Supplementary Methods and Information

Methods – Sequencing and assembly

Source DNA

DNA from a single female of the inbred UCD001 line of Red Jungle fowl was used to prepare all sequencing libraries, as well as several BAC libraries used for physical mapping[1]. A single inbred individual was chosen to minimize internal genetic polymorphism, and a female was used to provide sequences of both the Z and W chromosomes. The UCD001 line is fully interfertile with domestic chickens, and it was previously employed in a cross to inbred UCD003 White Leghorn birds to generate one of two widely shared families for genetic linkage map development [2]. The UCD001 RJF line was initiated in 1956 by brother x sister matings at the University of California-Davis from a stock at Cornell University that originated from a population of birds maintained at a zoo in Hawaii[3]. The zoo stock was established from birds from Malaysia. As with most captive RJF populations, it is probable that, during its breeding history, the UCD001 genome was contaminated with a limited amount of domesticated chicken germplasm. However, similar to wild RJF, UCD001 females lay small brown eggs on a seasonal basis and are small birds with brown feathering and shy behavior. The males have red and black feathering; the extent to which UCD001 males possess the eclipse plumage pattern, a hallmark of "true" RJF (i.e., during molting, the first set of replacement feathers are female-style and these are later replaced with male plumage) is unknown, as the line is typically maintained using artificial light conditions. RJF males are known for their fighting characteristics and the UCD001 males are also very aggressive.

Sequencing

Sequence data was obtained from paired-end plasmid subclone reads. Greater than 90% of the time high quality data were obtained from both ends of each plasmid subclone. Libraries were constructed using inserts of various sizes ligated into the pOT vector. A highly automated production pipeline consisting of a 384 well format ensured the integrity of the paired-end data[4]. For more information on specific protocol information, please visit the Genome Sequencing Center web site at http://genome.wustl.edu/ .

Whole genome shotgun assembly

A whole genome shotgun assembly of chicken was performed using PCAP [5], a parallel algorithm that takes advantage of both read-pairing constraint information and base quality values during sequence assembly. The assembly (main paper Table 1) was based on 6.6X phred 20[6] coverage (72% of total input bases were of phred 20) of the chicken genome (assuming a 1.06Gb genome; 9,762,824,258 bases in 11,241,008 reads, 4.4% unused); stringent parameters were used to avoid potential global assembly errors. Genetic Marker data were assigned for those markers with sequence data available via WU-BLAST (Warren R. Gish, unpublished, http://blast.wustl.edu).

Integration of Assembly and Fingerprint map

BAC-end sequences assembled in WGS assembly contigs were used to position fingerprint contigs from the chicken physical map (Wallis et al., this issue) "onto" the assembly and vice versa (requiring links of at least 6 BAC ends to define a fingerprint and assembly contig as linked). The fingerprint contig was assigned to a sequence contig

with the majority of end sequences. Further, we performed in silico digests for sequence contigs longer than BAC length and compared the *in silico* restriction map with the fingerprint map to further place the assembly onto the map.

All possible discrepancies between the assembly and fingerprint map were reviewed with the "problems" identified as follows: the resolution of the fingerprint-assembly comparison is a window of 10 BAC-end links. If the ends are distributed randomly, that is roughly 60 kb. Scanning the assembly, if there were >= 6 links within the window to fingerprint contigs that are not 'assigned' to the supercontig, that region was flagged for manual review. An FPC contig is not assigned to a supercontig unless there are a minimum of 6 links between them and it does not introduce a topological problem.

All possible merges in the fingerprint map as predicted by the sequence assembly were manually reviewed. Further all regions where the fingerprint map disagreed with the assembly were manually reviewed both in the fingerprint map and WGS assembly to check for possible discrepancies. Further discrepancies suggested by genetic map locations of assigned markers, such as a single supercontig being assigned to multiple chromosomes, were manually reviewed. After manual review of physical and genetic map discrepancies, no mis-assemblies in the WGS assembly were found. After any necessary changes were made to the fingerprint map, the fingerprint map was used to create "ultracontigs", clusters of multiple supercontigs, from the assembled "supercontigs". Supercontigs were ordered and oriented into ultracontigs using BAC end placement and marker information. A total of 772 Mbp of actual sequence of supercontigs (not including gaps between contigs) was integrated into 84 longer

"ultracontigs" (average non-gap basepair length of 10Mb per ultracontig). Of the 84 ultracontigs, 80 are localized to specific chromosomes or linkage groups. Additionally, 490 supercontigs not linked to the physical map were also localized to specific chromosomes or linkage groups. In the final sequence, there are a total of 11,000 ultra/supercontigs larger than 2kb that total 1.055Gb with 933Mb (actual bp not including gaps or 968Mb including estimated gap sizes), or almost 90% of the sequence anchored to a specific chromosome or linkage group. There is an additional 121Mb of non-N basepairs on chrUn (165Mb including estimated gap sizes).

Creation of ordered/oriented supercontigs along chromosomes or "AGP files"

We created chromosomal sequences as possible from the underlying sequence data. Genetic markers were assigned to both the fingerprint (Wallis et al., this issue) and the WGS maps. Initially, all ultracontigs and supercontigs anchored to the physical map were ordered and oriented along the chicken chromosomes using the start, end, and median positions based on the marker data. When possible location discrepancies of marker positions within ultracontigs were found, they were manually resolved. After initial anchoring via the physical map, other supercontigs not anchored to the physical map but assigned via BLAST assignment of the markers were integrated into the list of ordered and oriented chromosomal locations when order was clear based on the genetic map. In some cases, some read pairing information provided linking information which aided in verification of supercontig order. For supercontigs assigned to a chromosome, but where order was not clear, the supercontigs were assigned to the random portion of the chromosome. For those supercontigs where no marker information unambiguously

localized the sequence, they were placed onto the unlocalized chromosome (chrUn). In some cases where supercontig order was clear based on marker content, orientation was not clear. For example, if only a single marker was available or a pair of unordered markers was available, orientation could not be certain. In those cases, read pair information was checked for help in orientation.

