are correlated with transcriptionally active DNA[105–110], and have structural counterparts in the (G+C)-rich mammalian chromosomal R-bands (Table 8). Our current analysis shows that the different size classes of chicken chromosomes exhibit a number of correlated attributes. Although these trends are well documented in mammals (for example, see refs 1, 2, 111–113), the sequence of the chicken genome assembly demonstrates levels of complexity, breadth and resolution that previously could not be achieved.

Comparison of physical and genetic distances

A comparison of the physical distance along each chicken chromosome derived from the sequence assembly with the genetic distance between markers (based on the sex-averaged map[25,29]; see Methods) reveals wide variation in recombination rates that has a strong negative association with chromosome length (Fig. 10a). The recombination rate varies over an eightfold range among chromosomes (2.5 to 21 cM Mb^{-1}); rates are much higher on microchromosomes (median value of 6.4 cM Mb^{-1}) than on macrochromosomes (median value of 2.8 cM Mb^{-1}). These results contrast with the narrow range and overall lower rates found in mammalian chromosomes of only 1–2 cM Mb^{-1} for human and 0.5–1.0 cM Mb^{-1} for the mouse[1,2]. The increased recombination rate of microchromosomes is such that all are likely to have total genetic lengths between 50 and 100 cM, as was expected, given the obligatory chiasma per bivalent to ensure normal segregation during meiosis[114].

Chromosomal distributions of sequence features

Our analysis of the distributions of G+C content, CpG islands, and gene density and size was restricted to chromosomes with nearly complete sequence coverage; this excluded chromosomes GGA16, GGA22, sex chromosomes and microchromosomes smaller than GGA28 (Supplementary Table S2 and Fig. S8). We arbitrarily designated three chromosome size groups: large macrochromosomes (GGA1–5), intermediate chromosomes (GGA6–10) and 28 microchromosomes (GGA11–38). The overall G+C content declines markedly with increasing chicken chromosome length (Fig. 10b). The G+C content in macrochromosomes shows a narrow distribution, whereas for microchromosomes the distribution is broad and shifted to higher values (Supplementary Fig. S9a). Indeed, some individual chromosomes have almost no overlap in their distributions of G+C content (Supplementary Fig. S9b). These patterns differ from the considerable overlap observed in G+C content among individual mammalian chromosomes (Supplementary Fig. S9c; see also ref. 3).

Thus both G+C content and recombination increase with decreasing chromosome size. The association between G+C content and recombination may be viewed as a dynamic cycle. A possible explanation is that chromosomal regions with a high recombination rate might elevate the G+C content by "biased gene conversion"[115]. In biased gene conversion, mismatch repair within the heteroduplexes formed during recombination results in G+C-biased gene inclusion[116]. Biased gene conversion may also increase the neutral mutation rate of microchromosomes. Furthermore, studies on a wide range of species (summarized in ref. 111) have shown that (G+C)-rich DNA has an increased recombination rate. Thus, a cycle of high recombination may lead to increased G+C content, which then leads to a further increase in recombination rate until equilibrium is reached when selective forces prevent further change. This model suggests that changes in recombination brought about by a chromosome rearrangement may lead to changes in G+C content, as recently shown in mouse[117].

In the chicken genome 48% of CpG islands (see Methods and Supplementary Table S5), often associated with promoters and other sites of regulation[118,119], overlap a gene. We found that ~38% of chicken CpG islands are conserved in the human genome. However, taking into account proximity to protein-coding genes and overlaps with ESTs (see Methods and Supplementary Table S5), we find that 10% of CpG islands conserved with the human genome are not near a gene. These may represent distant regulatory regions, flank undetected genes, or serve another function.

CpG island density presents a strong negative association with chromosome length, being highest on the intermediate chromosomes and microchromosomes (Fig. 10c). Similarly, gene density shows a strong negative association with chromosome length (Fig. 10c), confirming, at higher resolution, earlier studies[102,104]. Furthermore, we find a strong correlation between the length of a gene and the size of the chromosome in which it is found, an effect that is determined largely by variation in intron size (Fig. 10d). Although exon lengths (Fig. 10d) and numbers (data not shown) do not vary significantly among chromosome types, intergenic distances do increase with chromosome size (Fig. 10e). Intron length in the chicken correlates negatively with recombination ($r_s = -0.774$, $P < 0.001$), G+C content ($r_s = -0.934$, $P < 0.001$) and gene density ($r_s = -0.961$, $P < 0.001$), as has been reported previously for other genomes[120–123]. The proportion of a chromosome covered by interspersed repeats increases with chromosome length (Fig. 10f; $r_s = 0.973$, $P < 0.001$, for the 14 largest chromosomes excluding GGA8; $r_s = 0.940$, $P < 0.001$, for all chromosomes when excluding the outliers GGA8, GGA24, GGA27 and GGA28). Not surprisingly, repeat density correlates positively with intron and intergenic gap length, as repeats primarily exist in these locations. Repeat density correlates negatively with recombination rate, G+C content and gene density (Fig. 10). The correlation with recombination rate is also expressed in the twofold higher than expected repeat density on sex chromosome GGAZ (Table 5) and an increased density with distance from the

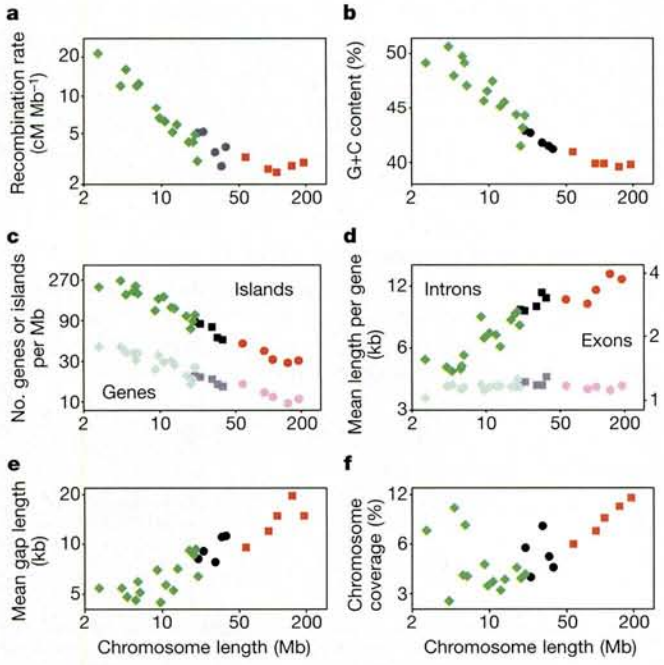


Figure 10 Relationships between chromosome sequence length and characteristics for chromosomes 1–28. **a–f**, Recombination rate (**a**), G+C content (**b**), densities of genes and CpG islands (**c**), total lengths of introns and coding exons per gene (**d**), intergenic gap lengths (**e**) and densities of interspersed repeat elements (**f**). All plots exclude GGA16 and GGA22, which have insufficient sequence, and panel **a** also excludes GGA23 and GGA25, which have insufficient genetic markers. Red, macrochromosomes; black, intermediate chromosomes; green, microchromosomes; additional paler colours indicate genes in **c** and exons in **d**.

Table 9 **The short arm of chicken chromosome 4 retains its ancestral microchromosome properties**

Characteristic	GGA4p	GGA4q
Physical length (bp)	18,442,167	70,692,738
Recombination rate (cM Mb^{-1})	4.44	2.51
G+C content (%)	42.41 ± 5.60	38.68 ± 3.36
Intron length (bp)	16,005 ± 26,850	28,503 ± 40,711
G+C content of intron (%)	46.00 ± 8.92	40.00 ± 6.22
Exon length (bp)	1,425 ± 1,089	1,567 ± 1,308
G+C content of exons (%)	51.28 ± 6.83	47.00 ± 6.26
Gene density (genes Mb^{-1})	19.5	10.6
CpG density (islands Mb^{-1})	80.96	30.64
Repeats (%)	4.84	9.13
DAPI stain	Dull, (G+C)-rich	Bright, (A+T)-rich

In some cases values are shown ±1 s.d.

centromere (Supplementary Fig. S10). Recombination rate will influence repeat density if transposable element insertions on average are slightly deleterious and selection is more effective in regions of high recombination[124]. Indeed, recently a strong negative correlation between recombination rate and repeat density was reported in primates and rodents, where recombination rate was found to be proportional to distance from the centromere[111]. The correlation may be more obvious in chicken because of the larger differences in chromosome size and recombination rates. The same mechanism can be proposed to explain the correlation between the size of introns and intergenic gaps and recombination rate, assuming that longer introns and intergenic gaps are slightly deleterious.

The distinctive characteristics of microchromosomes are retained even after fusion with a macrochromosome, as illustrated by an ancestral chromosome fusion leading to GGA4. The p arm of GGA4 is orthologous to an ancestral microchromosome, as shown by comparative mapping using chromosome painting and by sequence matches with microchromosome 9 in turkey (see Methods). Interestingly, the telomeric DNA signals from fluorescence *in situ* hybridization analyses show that GGA4p possesses an interstitial telomere adjacent to the centromere (Supplementary Fig. S11). Analysis of the sequence assembly shows that GGA4p has not taken on characteristics of the macrochromosome, but rather it still has the properties of its ancestral microchromosome, including high recombination rate and high gene density (Table 9).

Synonymous substitution rates

Neutral substitution rate variation was analysed by examining synonymous substitution rates (K_S), estimated using Codeml[125], in a set of 7,529 human–mouse–chicken 1:1:1 orthologues (Fig. 11). As expected from the greater phylogenetic distance, the level of nucleotide substitution deduced from human–chicken alignments is much higher than that observed in human–mouse comparisons. The median value of K_S for 1:1 orthologues between human and chicken is 1.66, indicating that synonymous sites on average have changed one to two times in the combined lineages to humans and chickens. For the analyses presented below, we acknowledge that with such high substitution rates, the conclusions should be corroborated with data from comparisons of much more closely related taxa, in particular from two or more bird lineages.