Following creation of the initial chicken chromosome files, several steps were taken to improve this initial assembly. For example, an attempt was made to localize centromeres within the chromosome files primarily using marker boundaries (FISH data) and the presence of CNM repeats (GGA23 and GGA28 specifically). Further, certain specific regions of the chromosome assemblies were also improved by other data. For example, knowledge of the order and orientation of the alpha and beta-globin complexes [7,8] were used to manually re-evaluate all underlying read pair data through that region of the assembly. Similarly, matches to other chicken EST and mRNAs were used to help in integrating small sequence supercontigs (usually ~1kb) usually from the unlocalized chromosome into their correct place in the assembly (~200 changes were manually introduced using mRNA information) or relocating contigs nearer each other along chrUn. Some linking information from the underlying read pair data was required (where, for example, only a single read pair suggested connection to the neighboring contig which normally is not enough overlap to accept), however, before moving any supercontig, and when underlying data were ambiguous, the change was not made. A majority of these suspected rearrangements will require additional sequencing to allow correct ordering and orientation of adjacent contigs. Comparison to finished clones from

the Rfp-Y (or Y) and B-complex regions were used not to order and orient but to identify additional supercontigs for GGA16.

Intensive efforts were applied to discriminate Z and W contigs and to localize further data to the W chromosome. Unfortunately, the W chromosome only had six markers placed on the genetic map at the time of assembly. However, additional W-specific genes had been identified (some where both a W-specific and Z-specific version had been sequenced) and those were used to aid in final supercontig placement (*AD012, CHD1, ATP5A1, PKCI, SPIN, EE0.6, MADH2,* and *FET1*). Further, W-specific repeats were identified in several large supercontigs and those were placed onto the W chromosome. Unfortunately, those were later to be determined to not be W-specific repeats (see below). Finally, the sequence presented at the UCSC browser (http://genome.ucsc.edu) for chrM (the mitochondrion, complete genome) was extracted from GenBank entry (gi|5834843|ref|NC_001323.1|, *Gallus gallus*, White Leghorn) and is not the result of this assembly project.

Localization of centromeres and telomeres

Cloning, sequencing and assembly of repetitive elements organized as tandem arrays are problematic with the net effect that the draft and "finished" genome sequences are found to be diminished in sequence content of such regions, e.g., centromeric and telomeric repeats and also coding regions such as the 5S and 18S-5.8S-28S rDNAs (see ncRNA section). For example, although centromere positions are indicated on the chicken draft sequence, little is known of their exact sequence. In only two cases were centromere assignments made by sequence identification, utilizing the CNM repeats [9] on GGA23 and

28. The centromeres of an additional 14 chromosomes were tentatively localized based on FISH hybridization using BAC clones, genetic markers flanking the centromeres in coordination with mapping gaps in the physical map, and analysis of proximity to the constrictions of the mitotic metaphase chromosomes. Macrochromosome centromere sizes were assigned arbitrarily to be 1.5 Mb and those of microchromosomes to be 0.5 Mb lengths in absence of any evidence as to their true lengths.

Despite these limitations, evidence for interstitial and terminal telomere repeat sequence (TTAGGG) was found in the chicken draft sequence. Fifteen sites containing telomeric repeats were identified within GGA 1-4 and Z, none were found on GGA 5. Table S8 summarizes the characteristics of these telomeric repeat sequence locations. An additional 11 sequences containing telomeric repeats were found on unplaced sequence. In all but one case, degenerate telomeric sequence was found adjacent to the repeats; degenerate telomere sequences (e.g., TTTGGG) are typical of regions adjacent to telomeres in the human genome but were not previously identified as a characteristic of chicken. A majority (15 of 26) of the sites (localized or unlocalized) were adjacent to contig gaps suggesting that the identified sequences could be at the edges of larger interstitial or terminal telomeric repeats. The interstitial telomeric repeats did not occur preferentially at the boundaries of conserved synteny blocks (*data not shown*). Repeats were found within 2 Mb of the "ends" of the assembled sequence of GGA 1, 3 and 4 (Table S8) suggestive that the terminal telomeres are resident in these regions. GGA 2 exhibited repeats at 1 Mb and 138.6 Mb (148 total sequence); notably, GGA 2 possesses a q-arm interstitial telomere by FISH and was found to have several telomere repeats

identified at ~88Mb during the initial screen although these were all within genes (*data not shown*).

GGA1 was used as a case study to examine the alignment of the chromosome map telomere positions with the draft sequence locations of telomeric DNA. GGA1 possesses five telomeric DNA blocks identified by FISH: p- and q-arm terminal locations and three interstitial sites including p-terminal-adjacent, p-centromere-adjacent, and q-interstitial (Fig. S16). Although speculative, it is interesting to consider the correspondence between the sequence and chromosome map, and whether the telomeric sequence locations found for in GGA1 in fact corresponds to some of the known cytogenetic locations: the 15 Mb repeat reflecting the q-terminal-adjacent site, the 40 Mb repeat reflecting the p-centromere-adjacent site, the 145 and/or 165 Mb repeats reflecting the q-interstitial site.

Assembly - coverage

We assessed the coverage of the current chicken assembly in several ways described in the main text. We also compared our assembly to 8.5Mb of "improved" draft sequence[10] generated in a BAC by BAC fashion (with underlying contigs ordered and oriented) from three regions of the genome that correlate primarily to regions orthologous to human chromosome 19. Those three regions represent areas of varying G+C content, and accordingly our assembly shows differing levels of coverage dependent on the G+C content. The 3.8Mb region with an overall %G+C content of 40% was covered at a level of over 98% by our assembly. A 1.1 Mb region with higher overall G+C content (49%) revealed 88% coverage and the 3.6Mb region at 52% G+C content had 82% coverage.

Assessment of individual 100kb windows along these larger segments revealed that in the most GC-rich regions, both assemblies tended to be more fragmentary (more gaps and shorter contigs in both the BAC by BAC assembly and in the WGS assembly). In general, there tended to be a higher percent discrepancy between the two assemblies in some, but not all, of the most GC-rich regions.

Comparison to clones sequenced from GGA16, the microchromosome that contains three large gene families, the nucleolar organiser region (NOR) encoding highly repetitive rRNA genes, the B locus containing the GC-rich major histocompatibility complex (MHC), and the Y locus including non-classical MHC genes [11-13], revealed two distinct patterns. Comparison of our assembly to clones in the MHC region (Marcia Miller, Shiina Takashi, personal communication) reveal coverage of two BAC clones (113kb and 139kb) from the Rfp-Y region ("Y") (55% GC) at 80% coverage, and one 123kb clone (52%GC) from the B-complex region at 86% coverage. In the B-complex region the WGS assembly was organized into one 84kb supercontig and two other smaller supercontigs all with good agreement in order and orientation to sequenced BAC clones for the region. However, in the two clones from the Y region, the WGS assembly contained short contigs (2kb on average) and read pair information was often conflicting. Comparison of the assembled contigs of the WGS sequence with several sequenced cosmid contigs from several B haplotypes (95 kb of the BF/BL region containing the classical MHC, AL023516 [11]; 20kb of the Y locus, AJ277927 [14]; 150 kb of the BG region, Salomonsen, unpublished) showed that some genes are faithfully represented, but in other cases only partial genes are present (TAP and CD1) or the homologous genes known to be present are represented by fewer genes. Since complete gene sequences for

the missing genes are not yet available from BACs or other libraries, it is not clear why they present problems. However, analysis of the sequence of PCR-generated fragments of *BLA* detects many subtelomeric repeats, suggesting a reason for its lack of representation. For the MHC cosmid (AL023516 [11]), G+C analyses revealed 60% G+C content with portions of many genes exhibiting higher G+C content, particularly at their 5' end.

Representation through these regions is complicated, therefore, by high percent G+C and also by complex repetitive gene structure. Because there was a reduction in coverage in regions of high G+C content, we assessed the G+C content of the total set of chicken WGS reads versus that of the unplaced reads. These analyses revealed the highest peak at 51% G+C content for the unplaced reads as compared to 38% for the set of all WGS chicken reads. The chicken genomic assembly does, however, contain 500 100kb windows (5% of the 100kb windows in the genome) with G+C content of greater than 50%, indicating that not all of these regions have been left out due to cloning or other issues.

We also examined the lists of possible "missing" genes based on the list of chicken mRNAs that did not find a match in the chicken genomic assembly and on the set of missing orthologs (see main text). For example, of missing chicken mRNAs, 53% have GC content higher than 60%, whereas only 12% of all chicken mRNAs share this property. As another example, simple sequence repeat (SSR) content of the non-exonic portion of unplaced reads that align with some of the missing mRNAs is much higher (5.46%) than the SSR content of the introns from genomic ENSEMBL predictions (0.38%). Of the rare unidentified chicken mRNAs that are genetically mapped, 30% map

to the microchromosomes and 70% to the macrochromosomes, almost identical to the ratio of markers placed on the genetic map, 33% and 67%, respectively.

We examined a subset of 400 of the "missing 5-10% of chicken genes" described in the protein section (see main text, assembly discussion and below). Of these 400 proteins, 72% had some hit against the chicken genome using tblastn, and 50% could be found using the representative EST placed against the chicken genome. While some were only partially found in the assembly or in the set of singleton reads, some were actually completely found within the assembly using methods that varied from the methods used by the Ensembl pipeline, and were simply failures of the gene building process. Further, we obtained intronic sequence for 16 of these missing genes by PCR against the genome. The average GC content for these regions was 64%. Interestingly, 21% of the "missing" genes are associated with genes on HSA19, an extreme outlier in terms of GC content in the human genome. Thus in general, missing genes have higher GC content than average and many, including some HSA19 orthologues, are associated with intronic simple sequence repeats.

Assembly validation - comparison of physical and genetic map and comparison to new marker data

Marker sequences were identified in the chicken sequence and compared to their locations on the genetic maps. Marker data obtained from the mapping consortium after release of the chicken assembly showed that only 6 Mb of sequence were assigned to the wrong chromosome. Additionally, 12 Mb of sequence was moved from the

"unlocalized" chromosome to another chromosome. Further, 34Mb of sequence was moved from the "random bin" of an individual chromosome to an anchored position on the same chromosome. These changes will be incorporated in later releases of the chicken genome.

Assembly validation - W chromosome

A large portion of the chromosome was assigned to W based on the presence of proposed W-specific repeats that, as has now been shown (Hans Ellegren, personal communication), are not truly W-specific. Thus, the only portions of GGAW that should currently be considered specific to W are: W: 1-195831; W: 4895452-4916845; W_random: all. Since the assembly, new mapping data (Martien Groenen, personal communication) allow additional fingerprint contigs to be localized to GGAW, but links will only allow us to relocate one supercontig currently on chromosome Un (contig514).