Synonymous substitution rates are elevated in two types of

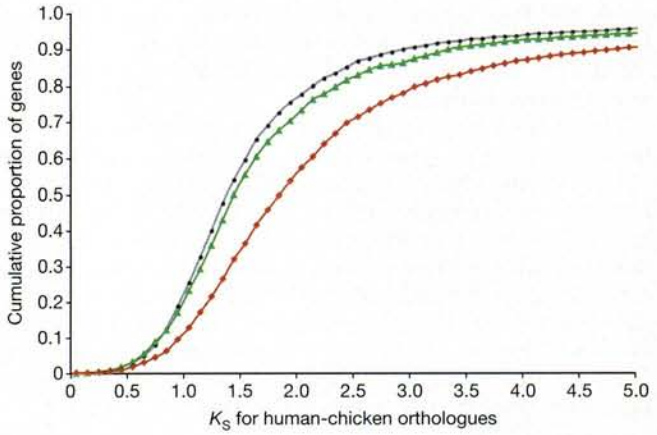


Figure 11 Genes on chicken microchromosomes possess higher synonymous substitution rates. We split the chromosomes into macrochromosomes (GGA1–5, grey circles), intermediate chromosomes (GGA6–10, green triangles) and microchromosomes (GGA11–32, red circles), and calculated their K_S values using Codeml. The cumulative distribution of genes in each chromosome size category is plotted versus the K_S values. Because large K_S values are methodologically suspect, we discarded K_S values >10.

chromosomal region. First, chicken–human K_S values are significantly higher among genes on microchromosomes (median $K_S = 2.01$) than they are on intermediate chromosomes (1.60) and on macrochromosomes (1.54) (Table 10; $P < 2.2 \times 10^{-16}$, Kolmogorov–Smirnov test). Despite significant differences in G+C content for the chicken chromosomes (Fig. 10b), the elevation in K_S is not entirely due to non-equilibrium base compositions of chicken or human sequences in the time since their ancestral genes diverged. Specifically, we found that K_S distribution differences between genes on chicken micro- and macrochromosomes also remain significant when only considering genes exhibiting similar G+C contents, or G+C percentage at fourfold degenerate sites[126] (Table 9). Higher neutral divergence rates are also seen on microchromosomes when genes are compared between chicken and turkey genomes[127].

Second, subtelomeric sequences, which are often duplicated and polymorphic[86,128], also show elevated neutral substitution rates. The synonymous substitution rates of chicken genes in regions up to 10 Mb from the ends of assembled macrochromosomes are elevated to levels that are indistinguishable ($P > 0.05$ in a Kolmogorov–Smirnov test) from those of microchromosomal genes when chicken–human alignments are considered (Table 10). It thus seems possible that the elevated K_S values associated with chicken microchromosomal genes result from the more general phenomenon of rate elevation in subtelomeres: the K_S value for a microchromosomal gene is, on average, higher than a gene on a macrochromosome because all microchromosomal genes lie within 10 Mb of one or both of their telomeres. However, it is important to note (Fig. 12) that genes on chicken microchromosomes are not, in general, subtelomeric in mammalian genomes. Nevertheless, the mammalian orthologues of chicken microchromosomal genes show significantly ($P = 0.003$) elevated K_S values even when those values

Table 10 **Chicken gene synonymous substitution rates**

Species	K_S microchromosomes			K_S intermediate chromosomes			K_S macrochromosomes		
	25%	Median	75%	25%	Median	75%	25%	Median	75%
Chicken–human	1.473	2.009	2.895	1.224	1.599	2.252	1.193	1.535	2.087
Chicken–human (0.48 < [GC] < 0.52)	1.409	1.867	2.657	1.172	1.479	1.876	1.196	1.556	2.088
Chicken–human (0.48 < [GC4d] < 0.52)	1.389	1.801	2.698	1.330	1.640	1.811	1.179	1.566	2.056
Chicken–human 5 Mb subtelomeric regions	1.519	2.076	3.011	1.235	1.623	2.229	1.474	2.162	3.080
Human–mouse	0.461	0.588	0.768	0.421	0.528	0.651	0.444	0.548	0.687

GC4d, G+C fraction at fourfold degenerate sites.

are derived from mouse–human comparisons. Thus, this gene set seems to have retained some characteristic other than chromosome location that has elevated neutral divergence during mammalian, as well as avian, evolution.

The overall increase in the rate of synonymous substitution for microchromosomal genes was not accompanied by significant changes in the chicken–human non-synonymous substitution rate, K_A (median values of 0.121, 0.118 and 0.116 for macro-, intermediate and microchromosomes, respectively). Microchromosomal genes on average have smaller K_A/K_S ratios (median value of 0.052, compared with 0.073 and 0.068 for macrochromosomes and intermediate chromosomes, respectively), indicating that they have been subjected to greater degrees of purifying selection. This highlights a hitherto unforeseen contribution of genomic location to coding sequence evolutionary constraint.

Segmental duplications

One unusual aspect of the human genome with respect to other sequenced genomes is the abundance (4%) of large (>20 kb), nearly identical duplications[129]. In marked contrast to observations for mammalian genome assemblies[129–132], of the confirmed segmental duplications in chicken (see above), few exceeded 10 kb in length (Fig. 13), and none greater than 50 kb were detected. Almost all mapped duplications (93%) within the chicken genome are intrachromosomal (excluding unplaced sequences) (Supplementary Fig. S12). Analysis of exon content of the duplicated regions revealed that only 3.7% of Ensembl predicted genes showed evidence of being recently duplicated—a slight enrichment as compared to the genome average of confirmed duplicated bases (3.0%). The proportion of uncharacterized genes within these recent segmental duplications is nearly fivefold greater than that observed for

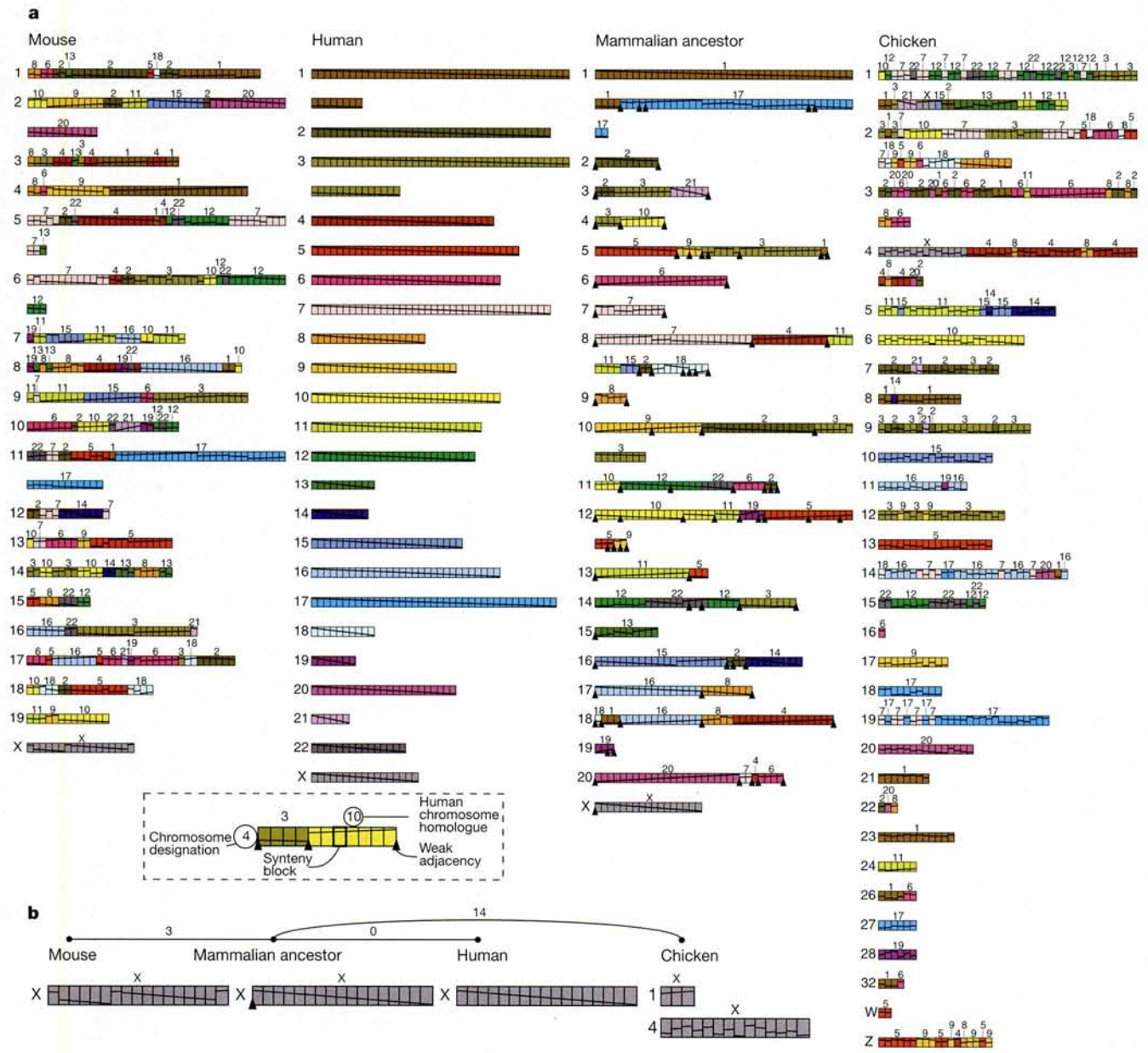


Figure 12 Putative mammalian ancestor recovered by GRIMM and MGR using the human, mouse, rat (not shown) and chicken genomes. **a**, Each genome is represented as an arrangement of 586 synteny blocks each drawn as one unit, regardless of its length in nucleotides. Each human chromosome is assigned a unique colour, and a diagonal line is drawn through the whole chromosome. In other genomes, this diagonal line indicates the relative order and orientation of the rearranged blocks. **b**, The recovered ancestral X chromosome is optimal and unique. Gene order of the ancestral X chromosome is identical to human. Numbers associated with the lines indicate the minimum number of rearrangements required to convert between two nodes.

other unique sequences. Characterization of these may uncover lineage-specific genes important for the adaptation of the chicken.

Evolution of vertebrate genomes

Amniote chromosomal rearrangements

The evolutionary history of genome organization can be inferred from the comparative analysis of gene orders on chromosomes. Recent studies of mammalian genomes suggest a larger number of rearrangements than previously estimated[1,2,133–136], largely due to the underestimation of intrachromosomal rearrangements. Using the chicken genome sequence as an outgroup provides an opportunity to understand better the rate of rearrangements in mammalian lineages and the architecture of the ancestral mammalian genome.