Assembly validation - comparative analysis

When comparing the chicken and human genome sequences, we have discovered several differences where additional sequence data will be required before it can be determined whether these are true evolutionary rearrangements or assembly problems in the sequence. Only a single erroneous inversion in the sequence has been confirmed to this point: the chicken assembly should have been reversed in order to agree with the human suggested order (chr1: 145299067-> 149609814, 149619815 ->154006799; should be 149619815 ->154006799: 145299067-> 149609814). In addition, we provide below some of the ambiguous examples:

a) Bases 76M-90M of human chromosome 16 align with a 14-19 Mb segment of chicken chromosome 11. This particular region contains one of the highest densities of inversions we have found. Here, we have again utilized the underlying sequence assembly, the marker data, the BAC fingerprint data, and alignments to the human sequence in an attempt to refine contig order and orientation. While one supercontig in the region could potentially be flipped, the read pair data are not unambiguous. Likewise, the fingerprint data do not provide additional clues. The sequence assembly is well supported by read pair data, and a misassembly here is not likely. Therefore, without generating additional data specifically aimed at closing the existing gaps, we currently are unable to resolve this region.

b) A region of chicken chromosome 20 contains five supercontigs assembled in the following order (s1: 10,051,000-10,060,000; s2: 10,060,000-10,070,000; s3: 10,070,000-10,095,000; s4: 10,095,000-10,106,000; s5: 10,106,000-10,115,000). An alignment of this region with human chromosome 20 suggests that the order should be s4, s1, s5, s2 (same orientation, and with the position of s3 undetermined). However, an examination of other data left us unable to accurately place s3, although we did find that potentially interweaving two supercontigs would result in an order that is more similar (but still not identical) to the human sequence through this region than in our original assembly.

c) Alignments spanning the myosin heavy chain region in human/chicken are also of interest. There is a 1.8 Mb region containing several genes of interest in subregions 1-5 :

1. 1-650,000 MyHC ;

2. 650,000-740,000 MAP2K4;

3. 740,000-990,000 MYCD;

4. 990,000-1,180,000 DNAH9;

5. 1,180,000-1,700,000 AK127379

Alignment with the human genome sequence predicts the order 1 5’ 4 2 3’ (with primes denoting reversal). At a finer scale, alignments predict the reversal of a fragment of approximately 50 kb with sub-region 3. In this case, after review of the underlying data, the physical map strongly supported our initial assembly through the region. There are some remaining questions as to the precise order of marker data, however, the overall placement is not far out of range. This particular region in our whole genome assembly is spanned by a single supercontig that has good supporting read pair information. Thus, by our usual criteria, the current assembly is acceptable. Assembling this region using alternate assembly algorithms has also not conclusively determined absolute order through the region. As previously stated, the availability of additional linkage markers and/or sequencing data would help to resolve the order and orientation of this region in the current chicken genome sequence. Specifically for the MyHC region, SNP markers (David Burt, personal communication) were created to determine the order through this region. Currently on the East Lansing map, the (low resolution) order is: MYE1-0.0-[MYHC, MAP2K4, MYCD, DNAH9 order unknown]-2.0-COM0049-6.1-AK127379. So we can say that AK127379 is not WITHIN this group but is outside. Crosses with n=500 are underway to determine a more accurate order.

d) During comparative analyses, regions were identified where there was conservation in mouse/rat/human and yet a difference in the chicken genome (G. Tesler and P. Pevzner, personal communication). We assessed nine of these regions of micro-rearrangement using underlying read pair and read depth data, correlation with the

physical map, mRNA content, and reassembly of the regions using other assembly algorithms. A majority of the regions (7) were localized to a single "supercontig" and as such were supported by underlying read pair data. Using all of the above methods for assessment, only one region showed a possible assembly problem based on some contradicting read pairing information, but even in that case, the single possible contradiction would not explain the entire rearrangement.

Assembly validation - gene detection and annotation

Analysis of the current assembly with the goal of creating an index of all chicken genes has been challenging. For example, in the current draft for any particular gene, one occasionally encounters the following scenarios: 1) Individual exons from the same gene localized to multiple regions of the genome. For example, the gene Hox-B9 (P17482) currently has three of its exons on two contigs localized to chromosome 2, another copy of one of those exons on a chromosome 27 contig, with at least two other exons missing. Obviously, this type of example will result in genes being missed or at least incomplete. In some cases, these could arise from paralogous duplication events. 2) Stretched/long introns. For example, the gene Q8N6G6 is mainly on chromosome Z, with three "tight" islands of exon structure (one of two exons, one of three exons, one of four exons) separated by introns of 10,000 bp and 5,000 bp. The 5' end of the gene appears to be copied on GGA10 (and/or there is a complicated paralog). This type of problem likely will lead to the islands being annotated as separate genes. 3) Complex duplications/paralogous structure. For example, it appears that the gene P50238 may be present in three copies within the same region on chromosome 8. Within this region,

exons are either missing or misplaced, leading to an inability to accurately reconstruct the intron-exon structure.

These examples represent the predominant errors. Largely, they are due to a ~6X draft with a significant number of remaining sequence gaps; this leads to "drop outs" that must be identified by other methods. Along these lines, some of the HOXC and HOXD cluster genes are not in the current assembly. While many of the human genes for this region are high in G+C (ranging from 51% to 69%), the overall G+C content of the regions in the human genome is not as high (36% GC for HOXC and 46% GC for HOXB regions as a whole).

A number of other interesting examples of specific genes that are missing or incomplete in the current draft sequence have been detected. One such example is the VKORC1 gene, which encodes the vitamin K epoxide reductase protein that recycles vitamin K. The VKORC1 gene is present in all available mammalian genomes, as well as the Fugu and zebrafish genomes, yet was apparently absent from the assembled chicken genome sequence. A TBLASTN search of the 440,000 unassembled reads revealed a single read that contains most of the second of three expected exons. A closer look suggests that the VKORC1 gene lies in a region of the genome that is underrepresented in the current assembly, perhaps due to cloning problems or simply for statistical reasons. In contrast, the paralog of this gene - VKORC1L1 - is present in the assembled genome. Another gene, SOX21, is only partially found (261 bp of similarity to an unplaced read), but this gene shows a GC content of 76%. Similarly, the PMEL1 gene is only partially found. This gene is 68% GC in content and has been difficult to sequence because of presumably polymorphic insertion/deletions.

Methods - Gene content of the chicken genome

Non-coding RNA Methods

Non-coding RNA genes were predicted using a variety of computational methods, including pairwise similarity to known ncRNAs, covariance model searches, and specific gene family finding algorithms. tRNA genes were predicted using tRNAscan-SE 1.23 in eukaryotic mode with the default threshold of 20 bits [15], in both the assembled genome sequence and all unplaced sequence reads. We searched the genome sequence for 18S and 26S ribosomal RNAs using WUBLASTN (2.0MP-WashU 01-Mar-2004) with a wordsize of 3, an E-value threshold of 0.01 and with sum statistics turned off (the -kap option) (http://blast.wustl.edu/). We collected a non-redundant set of 236 human, mouse and rat miRNA sequences from the microRNA Registry 3.1 (http://www.sanger.ac.uk/Software/Rfam/mirna/) [16]. Candidate chicken miRNAs match a mammalian precursor miRNA with E-value less than 10(-4) (using WUBLASTN), form a predicted hairpin structure with free energy of folding less than -20 kJ/mol (using RNAfold from the ViennaRNA package) [17], and match the mature ~22 nt miRNA with 2 or fewer mismatches. Chicken homologs of a previously annotated set of 245 human snoRNAs (Jones T and Eddy SR, ftp://ftp.genetics.wustl.edu/pub/eddy/annotation/human-hg16/) were detected using WUBLASTN as above, with an E-value threshold of 10(-4). Other ncRNAs were predicted using the Rfam 6.0 library of covariance models (http://www.sanger.ac.uk/Software/Rfam/) [18] and the INFERNAL 0.55 software suite [19].

Protein coding gene methods

TWINSCAN

TWINSCAN [20] recognizes statistical patterns characteristic of coding sequences, splice signals, and other features in the genome to be annotated to assign probability scores to each potential exon. TWINSCAN employs an independently developed extension of the GENSCAN [21] probability model in which the probability score assigned to each potential exon is modified by the presence and quality of genome alignments. TWINSCAN uses nucleotide alignment and has specific models for how alignments modify the scores of coding regions, UTRs, splice sites, and translation initiation and termination signals. For this study TWINSCAN 1.6 was used with specialized parameters trained from a set of 525 cDNA confirmed chicken genes mapped to the galGal2 assembly. The training set is based on a set of 1266 "provisional refseq" mRNA sequences that were downloaded from Genbank on March 27, 2004 and strictly filtered. Those mRNA sequences that were not placed on the galGal2 assembly were removed. Any sequences without an ungapped alignment in the coding regions and genes with in-frame stop codons and non-canonical splice sites were also removed. The training set is available at http://genes.cs.wustl.edu/private/chicken_refseqs/. TWINSCAN's alignments covered 3.8% of the chicken assembly and were created with WU-BLAST (http://blast.wustl.edu) using M=1 N=-1 Q=5 R=1 W=10 X=30 S=30 gapS2=30 filter=seg filter=dust against a human genome NCBI34 assembly database created as previously described [22] . TWINSCAN is open source software and can be run through a web interface or downloaded at http://genes.cs.wustl.edu.

SGP2

SGP2 training followed a hybrid approach. SGP2 was used with human parameters, whereas the score weights to reward the Human-Chicken homologies and penalize the lack of them were optimized using the same curated set of 525 Chicken RefSeqs used for Twinscan. SGP2 was then run on unsegmented chicken chromosomes using the TBLASTX alignments with the human genomic sequence (assembly NCBI34). These alignments, which covered approximately 3% of the chicken genome, were enriched with 391,610 extra HSPs obtained from the ungapped Exonerate (G. Slater, unpublished) alignments of human proteins from Ensembl (release NCBI34c), the Geneid prediction set for the same human assembly and the set of vertebrate RefSeq proteins (version of April 2004). The extra alignments covered 43% of the nucleotides in TBLASTX HSPs and 5% of their sequence represented 5840 non-redundant homology regions that had not overlap with the TBLASTX hits. The extra alignments produced a considerable improvement of the sensitivity and specificity at the gene level with respect to SGP2 predictions using only TBLASTX HSPs when tested against the Ensembl set and the aforementioned 525 RefSeqs. It also achieved a slight improvement of the sensitivity at the exon and nucleotide level.

ENSEMBL

The chicken Ensembl system employed the Ensembl system designed for mammalian systems [23] adapted for the chicken genome. We used the standard targeted system to place known genes (where both a protein and cDNA have been submitted to the database) on the chicken genome. As we expected more genes via protein similarity we relaxed the protein cutoff to 150 bits. The EST and cDNA information were processed through the EST based pipeline, which aligns cDNAs and ESTs to the genome, merges the resulting

partial transcript structures and then finds the longest open reading frame through the resulting set of exons. A selected set of cDNA and EST based transcripts were merged with the protein similarity if the cDNA/EST transcripts if the cDNA or EST transcript contributed more than 2 unique exons to an existing gene structure or at least a spliced structure if the cDNA/EST transcript lay outside of any protein similarity defined structure. We rejected single exon EST and cDNA transcripts where the only evidence was transcript based, the assessment being that the majority were genomic contaminations, although obviously a significant number will be real transcripts. However, in the absence of protein similarity, there is no easy way to distinguish cloning artifact from true single exon genes. We then performed a round of triage against missing mammalian orthologs using the Exonerate protein2genome (G.Slater, unpublished) system. When testing the sensitivity of Ensembl, we used only the confident cDNA set as our reference and tested just the protein similarity set as the exemplar of the Ensembl system. Given that additional cDNA and EST evidence is also used in the final gene set, one would expect the sensitivity to be higher.