We generated two maps of conserved synteny—defined as orthologous chromosomal segments with a conserved gene neighbourhood[133,134,137]—among chicken, human and mouse, using DNA- and protein-level methods (see Methods). These two sets of anchors generated highly similar maps, with comparable sizes of orthologous segments and amounts of coverage of the human and mouse genomes (Table 11). We then used two approaches to quantify chromosomal dynamics: (1) by counting the number of synteny breaks where the ancestral state is confirmed by conserved synteny to an outgroup species; and (2) by estimating the minimum number of rearrangements that could explain the current genome organization. Both approaches are consistent, revealing a slow rate of rearrangement in the human lineage, about one-third that of the rodent lineage (Fig. 14). The synteny analysis using orthologous gene pairs is more sensitive at long evolutionary distances than DNA-based methods, and it allowed us to use recently sequenced fish genomes of *Fugu*[67] and *Tetraodon*[138] as outgroups. This analysis revealed an even slower rate of rearrangements in the chicken lineage (Fig. 14). The synteny maps confirm an earlier observation[139] that the human genome is closer to the chicken than to rodents in terms of chromosomal organization of genes (Supplementary Figs S13 and S14). As indicated from the analysis of the *Tetraodon* genome, this surprising level of similarity arises from an unusually low rate of interchromosomal shuffling in the lineage leading to the earliest mammal.

We reconstructed the putative mammalian ancestor genome (more precisely, the common ancestor of human–mouse–rat and all the Euarchontoglires[140]) by looking for a scenario minimizing the number of rearrangement events on the evolutionary tree[141] (see Methods). Because the pairwise distances between the initial genome are substantial, it is possible to find alternative ancestors that also minimize the total number of rearrangement events. By exploring some of these alternative ancestors, we partitioned all the pairs of adjacent synteny blocks of the recovered ancestor into 'strong' and 'weak' pairs depending on whether they are present or not in all of the observed alternative ancestors (Fig. 12a). Many of the previously postulated chromosome associations of the placental ancestor correspond to strong pairs of adjacent blocks in the mammalian ancestor[136,142]. The synteny of six human chromosomes and of ten chicken chromosomes is preserved in the mammalian ancestor, whereas synteny only of the rodent X chromosome is preserved in the mammalian ancestor[141]. This is consistent with the suggestion that interchromosomal rearrangements have been more frequent in rodents[139]. The scenario recovered for the X chromosome also reveals variable rates of intrachromosomal rearrangements (Fig. 12b). There appear to have been no rearrangements between HSAX and the X chromosome of the mammalian ancestor (that is, human order is ancestral), yet there are three inversions between the mouse X chromosome (MMUX) and the mammalian ancestor X chromosome, and there are 14 rearrangements (13 inversions and 1 fusion) between the two chromosome segments in chicken (from GGA1 and GGA4) and the mammalian ancestor X chromosome.

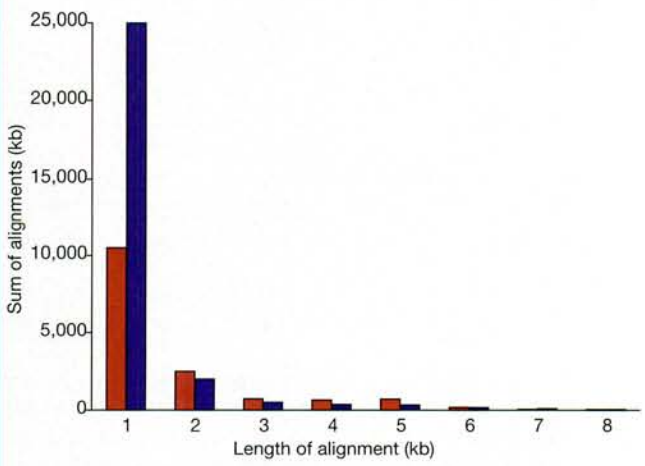


Figure 13 Segmental duplications within the chicken genome are small. The length distribution of intrachromosomal (blue) and interchromosomal (red) alignments that were confirmed by over-representation in the whole-genome shotgun sequence is shown for the chicken genome. The sum of alignments for alignments of lengths 9–15 kb are each less than 40 kb (data not shown).

Variation in human–chicken genome expansion ratio

The large variation in genome size between organisms[143] has been addressed previously by comparing entire genomes, thereby deriving an aggregate estimate for expansion or contraction in one lineage. The change in overall genome size between chicken and mammals coupled with the conserved synteny allows us to isolate different factors that contribute to this variation (see Methods). The chicken genome is about 40% the length of the human genome, but this ratio is not constant across all orthologous regions (Fig. 15a). The variation in length ratios for human–chicken alignments is much greater than that seen for human–rodent alignments (Fig. 15b). The higher density of interspersed repeats in the human genome accounts for ~10% of this variation (correlation coefficient 0.326, $P < 10^{-4}$), with the strongest positive association at intermediate ranges (Fig. 15c). G+C content also accounts for a portion of the residual variability: after subtracting repeats from human and chicken in each window, the resulting length and G+C content ratios are strongly and negatively related (correlation coefficient −0.256, $P < 10^{-4}$) (Fig. 15d).

More extensive multivariate regression analyses using these and other genomic parameters derived from human–chicken alignments explain only 20–25% of the variability in length ratios (see Methods and Supplementary Table S6). Features not measured in the current analysis are thus the main contributors to changes in genome size. One potential hidden variable is the density of ancient

Table 11 **Comparison of orthologous segments**

Feature	Anchors	
	Orthologous genomic segments	Orthologous gene pairs
Number of blocks	1,068	1,009
Average size in human (Mb)	2.10	2.03
Average size in chicken (Mb)	0.78	0.76
Average human:chicken ratio	2.67	2.67
Median size in human (kb)	177	566
Median size in chicken (kb)	74	185
Median human:chicken ratio	2.39	3.06
Longest block in human (Mb)	91.5	48.5
Longest block in chicken (Mb)	38.2	19.8
Coverage per chicken chromosome (%)	71.1	66.1
Coverage per human chromosome (%)	72.7	66.8

repeat sequences that have diverged beyond the ability to be found confidently by RepeatMasker[144]. For example, repeat expansions in G+C-poor regions during early mammalian evolution might contribute to length ratio variation.

Illuminating the human genome: conserved non-coding segments

One important use of the chicken sequence is to identify functional non-coding elements in the human genome. The high synonymous substitution rates described above would preclude neutral regions from aligning between human and chicken. Thus the regions that do align are likely to be subject to purifying selection and hence are predicted to be functional. Compared with approximately 5% of the human sequence estimated to be under purifying selection, based on human–mouse comparisons[2,145], only 2.5% of the human sequence aligned with chicken. Of aligned positions, 44% are in protein-coding regions, 25% are intronic and 31% are intergenic. Aligned sequences within genes were discussed above in relation to chicken gene content. Here we investigate the global properties of the complete set of non-coding aligned segments.

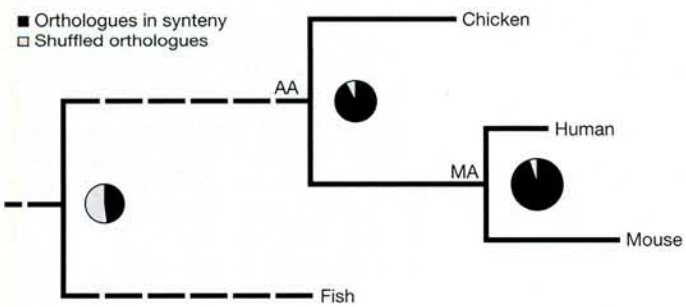
Conservation patterns in proximal *cis*-regulatory regions

We examined functional non-coding elements in the human genome to analyse how frequently they are conserved with chicken. For sets of known regulatory elements, functional promoters, predicted promoters, CpG islands, and predicted transcription-factor-binding sites that are conserved between human and rodents, we find that only 30–40% are conserved between human and chicken (Fig. 16). Even for coding exons, which are among the most highly constrained elements in the genome, only 75% aligned between human and chicken. Thus, conservation between mammalian and chicken genomes greatly increases specificity in the search for functional elements, but at a price in sensitivity that varies depending on the functional category. The conserved regulatory modules (CRMs; for example, promoters and enhancers) that still align between human and chicken are not a coincidental result of a lower local rate of neutral divergence because the distribution of neutral substitution values[2,126] for the regions containing conserved CRMs was indistinguishable from that for the diverged CRMs. We assessed whether the biological functions of genes regulated by the conserved CRMs might differ from those with divergent CRMs. We found that genes from three GO categories (development, metabolism and structural component of muscle) are over-represented in the set regulated by the conserved CRMs, whereas genes within the signal transducer and hydrolase activity GO categories are over-represented in the set with non-conserved regulatory elements (Fig. 17).

Conserved non-coding regions are clustered and far from genes

Several recent human–rodent comparisons[146–148] within limited regions of the genome have observed that conserved non-coding segments tend to lie relatively far from genes. To investigate whether this holds true across the genome in human–chicken comparisons, we partitioned the human genome into non-overlapping intervals, and in each interval (after excluding repetitive nucleotides) computed the fraction that falls in a coding exon annotated by RefSeq or Ensembl (the 'coding fraction') and the fraction that aligns with chicken but does not intersect coding exons (the 'conserved non-coding fraction' or CNF), taking care to keep the two quantities logically independent (see Methods). The two quantities are negatively associated (correlation coefficient -0.197, $P = 0.000$; Supplementary Fig. S15). This inverse correlation is robust to variation in the interval size (between 100 and 1,000 kb), the set of gene annotations, and the set of non-coding alignments, and it is not explained by G+C content alone (see Methods). Proximal *cis*-regulatory elements (that is, those within a few kilobases of an exon) thus seem not to be the primary contributors to overall non-coding conservation.

a



b

Divergence time (Myr)	Branch	All rearrangements					Interchromosomal				
		Normalized in time	Break count		Reconstruction		Normalized in time	Break count		Reconstruction	
			Raw	Normalized	Raw	Normalized		Raw	Normalized	Raw	Normalized
310	AA–chicken	0.84	35	10.2	389	5.3	0.5	>46	2.1	102	3.4
223	AA–MA	1.17	48								
87	MA–human	1	8	1	73	1	1	22	1	30	1
87	MA–mouse	2.3	21	2.6	160	2.2	3.2	74	3.4	94	3.1

Figure 14 Rates of genome structure divergence. **a**, Phylogenetic tree of chicken, human and mouse showing rates of genome structure divergence, where the branch length is proportional to the number of estimated chromosomal rearrangements. Details are given in the supporting table (**b**). AA, amniote ancestor; MA, mammalian ancestor. The pie charts show the fraction of orthologous genes that have retained their genomic neighbourhood; for example, about 85% of chicken and human orthologous genes reside in orthologous chromosomal segments, and their sizes are proportional to the number of recognizable orthologous genes. **b**, The table highlights a very low rate of interchromosomal shuffling in early mammalian and bird evolution, and an elevated rate of interchromosomal shuffling in the rodent lineage. It provides a summary of the number of chromosomal rearrangements estimated by counting synteny breaks where ancestral state is supported by synteny to an outgroup species, and by reconstruction of ancestral genomes through a combinatorial search for the most parsimonious rearrangement pattern. The two methods agree reasonably and have been used to estimate the relative branch lengths of the genome structure divergence tree in **a**. Normalization to the length of the MA–human branch and to the time of independent evolution is presented for easy comparison.