The Ensembl chicken gene set consists of 28,416 transcripts in 17,709 genes (1.6 transcripts per gene). There are 5,281 genes (~30%) that have multiple transcripts with 4,883 (92%) represented by 5 or fewer alternative transcripts.

Estimating gene number

As was done with the mouse genome, we estimated the gene number by first estimating the total number of exons in the genome and then dividing by the average number of exons per gene. To estimate the total number of exons, we stratified our exon predictions into 5 sets, being the 3-way intersection of the prediction set, the 2-way intersection and

then the unique Ensembl set. The unique Twinscan and SGP-2 set was not used because we expected low specificity which, even in best-case scenarios, would be hard to measure. The specificity of these sets was measured by RT-PCR experiments (see the companion paper by Eyras et al), and the sensitivity assessed by using a confident set of cDNAs. The resulting sensitivity/specificity analysis is given in Table S9. The variation in predicted exon total is large; this is somewhat to be expected as this estimation assumes random behaviour of all three groups, in particular with respect to the reference set of cDNAs. A notable outlier is the (T&E)!S, which has a very low level of exons compared to the other sets, but seems to find relatively many reference cDNA exons. Despite these issues, when taking the weighted (by exon number) average of these estimates we arrive at a total number of estimated coding exons of 183,812. Using a coding exon per locus number between 9.6 and 8.0, one gets a gene number between 20,000 to 23,000 (rounding the nearest 1,000). This is a similar number to the straightforward extrapolation of the Ensembl sensitivity and specificity numbers. There are many potentially errors in this estimation, in particular the assumption of randomness between the stratified exon set and the reference set of exons (i.e., partial cDNAs will bias the numbers towards central core exons, which will also be biased towards the intersection sets. However, the total exon/gene number is calculated using longer cDNAs which represent edge exons as well as core exons).

Selenocysteine-containing genes

Selenoproteins are encoded by a small group of genes represented among 20 families in metazoans that have not yet been identified in either yeast or higher plants. They encode selenocysteine, the 21st amino acid, by usurping a specific TGA codon through the

presence of a downstream stem-loop structure in the mRNA. As TGA usually codes for translation termination, these proteins present a challenge for nearly all computational gene finders, which rely heavily on the absence of stop codons. There are two ways to find selenocysteine proteins: by homology to existing proteins or via ab initio techniques that look for the stem-loop structure in the mRNA. Selenocysteines are not perfectly conserved during evolution but rather can be substituted with standard cysteine codons. In a particular family of selenocysteine proteins there will be at least one instance of a selenocysteine at a particular position. Using adapted homology rules, we predicted 33 chicken selenocysteine proteins organized into 19 families, of which 18 contained selenocysteines and one contained a normal cysteine codon at the homologous selenocysteine position.

Methods - Interspersed repeat content of the chicken genome

Identification of repeats

Building on an existing set of reconstructed transposable elements from the chicken genome[Thomas, 2003 #25], we constructed a database of interspersed repeat consensus sequences with the aid of the program RECON[24]. After masking the previously known repeats with RepeatMasker (A.F.A. Smit, R. Hubley, & P. Green, http://www.repeatmasker.org), the sequences were subjected to an all-vs-all pairwise comparison using WUBLAST (W. R. Gish, unpublished, http://blast.wustl.edu), with options "-kap E=0.00001 wordmask=dust wordmask=seg maskextra=20 –hspmax 5000 M=5 N=-11 Q=22 R=11". The resulting pairwise alignments were subjected to RECON with default options. For families with 10 or more copies, a simple majority-rule

consensus was constructed as previously described[24]. The whole genome was analyzed in three iterations, with 5% randomly selected, 30% randomly selected and 100% of the total sequences, respectively. For the second and third iterations, consensus sequences defined in the previous iteration(s) were used to mask the selected genomic sequence using RepeatMasker, in addition to the masking of the known repeats. A total of 983 repeat sequences were thus identified.

RECON-derived repeat libraries contain close to full-length consensus sequences for transposable elements, but mostly fragments, rearrangements or hybrids of such, and also sequences representing gene families or other genomically functional repetitive DNA. Thus, further analysis is needed before RECON libraries can be used for repeat analysis. For this, those repeat sequences represented by more than 300 copies or with strong BLASTX similarities to known transposable element protein products were further characterized. Consensus sequences were improved and extended until they comprised complete elements, usually defined by the presence of flanking target site duplications in the genome. When a single RECON repeat clearly represented several distinct groups, we derived subfamily consensus sequences based on multiple shared ("diagnostic") substitutions and deletions.

Since the great majority of interspersed repeats are formed by CR1 copies, which were partially or completely masked by the few previously derived CR1 consensus sequences, we employed a second strategy to create a representative set of CR1 subfamilies. Detailed subfamily consensus sequences were derived for the abundant and conserved CR1 3' ends and each was extended towards the 5' end of the CR1 element as far as possible. More 5' fragments of the CR1 subfamilies thus derived often had been

recognized in the RECON analysis as lower copy number repeats, a result of the faster evolution of these region compared to the conserved CR1 3' ends.

The resulting database of classified repetitive elements, used for repeat analysis of the chicken genome with the March 2004 version of RepeatMasker contained 89 different consensus sequences comprising 189 kb of DNA. These covered about 180 of the RECON derived repeats, and included 4 DNA transposons, 22 CR1s, 53 LTR elements, and 10 satellites. The consensus sequences have been submitted to RepBase Update (http://www.girinst.org).

Methods - Evolution of the protein content of chicken and mammalian genomes

Domain matching and ranking

To identify known families of genes and domains we scanned respective proteomes for characteristic HMM profile signatures from Pfam [25] and SMART[26] databases using HMMER (http://hmmer.wustl.edu/) software and applied corresponding family specific cut-offs. The identified families were ranked by the number of matching genes requiring at least one matching transcript and counting once repetitive matches.