Non-coding conservation is not uniformly distributed across the human genome. Fifty-seven segments of high non-coding conservation (average length 1.176 Mb and average CNF 13.1%, compared with a genome average of 1.7%; see Methods and Supplementary Table S7) were found to be gene poor (they cover 2.3% of the human genome but contain only 0.3% of the exons in RefSeq genes). They were also G+C poor (38.8% overall) and depleted for all classes of interspersed repeats, as are their chicken orthologues. They contain none of the 731 break points identified through an analysis of regions of conserved gene order between human and chicken with length exceeding 200 kb, and—from human–mouse neutral substitution rates estimated using interspersed repeats[126]—do not seem to have experienced particularly low mutation rates. The genes within or overlapping these 57 high-CNF segments, however, are significantly enriched for the GO categories associated with gene regulation (Fig. 17). Thus, the degree of non-coding sequence conservation is related to the biological function of the genes in the general genome neighbourhood. This enrichment is not merely a by-product of the gene-poor nature of these segments, because human gene-poor regions with low CNF are not enriched for the same GO categories[149].

The 57 high-CNF segments were found to contain 60.7% of 417 human–chicken ultraconserved elements (UCEs): sequences defined as being 200 bp or longer in orthologous chromosomal locations and 100% identical between the two species. Only 27.3% of these human–chicken UCEs overlap the 481 previously studied human–rodent UCEs[150], although all of them are found in the draft genomes of either mouse or rat, at 87–100% identity in the more conserved rodent. Human–chicken UCEs differ from human–rodent UCEs in containing far fewer exon-intersecting elements (7.9% versus 23%) and in showing only weak enrichment for proximity to genes whose products have a GO and InterPro classification associated with RNA splicing. The two UCE sets share a strong bias against harbouring human-verified single-nucleotide polymorphisms ($P < 10^{-40}$ compared to the genome-wide average in both cases). In both, the set of elements with no expression evidence (non-exonics in ref. 150) is found in or next to genes whose products are highly enriched for transcriptional regulation, DNA binding, homeodomains and developmental functions; many of these UCEs are found more than 100 kb away from the corresponding genes. Finally, in both sets, no UCEs except a handful of coding regions could be traced back through sequence similarity to *Ciona intestinalis*, *Caenorhabditis elegans* or *Drosophila melanogaster*. At the moment, little is understood about the functional significance of either the UCEs or the high-CNF segments that are far from genes.

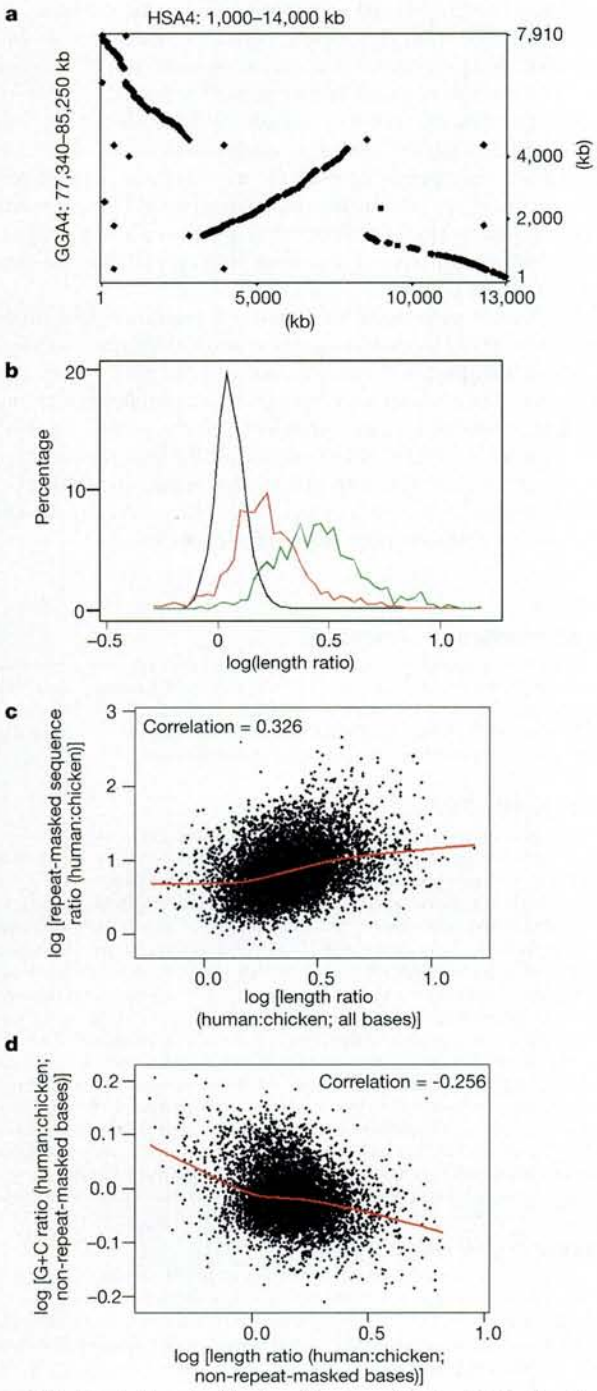


Figure 15 Variation in the ratio of lengths of human and chicken DNA in aligned segments. **a**, Dot plot comparing orthologous regions of human and chicken, showing variable slope. **b**, Variation in log-transformed ratio of lengths of aligned segments, comparing human–chicken (green, all; non-repetitive, red) and human–mouse (black). **c**, Scatter plot of the log-transformed ratio of lengths of orthologous human and chicken segments versus ratio of repeat-masked sequence in the two species. **d**, Scatter plot of the log-transformed ratio of lengths versus ratio of G+C contents after removal of masked bases. Lowess smooths (locally weighted scatterplot smoothing) are superimposed (red curves; smoothing parameter 0.5). See Methods for details.

Conclusion

The chicken genome represents an intermediate test case as a target

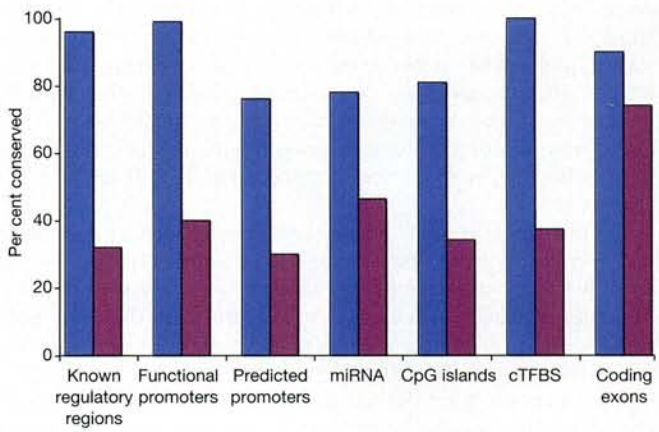


Figure 16 All sets of functional elements in human–chicken alignments show reduced representation relative to human–mouse–rat alignments. We examined functional elements containing known regulatory regions[2,164], functional promoters and predicted promoters[165], miRNA[34], CpG islands, conserved transcription factor binding sites[3], and coding exons of known genes. The per cent of each category that aligns is shown for the human–mouse–rat alignments (HMR, blue) or human–chicken alignments (HC, red).

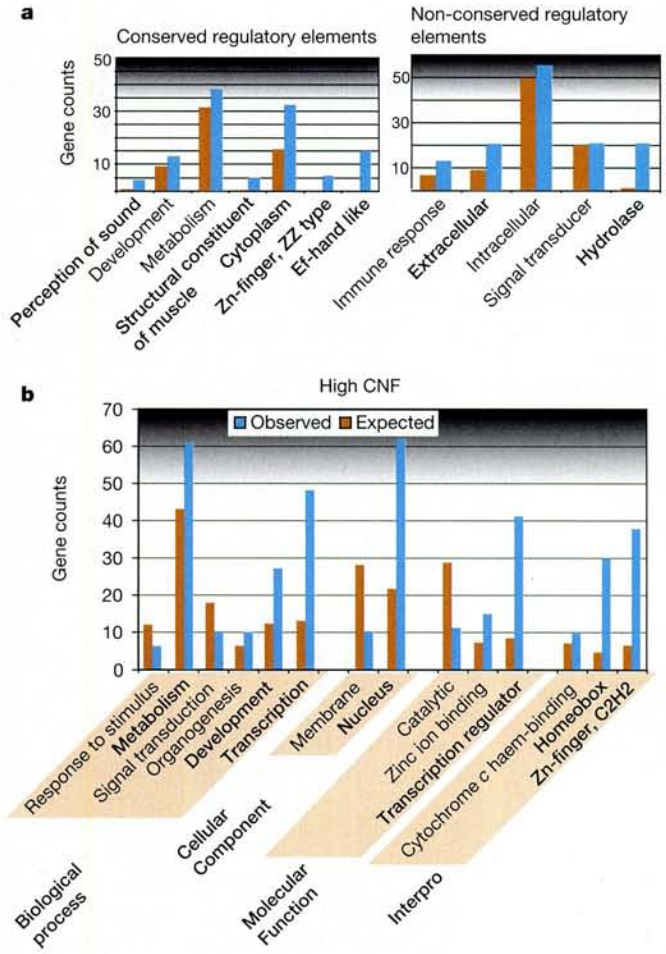


Figure 17 Enrichment of particular GO categories in genes regulated by conserved or non-conserved *cis*-regulatory modules (**a**) and in high-CNF regions (**b**). The observed and expected gene counts are shown for GO categories with at least five genes from high-CNF segments. Significantly enriched categories are in bold; the largest significant *P*-value is 0.0022 and the others range from 10^{-10} to 10^{-25}.

for genome sequence assembly and analysis. Although less than half the size of mammalian genomes, it is still much larger than those of *D. melanogaster*, *C. elegans* and even *T. rubripes*, and it lacks the dense linkage map platform that helped to assemble the first two. Unlike the rat and chimpanzee genomes, there was no closely related, high-quality genome sequence already available to provide a framework for assembly. Nevertheless, a relatively high-quality draft of the chicken genome has been achieved on the basis of only 6.6 × whole-genome shotgun coverage, owing in part to the remarkably low level of recent transposon activity it has endured (Table 5).

The quality of this draft genome sequence makes it a key resource for comparative genomics. Natural selection and evolution provide us with many perspectives from which to view our own genome. Genomes of distant species resolve key processes that have been conserved over millennia, whereas those of our close relatives allow an analysis of rapidly changing sequence. In many respects, the chicken genome provides insights that were unavailable from previous sequences. For nearly every aspect of biology, it allows us to distinguish features of mammalian biology that are derived or ancient, and it reveals examples of mammalian innovation and adaptation.

The chicken is sufficiently distant that little unselected sequence has survived unchanged along the separate evolutionary paths to birds and mammals from their last common ancestor. Against this background, conserved non-coding sequences stand out clearly. Some of these represent known regulators (Fig. 16) and others use novel mechanisms yet to be identified. On the other hand, the counterparts of many functional mammalian sequence elements could not be identified in the chicken sequence. Either these represent mammalian innovations or else any commonality has been lost over the course of >310 Myr of separate mutation and fixation.

Chicken breeding, based on quantitative genetic methods, represents one of the most remarkable examples of directed evolution. Even after 50 yr of intensive selection, annual genetic progress in production traits remains undiminished[151]. An impressive list of chicken quantitative trait loci has already been identified[152], many with combined agricultural and medical relevance. The chicken genome sequence promotes both the development of more refined polymorphic maps (see the accompanying paper[153]) and the framework for discovering the functional polymorphisms underlying interesting quantitative traits, thus fully exploiting the genetic potential of the chicken.

The chicken genome is invaluable for shedding light on functional elements of the human genome and our unique evolutionary history. It also points the way forward to the great utility we can expect from the genome sequences of other carefully chosen species. The data presented here demonstrate both the unique value of the chicken as a model species and emphasize the incomplete nature of our collective understanding of complex organisms. This chicken genome sequence will both integrate and stimulate the expanding array of contributions from this versatile species. □

Methods

Domain matching and ranking

To identify known families of genes and domains we scanned respective proteomes for characteristic HMM profile signatures from Pfam[69] and SMART databases using HMMER (http://hmmer.wustl.edu/) software and applying corresponding family-specific cutoffs. The identified families were ranked by the number of matching genes requiring at least one matching transcript, and only counting repetitive matches once.

Orthology detection

Orthologous relationships between genes of chicken, human, *Fugu* and others were inferred through systematic similarity searches at the level of the predicted proteins. We retained only the largest predicted ORF per locus, and compared those in an all-against-all fashion using the Smith–Waterman algorithm. We then formed orthologous groups using a variant of a strategy used earlier[137,154,155]. First, we grouped recently duplicated sequences within genomes into 'paralogous groups', to be treated as single sequences subsequently. For this, there was no fixed cutoff in similarity, but instead we started with a stringent similarity cutoff and relaxed it on each successive step, until all paralogous proteins were joined, thereby satisfying the following criteria: all members of a group had to be more similar to each other than to any other protein in any other genome; and all members of the group had to have hits that overlapped by at least 20 residues, to avoid 'domain walking'. After grouping paralogous proteins, we started to assign orthology between proteins by joining triangles of reciprocal best hits involving three different species (here, paralogous groups were represented by their best-matching member). Again, a stringent similarity cutoff was used first and relaxed on each successive step, and all proteins in a group were required to have hits overlapping by at least 20 residues. Finally, we joined any remaining nodes by allowing not only reciprocal triangles, but also reciprocal tuples.

Detection of gene loss in mammals

The orthologous relations defined above were used to infer losses when a gene was found in chicken and in at least one earlier-branching animal, but not in any mammal. Of 122 candidate losses obtained in this manner, many were manually discounted after Blastp searches in mammalian genomes (thus hinting that several as-yet-unannotated genes in mammals remain to be predicted).

Detection of orthologous introns

For each orthologous group we created a multiple alignment and mapped intron positions and protein features onto it. This procedure is incorporated into the SMART web server. To minimize errors due to erroneous alignments, introns flanking alignment gaps were discarded (less than 1% of all introns). To compensate for effects of intron sliding and to reduce further the impact of possible alignment errors, we allowed a window of 12 nucleotides in which we considered a position as conserved. Previous estimates indicated that the chance of independent intron insertion in such a window is <1%. To avoid biases due to incomplete gene predictions, we omitted 18,910 introns in regions that were missing from some of the predicted genes.

Deriving tissue expression data

Chicken ESTs were mapped to the assembly, and to Ensembl genes (±1 kb), using BLAT and a 95% identity threshold, and were partitioned into ten (brain; fat and skin; bone and connective tissues; heart; kidney and adrenal tissues; immune; liver; female reproduction; alimentary; testis) distinct tissue types. Percentage amino acid sequence identities of 1:1 chicken–human orthologues were calculated as described previously (Fig. 6). Note that single genes may be assigned to multiple tissues.

Whole-genome alignments

Human–chicken whole-genome alignments were obtained by using the program BLASTZ[156] to produce short (typically 100–1,000 bp) local alignments, and then assembling gap-free segments of those alignments into 'chains' in which aligned segments occur in the same order and orientation in both species[134]. These alignments—which were used to generate data for Figs 2, 15–17 and Table 4, and Supplementary Figs S1, S15 and Supplementary Tables S6 and S7—can be obtained from the U.C. Santa Cruz Browser (http://genome.ucsc.edu/). To compute the CNF of a human genomic interval, we limited consideration to non-repetitive bases that are not in a local alignment that intersects a protein-coding region, and determined the fraction of those bases that are within an alignment.

Evolution of vertebrate genomes

The maps of conserved synteny (orthologous chromosomal segments with a conserved gene neighbourhood[133,134,137]) between chicken, human and mouse were produced using whole-genome DNA alignments post-processed into chains and nets[134] as well as looking for a conserved neighbourhood of orthologous genes as described previously[133,137]. We used gene-based synteny as input to MGR[157] and GRIMM[133,158] to look for parsimonious scenarios of rearrangements, starting from a set of 6,447 four-way orthologous genes pre-filtered for evidence of conserved pairwise synteny using SyntQL (E. Zdobnov, unpublished program) and applying GRIMM-Synteny[133] for more stringently defined 586 four-way human–mouse–rat–chicken synteny blocks.

Detailed descriptions of all methods are provided in the Supplementary Information.

Received 19 July; accepted 1 November 2004; doi:10.1038/nature03154.