Orthology detection

Orthologous relations 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 employed earlier [27-29] . 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 step-wise, until all paralogous proteins were joined – 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 step-wise, 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.

Estimate of genes missing from Ensembl chicken gene set

We estimated the number of genes missing from the current Ensembl chicken gene set. The number of represented genes is sensitive to artifacts introduced by gaps in sequencing coverage, assembly or gene prediction: for the current state of the chicken gene set, we estimate that roughly 5-10% of genes are actually missing, or at least substantially truncated. This estimate is based on an analysis of 3555 widely-conserved genes which are found as single copy orthologs in each of five diverse genomes (from human, mouse, rat, fugu, and fruit fly), and which are therefore expected to be additionally present in chicken. Of these, 546 were not detected as orthologs in the chicken assembly, a fraction (~15%) that is larger than for other completed vertebrate genomes (2-7%). Only a minority of these absent genes reflect true biological losses

however because at least 317 genes appear to be represented among EST data (identified using bidirectional hits to human sequences, with scores > 200 bits).

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 following TBLASTN searches in mammalian genomes (thus hinting that several as yet unannotated genes in mammals remain to be predicted).

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 10 (brain; fat & skin; bone & connective; heart; kidney & adrenal; immune; liver; female reproduction; alimentary; testis) distinct tissue types. Percentage amino acid sequence identities of 1:1 chicken-human orthologs were calculated as previously (Fig. 6). Note that single genes may be assigned to multiple tissues.

Duplications

To detect duplicate genes within the EnsEMBL proteomes of chicken and human, homology searches of translations of each annotated transcript against all other transcripts were performed with FASTA [30]. Only hits with greater than 50% amino acid identity and more than 75% reciprocal alignment length were considered. Pairs were considered where each was the best hit in terms of number of aligned amino acids for the other (reciprocal best hits).

Comparative Genomics

Unless stated otherwise, alignments or reports of comparisons to other genomes were done using the following versions of the genome databases: human, July 2003 (hg16); mouse, Feb. 2003 (mm3); rat, Jun. 2003 (rn3); chicken, Feb. 2004 (galGal2).

Methods for Fig. 5

The following gene sets were used to support the figure: Chicken – Ensembl v22.1.1 (official set supporting this manuscript); Human – Ensembl v19.34a ; Fugu – Ensembl v21.2c.1.

Orthologous groups were assembled using all proteins from all three organisms, as described in the main methods part of the manuscript. Genes not covered by orthologous groups were systematically searched against the other genomes using the Smith-Waterman algorithm and placed into the 'homology' section if they had at least one hit scoring 50 bits or better (50 bits correspond to an e-value of 10^{-6} when searching against the human genome, and for the other genomes it corresponds to even better e-values). All remaining genes were placed into the category 'unique'. All similarity searches were restricted to the translation of the single longest predicted transcript per locus.

Methods for Fig. 6

The following gene sets were used to support the figure: Chicken – Ensembl v22.1.1 (official set supporting this manuscript); Human – Ensembl v19.34a ; Fugu – Ensembl v21.2c.1.

The identity values (percent amino acid sequence identity) refer only to those portions of the 1:1 orthologous proteins that were aligned by Smith-Waterman searches.

Gene Ontology assignments for the human protein set were obtained from EnsMart (v21). The GO hierarchy 'biological process' was simplified to accommodate no more than 20 categories (including the category 'unassigned') using an automated procedure described earlier [28]. Backtracking was used to translate the actual GO-terms to one of the 20 categories for each protein.

For B) and C), the distributions are shown using color-density plots; these were derived from dot-clouds, after smoothing with a Gaussian smoother having a smoothing width of 20% and mirror-like backfolding at 100% and 0%.

Methods for Fig. 7

The following gene sets were used to support the figure: Chicken - Ensembl v22.1.1 (official set supporting this manuscript); Human - Ensembl v19.34a; Fugu - Ensembl v21.2c.1

A) Domain counts: Protein domains were ranked by the number of matching genes. To depict the most prominent cases of domain family loss, innovation, expansions and contractions, we required at least two-fold difference in the number of chicken and human genes, having at least 20 matching genes in either human or chicken for families with at least 5 members for loss and innovation cases, and restricting the list to the 25 most differing families.

B) Orthologous relations: Orthologous groups were assembled as described above. As in A) we required at least a two-fold difference in the number of chicken and human genes, having at least 5 orthologous group members in either human or chicken and restricted the list to the top 25 examples.

Methods for Fig. 8

The following gene sets were used to support the figure: Chicken – Ensembl v22.1.1 (official set supporting this manuscript); Human – Ensembl v19.34a ; Mouse – Ensembl v19.30 ; Rat – Ensembl v19.3a ; Fugu – Ensembl v21.2c.1; Fruitfly – Ensembl v19.3a ; Mosquito – Ensembl v19.2a ; C.elegans – Ensembl v19.102 ; C.briggsae – Ensembl v19.25 ; Arabidopsis – TAIR database, version R5v01212004

Orthologous groups were assembled using proteins from all ten organisms (see above, "orthology detection". Each orthologous group was counted as a single gain in the one ancestral organism whose descendants are needed to cover all of the proteins in the group; orthologous groups were potentially counted multiple times as losses (depending on their pattern of species coverage), assuming a parsimonious scenario with as few losses as possible in order to accommodate the observed pattern.

To estimate the number of genes in extant genomes, we counted, for each genome, all the genes present in the orthologous groups (i.e., having at least one recognizable ortholog in any of the other genomes). In addition, we considered those genes that had paralogy support within the genome – but only if the similarity at the protein level was found to be sufficiently strong to rule out most cases of fragmentation, pseudogenes or gene-prediction artifacts. Specifically, similarity within the genome had to be above 200 bits in Smith-Waterman searches, covering at least 200 amino acids in each of the proteins, and the similarity within the genome had to be higher than the similarity towards any protein in any of the other genomes.