1. The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
3. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
4. Hedges, S. B. The origin and evolution of model organisms. *Nature Rev. Genet.* **3,** 838–849 (2002).
5. Reisz, R. R. & Muller, J. Molecular timescales and the fossil record: a paleontological perspective. *Trends Genet.* **20,** 237–241 (2004).
6. Duret, L. & Bucher, P. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* **7,** 399–406 (1997).
7. Gottgens, B. *et al.* Analysis of vertebrate SCL loci identifies conserved enhancers. *Nature Biotechnol.* **18,** 181–186 (2000).
8. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424,** 788–793 (2003).
9. Ostrum, J. H. The origin of birds. *Annu. Rev. Earth Planet. Sci.* **3,** 55–77 (1975).
10. Sereno, P. C. The evolution of dinosaurs. *Science* **284,** 2137–2147 (1999).
11. Shinan, R. Several major achievements in early Neolithic China, ca. 5000 BC. *Kaogu–Archaeology* **1996,** 37–49 (1996); (trans. Cheung, W. K.) (ed. Gordon, B.) (http://www.carleton.ca/~bgordon/Rice/papers/REN96.htm).
12. Fitzpatrick, D. M. & Ahmed, K. Red roving fowl. *Down Earth* **9,** 28 (2000).
13. Crawford, R. D. (ed.) *Poultry Breeding and Genetics* (Elsevier, Amsterdam, 1995).
14. Darwin, C. *The Variation of Animals and Plants Under Domestication* (D. Appleton and Co., New York, 1896).
15. Fumihito, A. *et al.* One subspecies of the red jungle fowl (*Gallus gallus gallus*) suffices as the matriarchic ancestor of all domestic breeds. *Proc. Natl Acad. Sci. USA* **91,** 12505–12509 (1994).
16. Punnett, R. C. *Heredity in Poultry* (Macmillan, London, 1923).
17. Bateson, W. & Saunders, E. R. Experimental studies in the physiology of heredity. *Rep. Evol. Commun. R. Soc.* **1,** 1–160 (1902).
18. Pisenti, J. M. *et al.* Avian genetic resources at risk: An assessment and proposal for conservation of genetic stocks in the USA and Canada. *Avian Poultry Biol. Rev.* **12,** 1–102 (2001).
19. Brown, W. R., Hubbard, S. J., Tickle, C. & Wilson, S. A. The chicken as a model for large-scale analysis of vertebrate gene function. *Nature Rev. Genet.* **4,** 87–98 (2003).
20. Vogt, P. K. *Historical Introduction to the General Properties of Retroviruses* (eds Coffin, J. M., Hughes, S. H. & Varmus, H. E.) (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1997).
21. Stehelin, D., Varmus, H. E., Bishop, J. M. & Vogt, P. K. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260,** 170–173 (1976).
22. Cooper, M. D., Raymond, D. A., Peterson, R. D., South, M. A. & Good, R. A. The functions of the thymus system and the bursa system in the chicken. *J. Exp. Med.* **123,** 75–102 (1966).
23. Hutt, F. B. *Genetics of the Fowl* (McGraw-Hill, New York, 1949).
24. Bitgood, J. J. & Somes, R. G. J. in *Genetic Maps* (ed. O'Brien, S.) 4333–4342 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1993).
25. Groenen, M. A. *et al.* A consensus linkage map of the chicken genome. *Genome Res.* **10,** 137–147 (2000).
26. Bloom, S. E., Delany, M. E. & Muscarella, D. E. *Constant and Variable Features of Avian Chromosomes* (eds Gibbins, A. & Etches, R. J.) (CRC Press, Boca Raton, Florida, 1993).
27. Huang, X., Wang, J., Aluru, S., Yang, S. P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Res.* **13,** 2164–2170 (2003).
28. Wallis, J. W. *et al.* A physical map of the chicken genome. *Nature* doi:10.1038/nature03030 (this issue).
29. Schmid, M. *et al.* First report on chicken genes and chromosomes 2000. *Cytogenet. Cell Genet.* **90,** 169–218 (2000).
30. Romanov, M. N., Price, J. A. & Dodgson, J. B. Integration of animal linkage and BAC contig maps using overgo hybridization. *Cytogenet. Genome Res.* **102,** 277–281 (2003).
31. Burt, D. W. *Comparative Genomics in Poultry Breeding and Biotechnology* (eds Muir, W. M. & Aggrey, S. E.) (CAB International, Wallingford, Oxon, 2003).
32. Hubbard, S. J. *et al.* Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 Expressed Sequence Tags. *Genome Res.* (in the press).
33. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. *Nucleic Acids Res.* **31,** 439–441 (2003).
34. Griffiths-Jones, S. The microRNA registry. *Nucleic Acids Res.* **32** (Database issue), D109–D111 (2004).
35. Hirose, T. & Steitz, J. A. Position within the host intron is critical for efficient processing of box C/D snoRNAs in mammalian cells. *Proc. Natl Acad. Sci. USA* **98,** 12914–12919 (2001).
36. Curwen, V. *et al.* The Ensembl automatic gene annotation system. *Genome Res.* **14,** 942–950 (2004).
37. Flicek, P., Keibler, E., Hu, P., Korf, I. & Brent, M. R. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Res.* **13,** 46–54 (2003).
38. Parra, G. *et al.* Comparative gene prediction in human and mouse. *Genome Res.* **13,** 108–117 (2003).
39. Ashurst, J. L. & Collins, J. E. Gene annotation: prediction and testing. *Annu. Rev. Genomics Hum. Genet.* **4,** 69–88 (2003).
40. Birney, E. *et al.* An overview of Ensembl. *Genome Res.* **14,** 925–928 (2004).
41. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence project: update and current status. *Nucleic Acids Res.* **31,** 34–37 (2003).
42. Strausberg, R. L. *et al.* Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA* **99,** 16899–16903 (2002).
43. Imanishi, T. *et al.* Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* **2,** E162 (2004).
44. Wu, J. Q., Shteynberg, D., Arumugam, M., Gibbs, R. A. & Brent, M. R. Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing. *Genome Res.* **14,** 665–671 (2004).
45. Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
46. Kozak, M. How do eukaryotic ribosomes select initiation regions in messenger RNA? *Cell* **15,** 1109–1123 (1978).
47. Reymond, A. *et al.* Nineteen additional unpredicted transcripts from human chromosome 21. *Genomics* **79,** 824–832 (2002).
48. Abril, J. F., Castelo, R. & Guigo, R. Comparison of splice sites in mammals and chicken. *Genome Res.* (in the press).
49. Aruscavage, P. J. & Bass, B. L. A phylogenetic analysis reveals an unusual sequence conservation within introns involved in RNA editing. *RNA* **6,** 257–269 (2000).
50. Hoopengardner, B., Bhalla, T., Staber, C. & Reenan, R. Nervous system targets of RNA editing identified by comparative genomics. *Science* **301,** 832–836 (2003).
51. Long, M., Betran, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nature Rev. Genet.* **4,** 865–875 (2003).
52. Torrents, D., Suyama, M., Zdobnov, E. & Bork, P. A genome-wide survey of human pseudogenes. *Genome Res.* **13,** 2559–2567 (2003).
53. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24,** 363–367 (2000).
54. Burch, J. B., Davis, D. L. & Haas, N. B. Chicken repeat 1 elements contain a pol-like open reading frame and belong to the non-long terminal repeat class of retrotransposons. *Proc. Natl Acad. Sci. USA* **90,** 8199–8203 (1993).
55. Haas, N. B. *et al.* Subfamilies of CR1 non-LTR retrotransposons have different 5′ UTR sequences but are otherwise conserved. *Gene* **265,** 175–183 (2001).
56. Olofsson, B. & Bernardi, G. The distribution of CR1, and Alu-like family of interspersed repeats, in the chicken genome. *Biochim. Biophys. Acta* **740,** 339–341 (1983).
57. Haas, N. B., Grabowski, J. M., Sivitz, A. B. & Burch, J. B. Chicken repeat 1 (CR1) elements, which define an ancient family of vertebrate non-LTR retrotransposons, contain two closely spaced open reading frames. *Gene* **197,** 305–309 (1997).
58. Adey, N. B., Tollefsbol, T. O., Sparks, A. B., Edgell, M. H. & Hutchison, C. A. III Molecular resurrection of an extinct ancestral promoter for mouse L1. *Proc. Natl Acad. Sci. USA* **91,** 1569–1573 (1994).
59. Smit, A. F., Toth, G., Riggs, A. D. & Jurka, J. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* **246,** 401–417 (1995).
60. Ohshima, K., Hamada, M., Terai, Y. & Okada, N. The 3′ ends of tRNA-derived short interspersed repetitive elements are derived from the 3′ ends of long interspersed repetitive elements. *Mol. Cell. Biol.* **16,** 3756–3764 (1996).
61. Cordonnier, A., Casella, J. F. & Heidmann, T. Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *J. Virol.* **69,** 5890–5897 (1995).
62. Smit, A. F. Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res.* **21,** 1863–1872 (1993).
63. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9,** 657–663 (1999).
64. Duret, L. & Mouchiroud, D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17,** 68–74 (2000).
65. Winter, E. E., Goodstadt, L. & Ponting, C. P. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.* **14,** 54–61 (2004).
66. Zhang, L. & Li, W. H. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.* **21,** 236–239 (2004).
67. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297,** 1301–1310 (2002).
68. Mulder, N. J. *et al.* The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31,** 315–318 (2003).
69. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32** (Database issue), D138–D141 (2004).

70. Letunic, I. *et al.* SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* **32** (Database issue), D142–D144 (2004).
71. Fickenscher, H. & Pirzer, H. Interleukin-26. *Int. Immunopharmacol.* **4,** 609–613 (2004).
72. Copley, R. R., Goodstadt, L. & Ponting, C. Eukaryotic domain evolution inferred from genome comparisons. *Curr. Opin. Genet. Dev.* **13,** 623–628 (2003).
73. Kawasaki, K. & Weiss, K. M. Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster. *Proc. Natl Acad. Sci. USA* **100,** 4060–4065 (2003).
74. Williams, A. J., Blacklow, S. C. & Collins, T. The zinc finger-associated SCAN box is a conserved oligomerization domain. *Mol. Cell. Biol.* **19,** 8526–8535 (1999).
75. Sander, T. L. *et al.* The SCAN domain defines a large family of zinc finger transcription factors. *Gene* **310,** 29–38 (2003).
76. Hughes, A. L. & Friedman, R. Differential loss of ancestral gene families as a source of genomic divergence in animals. *Proc. R. Soc. Lond. B* **271** (suppl. 3), S107–S109 (2004).
77. Krylov, D. M., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* **13,** 2229–2235 (2003).
78. Kang, W. & Reid, K. B. DMBT1, a regulator of mucosal homeostasis through the linking of mucosal defense and regeneration? *FEBS Lett.* **540,** 21–25 (2003).
79. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14,** 988–995 (2004).
80. Shiina, T. *et al.* Genomic anatomy of a premier major histocompatibility complex paralogous region on chromosome 1q21-q22. *Genome Res.* **11,** 789–802 (2001).
81. Amadou, C. *et al.* Co-duplication of olfactory receptor and MHC class I genes in the mouse major histocompatibility complex. *Hum. Mol. Genet.* **12,** 3025–3040 (2003).
82. Malnic, B., Godfrey, P. A. & Buck, L. B. The human olfactory receptor gene family. *Proc. Natl Acad. Sci. USA* **101,** 2584–2589 (2004).
83. Alcock, J. *Animal Behaviour* (Sinauer Associates, Sunderland, Massachusetts, 1989).
84. Jones, R. B. & Roper, T. J. Olfaction in the domestic fowl: a critical review. *Physiol. Behav.* **62,** 1009–1018 (1997).
85. Mefford, H. C., Linardopoulou, E., Coil, D., van den Engh, G. & Trask, B. J. Comparative sequencing of a multicopy subtelomeric region containing olfactory receptor genes reveals multiple interactions between non-homologous chromosomes. *Hum. Mol. Genet.* **10,** 2363–2372 (2001).
86. Mefford, H. C. & Trask, B. J. The complex structure and dynamic evolution of human subtelomeres. *Nature Rev. Genet.* **3,** 91–102 (2002).
87. Rogers, M. A. *et al.* Characterization of a cluster of human high/ultrahigh sulfur keratin-associated protein genes embedded in the type I keratin gene domain on chromosome 17q12–21. *J. Biol. Chem.* **276,** 19440–19451 (2001).
88. Hesse, M., Zimek, A., Weber, K. & Magin, T. M. Comprehensive analysis of keratin gene clusters in humans and rodents. *Eur. J. Cell Biol.* **83,** 19–26 (2004).
89. Altmann, S. M., Mellon, M. T., Distel, D. L. & Kim, C. H. Molecular and functional analysis of an interferon gene from the zebrafish, *Danio rerio*. *J. Virol.* **77,** 1992–2002 (2003).
90. Hughes, A. L. & Roberts, R. M. Independent origin of IFN-α and IFN-β in birds and mammals. *J. Interferon Cytokine Res.* **20,** 737–739 (2000).
91. Smale, L., Lee, T. & Nunez, A. A. Mammalian diurnality: some facts and gaps. *J. Biol. Rhythms* **18,** 356–366 (2003).
92. Thoma, F. Light and dark in chromatin repair: repair of UV-induced DNA lesions by photolyase and nucleotide excision repair. *EMBO J.* **18,** 6585–6598 (1999).
93. Reverchon, S., Rouanet, C., Expert, D. & Nasser, W. Characterization of indigoidine biosynthetic genes in *Erwinia chrysanthemi* and role of this blue pigment in pathogenicity. *J. Bacteriol.* **184,** 654–665 (2002).
94. Shannon, M., Hamilton, A. T., Gordon, L., Branscomb, E. & Stubbs, L. Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. *Genome Res.* **13,** 1097–1110 (2003).
95. Zhao, G. Q. *et al.* The receptors for mammalian sweet and umami taste. *Cell* **115,** 255–266 (2003).
96. Bradbury, J. Taste perception: cracking the code. *PLoS Biol.* **2,** E64 (2004).
97. Bufe, B., Hofmann, T., Krautwurst, D., Raguse, J. D. & Meyerhof, W. The human TAS2R16 receptor mediates bitter taste in response to β-glucopyranosides. *Nature Genet.* **32,** 397–401 (2002).
98. Shi, P., Zhang, J., Yang, H. & Zhang, Y. P. Adaptive diversification of bitter taste receptor genes in mammalian evolution. *Mol. Biol. Evol.* **20,** 805–814 (2003).
99. Nordling, E., Persson, B. & Jornvall, H. Differential multiplicity of MDR alcohol dehydrogenases: enzyme genes in the human genome versus those in organisms initially studied. *Cell. Mol. Life Sci.* **59,** 1070–1075 (2002).
100. Hjelmqvist, L., Estonius, M. & Jornvall, H. The vertebrate alcohol dehydrogenase system: variable class II type form elucidates separate stages of enzymogenesis. *Proc. Natl Acad. Sci. USA* **92,** 10904–10908 (1995).
101. Tamir, H. & Ratner, S. Enzymes of arginine metabolism in chicks. *Arch. Biochem. Biophys.* **102,** 249–258 (1963).
102. McQueen, H. A. *et al.* CpG islands of chicken are concentrated on microchromosomes. *Nature Genet.* **12,** 321–324 (1996).
103. Andreozzi, L. *et al.* Compositional mapping of chicken chromosomes and identification of the gene-richest regions. *Chromosome Res.* **9,** 521–532 (2001).
104. Smith, J. *et al.* Differences in gene density on chicken macrochromosomes and microchromosomes. *Anim. Genet.* **31,** 96–103 (2000).
105. McQueen, H. A., Siriaco, G. & Bird, A. P. Chicken microchromosomes are hyperacetylated, early replicating, and gene rich. *Genome Res.* **8,** 621–630 (1998).
106. Grutzner, F. *et al.* Chicken microchromosomes are hypermethylated and can be identified by specific painting probes. *Cytogenet. Cell Genet.* **93,** 265–269 (2001).
107. Schmid, M., Enderle, E., Schindler, D. & Schempp, W. Chromosome banding and DNA replication patterns in bird karyotypes. *Cytogenet. Cell Genet.* **52,** 139–146 (1989).
108. Ponce de Leon, F. A., Li, Y. & Weng, Z. Early and late replicative chromosomal banding patterns of *Gallus domesticus*. *J. Hered.* **83,** 36–42 (1992).
109. Habermann, F. A. *et al.* Arrangements of macro- and microchromosomes in chicken cells. *Chromosome Res.* **9,** 569–584 (2001).
110. Holmquist, G. P. Evolution of chromosome bands: molecular ecology of noncoding DNA. *J. Mol. Evol.* **28,** 469–486 (1989).
111. Jensen-Seaman, M. I. *et al.* Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* **14,** 528–538 (2004).
112. Meunier, J. & Duret, L. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21,** 984–990 (2004).
113. Yang, S. *et al.* Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes. *Genome Res.* **14,** 517–527 (2004).
114. Rodionov, A. V. Micro vs. macro: structural-functional organization of avian micro- and macrochromosomes. *Genetika* **32,** 597–608 (1996).
115. Marais, G. Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* **19,** 330–338 (2003).
116. Eyre-Walker, A. & Hurst, L. D. The evolution of isochores. *Nature Rev. Genet.* **2,** 549–555 (2001).
117. Montoya-Burgos, J. I., Boursot, P. & Galtier, N. Recombination explains isochores in mammalian genomes. *Trends Genet.* **19,** 128–130 (2003).
118. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8,** 1499–1504 (1980).
119. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196,** 261–282 (1987).
120. Carvalho, A. B. & Clark, A. G. Intron size and natural selection. *Nature* **401,** 344 (1999).
121. Duret, L., Mouchiroud, D. & Gautier, C. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.* **40,** 308–317 (1995).
122. Hurst, L. D., Brunton, C. F. & Smith, N. G. Small introns tend to occur in GC-rich regions in some but not all vertebrates. *Trends Genet.* **15,** 437–439 (1999).
123. Versteeg, R. *et al.* The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* **13,** 1998–2004 (2003).
124. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8,** 269–294 (1966).
125. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13,** 555–556 (1997).
126. Hardison, R. C. *et al.* Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13,** 13–26 (2003).
127. Axelsson, E., Webster, M. T., Smith, N. G. C., Burt, D. W. & Ellegren, H. Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. *Genome Res.* (in the press).
128. Wilkie, A. O. *et al.* Stable length polymorphism of up to 260 kb at the tip of the short arm of human chromosome 16. *Cell* **64,** 595–606 (1991).
129. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002).
130. Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M. & Eichler, E. E. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res.* **14,** 789–801 (2004).
131. Tuzun, E., Bailey, J. A. & Eichler, E. E. Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res.* **14,** 493–506 (2004).
132. Cheung, J. *et al.* Recent segmental and gene duplications in the mouse genome. *Genome Biol.* **4,** R47 (2003).
133. Pevzner, P. & Tesler, G. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.* **13,** 37–45 (2003).
134. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **100,** 11484–11489 (2003).
135. Bourque, G., Pevzner, P. A. & Tesler, G. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.* **14,** 507–516 (2004).
136. Murphy, W. J., Bourque, G., Tesler, G., Pevzner, P. A. & O'Brien, S. J. Reconstructing the genomic architecture of mammalian ancestors using multispecies comparative maps. *Human Genomics* **1,** 30–40 (2003).
137. Zdobnov, E. M. *et al.* Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298,** 149–159 (2002).
138. Jaillon, O. *et al.* Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431,** 946–957 (2004).
139. Burt, D. W. *et al.* The dynamics of chromosome evolution in birds and mammals. *Nature* **402,** 411–413 (1999).
140. Murphy, W. J. *et al.* Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294,** 2348–2351 (2001).
141. Bourque, G., Zdobnov, E. M., Bork, P., Pevzner, P. A. & Tesler, G. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res.* (in the press).
142. Stanyon, R., Stone, G., Garcia, M. & Froenicke, L. Reciprocal chromosome painting shows that squirrels, unlike murid rodents, have a highly conserved genome organization. *Genomics* **82,** 245–249 (2003).
143. Gregory, T. R. Insertion-deletion biases and the evolution of genome size. *Gene* **324,** 15–34 (2004).
144. Smit, A. & Green, P. RepeatMasker ⟨http://ftp.genome.washington.edu/RM/RepeatMasker.html⟩ (1999).
145. Chiaromonte, F. *et al. The Genome of* Homo sapiens Vol. LXVIII (Cold Spring Harbor Press, Cold Spring Harbor, New York, 2003).
146. Dermitzakis, E. T. *et al.* Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420,** 578–582 (2002).
147. Dunham, A. *et al.* The DNA sequence and analysis of human chromosome 13. *Nature* **428,** 522–528 (2004).
148. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428,** 529–535 (2004).
149. Ovcharenko, I. *et al.* Evolution and functional classification of vertebrate gene deserts. *Genome Res.* (in the press).
150. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304,** 1321–1325 (2004).
151. Muir, W. M. & Aggrey, S. E. (eds) *Industrial Perspective on Problems and Issues Associated with Poultry Breeding* (CAB International, Wallingford, Oxon, 2003).

152. Andersson, L. & Georges, M. Domestic-animal genomics: deciphering the genetics of complex traits. *Nature Rev. Genet.* **5,** 202–212 (2004).
153. International Chicken Polymorphism Map Consortium. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature* doi:10.1038/nature03156 (this issue).
154. Koonin, E. V. A non-adaptationist perspective on evolution of genomic complexity or the continued dethroning of man. *Cell Cycle* **3,** 280–285 (2004).
155. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31,** 258–261 (2003).
156. Schwartz, S. *et al.* Human-mouse alignments with *Blastz*. *Genome Res.* **13,** 103–105 (2003).
157. Bourque, G. & Pevzner, P. A. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.* **12,** 26–36 (2002).
158. Tesler, G. GRIMM: genome rearrangements web server. *Bioinformatics* **18,** 492–493 (2002).
159. Benton, M. J. *Vertebrate Palaeontology* (Blackwell Science, Oxford, 2000).
160. Kapitonov, V. V. & Jurka, J. The esterase and PHD domains in CR1-like non-LTR retrotransposons. *Mol. Biol. Evol.* **20,** 38–46 (2003).
161. Vandergon, T. L. & Reitman, M. Evolution of chicken repeat 1 (CR1) elements: evidence for ancient subfamilies and multiple progenitors. *Mol. Biol. Evol.* **11,** 886–898 (1994).
162. Kedishvili, N. Y. *et al.* cDNA sequence and catalytic properties of a chick embryo alcohol dehydrogenase that oxidizes retinol and 3β,5α-hydroxysteroids. *J. Biol. Chem.* **272,** 7494–7500 (1997).
163. Estonius, M. *et al.* Avian alcohol dehydrogenase: the chicken liver enzyme. Primary structure, cDNA-cloning, and relationships to other alcohol dehydrogenases. *Eur. J. Biochem.* **194,** 593–602 (1990).
164. Elnitski, L. *et al.* Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13,** 64–72 (2003).
165. Trinklein, N. D., Aldred, S. J., Saldanha, A. J. & Myers, R. M. Identification and functional analysis of human transcriptional promoters. *Genome Res.* **13,** 308–312 (2003).
166. Caldwell, R. B. *et al.* Full-length cDNAs from bursal lymphocytes to facilitate gene function analysis. *Genome Biol.* (in the press).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements The *Gallus gallus* sequence and map generation at Washington University School of Medicine's Genome Sequencing Center was supported by grants from the National Human Genome Research Institute (NHGRI). For work from other groups, we acknowledge the support of the Biotechnology and Biological Sciences Research Council, Center for Integrative Genomics Funds, Childcare and Lejeune Foundations, Chinese Academy of Sciences and Ministry of Science and Technology, Department of Energy, Desiree and Niels Yde Foundation, European Union, European Molecular Biology Laboratory, Fonds Quebeçois de la Recherche sur la Nature et les Technologies, Howard Hughes Medical Institute, National Institute for Diabetes and Digestive and Kidney Diseases, NHGRI, National Institutes of Health, National Natural Science Foundation of China, National Science Foundation, Novo Nordisk Foundation, Stowers Institute for Medical Research, Swedish Research Council, Swiss NCCR Frontiers in Genetics, Swiss National Science Foundation, USDA/CSREES National Research Initiative, USDA/CSREES National Animal Genome Research Program, Wallenberg Consortium North and the AgriFunGen program at the Swedish University of Agricultural Sciences, UK Medical Research Council, University of California Presidential Chair Fund, University of Pennsylvania Genomics Institute Award, University of Texas at Arlington, and the Wellcome Trust. Resources for exploring the sequence and annotation data are available on browser displays available at Ensembl (http://www.ensembl.org), UCSC (http://genome.ucsc.edu) and the NCBI (http://www.ncbi.nlm.nih.gov). We thank R. Waterston for advice regarding the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.K.W. (rwilson@watson.wustl.edu).

This *G. gallus* whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the project accession number AADN00000000. The version described in this paper is the first version, AADN01000000.

International Chicken Genome Sequencing Consortium

Overall coordination: LaDeana W. Hillier[1], Webb Miller[2], Ewan Birney[3], Wesley Warren[1], Ross C. Hardison[2], Chris P. Ponting[4], Peer Bork[5,6], David W. Burt[7], Martien A. M. Groenen[8], Mary E. Delany[9], Jerry B. Dodgson[10]

Genome fingerprint map, sequence and assembly: Asif T. Chinwalla[1], Paul F. Cliften[1], Sandra W. Clifton[1], Kimberly D. Delehaunty[1], Catrina Fronick[1], Robert S. Fulton[1], Tina A. Graves[1], Colin Kremitzki[1], Dan Layman[1], Vincent Magrini[1], John D. McPherson[1], Tracie L. Miner, Patrick Minx[1], William E. Nash[1], Michael N. Nhan[1], Joanne O. Nelson[1], Lachlan G. Oddy[1], Craig S. Pohl[1], Jennifer Randall-Maher[1], Scott M. Smith[1], John W. Wallis[1], Shiaw-Pyng Yang[1]

Mapping: Michael N. Romanov[10], Catherine M. Rondelli[10], Bob Paton[7], Jacqueline Smith[7], David Morrice[7], Laura Daniels[9], Helen G. Tempest[11], Lindsay Robertson[11], Julio S. Masabanda[11], Darren K. Griffin[11], Alain Vignal[12], Valerie Fillon[12], Lina Jacobbson[13], Susanne Kerje[13], Leif Andersson[13], Richard P. M. Crooijmans[8], Jan Aerts[8], Jan J. van der Poel[8], Hans Ellegren[14]

cDNA sequencing: Randolph B. Caldwell[15], Simon J. Hubbard[16], Darren V. Grafham[17], Andrzej M. Kierzek[18], Stuart R. McLaren[17], Ian M. Overton[16], Hiroshi Arakawa[15], Kevin J. Beattie[19], Yuri Bezzubov[15], Paul E. Boardman[16], James K. Bonfield[17], Michael D. R. Croning[17], Robert M. Davies[17], Matthew D. Francis[17], Sean J. Humphray[17], Carol E. Scott[17], Ruth G. Taylor[17], Cheryll Tickle[19], William R. A. Brown[20], Jane Rogers[17], Jean-Marie Buerstedde[15], Stuart A. Wilson[21]

Other sequencing and libraries: Lisa Stubbs[22], Ivan Ovcharenko[22], Laurie Gordon[22], Susan Lucas[23], Marcia M. Miller[24], Hidetoshi Inoko[25], Takashi Shiina[25], Jim Kaufman[26], Jan Salomonsen[27], Karsten Skjoedt[28], Gane Ka-Shu Wong[29,30,31], Jun Wang[29,30], Bin Liu[29], Jian Wang[29,30], Jun Yu[29,30], Huanming Yang[29,30], Mikhail Nefedov[32], Maxim Koriabine[32], Pieter J. deJong[32]

Analysis and annotation: Leo Goodstadt[4], Caleb Webber[4], Nicholas J. Dickens[4], Ivica Letunic[6], Mikita Suyama[6], David Torrents[6], Christian von Mering[6], Evgeny M. Zdobnov[6], Kateryna Makova[2], Anton Nekrutenko[2], Laura Elnitski[2], Pallavi Eswara[2], David C. King[2], Shan Yang[2], Svitlana Tyekucheva[2], Anusha Radakrishnan[2], Robert S. Harris[2], Francesca Chiaromonte[2], James Taylor[2], Jianbin He[2], Monique Rijnkels[33], Sam Griffiths-Jones[17], Abel Ureta-Vidal[3], Michael M. Hoffman[3], Jessica Severin[3], Stephen M. J. Searle[17], Andy S. Law[7], David Speed[7], Dave Waddington[7], Ze Cheng[34], Eray Tuzun[34], Evan Eichler[34], Zhirong Bao[34], Paul Flicek[35], David D. Shteynberg[35], Michael R. Brent[35], Jacqueline M. Bye[17], Elizabeth J. Huckle[17], Sourav Chatterji[36], Colin Dewey[36], Lior Pachter[36], Andrei Kouranov[37], Zissimos Mourelatos[37], Artemis G. Hatzigeorgiou[37], Andrew H. Paterson[38], Robert Ivarie[38], Mikael Brandstrom[14], Erik Axelsson[14], Niclas Backstrom[14], Sofia Berlin[14], Matthew T. Webster[14], Olivier Pourquie[39], Alexandre Reymond[40], Catherine Ucla[40], Stylianos E. Antonarakis[40], Manyuan Long[41], J. J. Emerson[41], Esther Betrán[42], Isabelle Dupanloup[43], Henrik Kaessmann[43], Angie S. Hinrichs[44], Gill Bejerano[44], Terrence S. Furey[44], Rachel A. Harte[44], Brian Raney[44], Adam Siepel[13], W. James Kent[44], David Haussler[44,45], Eduardo Eyras[46], Robert Castelo[46], Josep F. Abril[46], Sergi Castellano[46], Francisco Camara[46], Genis Parra[46], Roderic Guigo[46], Guillaume Bourque[47], Glenn Tesler[48], Pavel A. Pevzner[49], Arian Smit[50]

Project management: Lucinda A. Fulton[1], Elaine R. Mardis[1] & Richard K. Wilson[1]

Affiliations for participants: 1, Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA; 2, Center for Comparative Genomics and Bioinformatics, Departments of Biology, Statistics, Biochemistry and Molecular Biology, Computer Science and Engineering, and Health Evaluation Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 3, EMBL-EBI, Wellcome Trust

Genome Campus, Hinxton, Cambridge CB10 1SD, UK; 4, MRC Functional Genetics Unit, University of Oxford, Department of Human Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, UK; 5, Max-Delbrueck-Center for Molecular Medicine, 13025 Berlin, Robert-Roessle-Strasse 10, Germany; 6, EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany; 7, Genomics and Genetics and Bioinformatics, Roslin Institute (Edinburgh), Midlothian EH25 9PS, UK; 8, Animal Breeding and Genetics Group, Wageningen University, Marijkeweg 40, 6709PG Wageningen, The Netherlands; 9, Department of Animal Science, 2131D Meyer Hall, One Shields Avenue, University of California, Davis, California 95616, USA; 10, Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan 48824, USA; 11, Cell and Chromosome Biology Group, Department of Biological Sciences, Brunel University, Uxbridge, Middlesex UB8 3PH, UK; 12, Laboratoire de Genetique Cellulaire, Centre INRA de Toulouse, BP 27 Auzeville, 31326 Castanet Tolosan, France; 13, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala Biomedical Center, Box 597, SE-751 24 Uppsala, Sweden; 14, Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvegen 18D, SE-752 36 Uppsala, Sweden; 15, Institut fuer Molekulare Strahlenbiologie, GSF–Forschungszentrum, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany; 16, Department of Biomolecular Sciences, UMIST, PO Box 88, Manchester, M60 1QD, UK; 17, The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK; 18, Laboratory of Systems Biology, Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Pawinskiego 5a, 02-106 Warszawa, Poland; 19, Division of Cell & Developmental Biology, School of Life Sciences, University of Dundee, Dundee DD15EH, UK; 20, Institute of Genetics, Nottingham University, Queen's Medical Centre, Nottingham NG7 2UH, UK; 21, Department of Molecular Biology and Biotechnology, University of Sheffield, Firth Court, Western Bank, Sheffield S10 2TN, UK; 22, EEBI Division and Genome Biology Division, Lawrence Livermore National Laboratory, Livermore, California 94550, USA; 23, DOE Joint Genome Institute, Walnut Creek, California 94598, USA; 24, Division of Molecular Biology, Beckman Research Institute, City of Hope National Medical Center, 1450 E. Duarte Road, Duarte, California 91010, USA; 25, Department of Molecular Life Science, Division of Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, 143 Shimokasuya, Isehara 259-1143, Japan; 26, Institute for Animal Health, Compton, Berkshire RG20 7NN, UK; 27, The Royal Veterinary and Agricultural University, Department of Veterinary Pathobiology, Laboratory of Immunology, Stigboejlen 7, Frederiksberg, Copenhagen DK-1870, Denmark; 28, Department of Immunology and Medical Microbiology, University of Odense, Winslovparken 19, Odense, Copenhagen DK-5000, Denmark; 29, Beijing Institute of Genomics of Chinese Academy of Sciences, Beijing Genomics Institute, Beijing Proteomics Institute, Beijing 101300, China; 30, James D. Watson Institute of Genome Sciences of Zhejiang University, Hangzhou Genomics Institute, Key Laboratory of Bioinformatics of Zhejiang Province, Hangzhou 310007, China; 31, Genome Center, Department of Medicine, University of Washington, Seattle, Washington 98195, USA; 32, Children's Hospital Oakland Research Institute, 747 52nd Street, Oakland, California 94609, USA; 33, Children's Nutrition Research Center, Baylor College of Medicine, 1100 Bates Street, Houston, Texas 77030-2600, USA; 34, Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA; 35, Laboratory for Computational Genomics, Campus Box 1045, Washington University, St Louis, Missouri 63130, USA; 36, Departments of Computer Science and Mathematics, U.C. Berkeley, Berkeley, California 94720-3840, USA; 37, Center for Bioinformatics, Departments of Genetics and Pathology, University of Pennsylvania, Medical School, Philadelphia, Pennsylvania 19104-6021, USA; 38, Plant Genome Mapping Laboratory and Department of Genetics, University of Georgia, Athens, Georgia 30602, USA; 39, Stowers Institute for Medical Research, 1000 East 50th Street, Kansas City, Missouri 64110, USA; 40, Department of Genetic Medicine and Development, University of Geneva Medical School, 1 rue Michel-Servet, 1211 Geneva, Switzerland; 41, Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA; 42, Department of Biology, University of Texas at Arlington, Arlington, Texas 76019, USA; 43, Center for Integrative Genomics, BEP, University of Lausanne, CH-1015 Lausanne, Switzerland; 44, UCSC Genome Bioinformatics Group, Center for Biomolecular Science & Engineering, Mailstop SOE, Baskin School of Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA; 45, Howard Hughes Medical Institute, Center for Biomolecular Science & Engineering, Mailstop SOE, Baskin School of Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA; 46, Grup de Recerca en Informatica Biomedica, Institut Municipal d'Investigacio Medica, Universitat Pompeu Fabra, and Programa de Bioinfomatica i Genomica, Centre de Regulacio Genomica, C/Dr. Aiguader 80, 08003 Barcelona, Catalonia, Spain; 47, Genome Institute of Singapore, 60 Biopolis Street, 02-01 Genome, 138672, Singapore; 48, University of California, San Diego, Department of Mathematics, 9500 Gilman Drive, La Jolla, California 92093-0112, USA; 49, University of California, San Diego, Department of Computer Science and Engineering, 9500 Gilman Drive, La Jolla, California 92093-0114, USA; 50, Computational Biology Group, The Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103, USA