Movement of genes between sex chromosomes and autosomes

In studies of interchromosomal gene traffic in mammals[31] and *Drosophila* [32], it was shown that there is greater export (and import) of X chromosome genes to other locations

in the genome, relative to autosomal genes. The results were explained by male selection bias; either X chromosome inactivation during male meiosis encourages autosomal gene redundancy, or perhaps sexual antagonism promotes export of genes beneficial to males vs. females off of the X due to the greater effective population size of X in females vs. males. An analogous analysis of the chicken genome might help to resolve these hypotheses, since birds show a reversed heterogametic sex system compared to mammals and lack a corresponding Z chromosome inactivation pattern in oogenesis. If an excess of genes have moved from the Z to the autosomes, the explanation cannot be chromosome inactivation and must be something else, like sexual antagonism [33]. However, no significant excess of retroposed genes in the chicken from parental genes on the Z chromosome was detected (no Z-derived functional retrogenes were identified, see Table S4). This is consistent with sex chromosome inactivation explaining X gene export in mammals (see [33] for a discussion). However, it can also be explained by the fact (main text) that retroposed copies are very rare in the chicken genome, and most, if not all, retropositions occurred before the relatively recent emergence (102-170 million years ago) of sex chromosomes in birds [34] .

Methods – Exploring genome architecture

Correlation of physical and genetic maps

A molecular cytogenetic definition for the chicken karyotype has recently been proposed based on chromosome morphology and FISH, using specific chromosome paints and BAC clones [35] that, in most cases, correlates well with the length of sequence assembled for each chromosome (Table S2). Exceptions are GGA16, 22 and the smallest

chromosomes, GGA32-35 and the sex chromosomes Z and W. Sequence coverage of these chromosomes is under-represented, so they were not considered in the following analyses of autosomes. One of the most interesting features of the chicken genome landscape is the extreme variability of chromosome size. The macrochromosome group is 4 and 13-fold longer than the intermediate and microchromosome groups, on average. The 10 macro- and intermediate chromosomes contain 82% of the current autosomal sequence.

Physical vs genetic map position

The availability of genetic marker data and the genome sequence enables a direct comparison between genetic length (cM) and physical distance on the chicken genome. Genetic and physical maps have been integrated using 1471 polymorphic loci (Table S2) from the chicken sex-averaged map [36,37], with positions in centimorgans (sex-averaged distances in cM) and megabase pairs (Mb), respectively (Table S2). The total genetic map length is ~3700 cM for a genome of 1.07×10^9 bp [38]. In contrast to mammals [39], there is only ~1% difference between the genetic maps arising from meioses in male and female chickens [38] . The male map is larger; however, the overall difference is within the limits of error. This suggests that the sex-specific maps are not in agreement with "Haldane's rule" [40], that the homogametic sex should be the longer map. Genetic length declines linearly with physical length until the threshold for an obligatory recombination per chromosome is reached. For these small chromosomes the information in genetic maps is insufficient to distinguish between an obligatory recombination per chromosome, or per chromosome arm (microchromosomes have been characterised as acrocentric) (Fig. S17).

A robust estimate of the rate of recombination (cM/Mb) has been calculated across each chromosome by taking the median of the slopes between all possible pairs of genes (Table S2) and, where the location of the centromeres is known (see assembly section), the approximate genetic length of each chromosome arm has been derived from a robust (lowess based) smoothed fit of genetic versus sequence distance over the whole chromosome.

Comparison of methods for identifying CpG-islands

To reduce the number of spurious CpG-islands in the human genome, Takai and Jones [41] modified the Gardiner-Garden and Frommer approach to define CpG-islands as regions greater than 500 bp in length with a GC-content of greater than 55% and an observed CpG/expected CpG of 0.65. Because this algorithm misses CpG-islands smaller than 500 bp in length, we adopted a different approach. First, the genome sequence was masked using the output of RepeatMasker [42] (with the -sensitive setting) and Tandem Repeats Finder (simple repeats with period of 12 or less). Then CpG islands were identified using thresholds employed by Gardiner-Garden and Frommer - "GGF" [43]: length >= 200bp, GC content >= 50%, ratio of observed CpG to expected CpG (obs/exp) >= 0.6. All CpG locations in each chromosome, as well as unmasked C's and G's up to and including each CpG, were counted. A sliding window search was performed on the set of CG locations (as opposed to a sliding window of genomic sequence) as follows. Using CG locations as window boundaries, a window of less than 200bp would be expanded to include the next CG to the right. If expanded from less than 200bp to at least 200bp, but then having less than 50% GC content or less than 0.6 obs/exp, it would be contracted by moving the window start up to the next CG in the window until the size fell back below 200bp. A

window of 200bp or more, with at least 50% GC content and obs/exp at least 0.6, would be expanded to include the next CG to the right, unless that addition would cause the window to fall below the GC and obs/exp thresholds, in which case the window would be declared an island and that next CG used as the potential start of the next island.

The islands predicted by our methods were re-evaluated by a separate program (*cpg-score*, A. Hinrichs), that verified that the sequence at each annotated island location was >=200bp, >=50% GC, >= 0.6 obs/exp and thus met the definition of a CpG-island. Ideally we need experimental proof that these regions are methyl-CpG-free, e.g., using methylation sensitive restriction enzymes. However, it is likely that this bioinformatic approach gives the best current estimate for locations and numbers of CpG-islands. In the main text, we examine the characteristics of 51,153 predicted CpG-islands that have been aligned with specific autosomes on which 12,483 Ensembl gene predictions have been placed.

We believe the method used herein, based on using CpG locations as window boundaries, is the most reliable way to detect putative CpG-islands. In the past most methods have been based on a genomic window search, that underestimates the number of CpG-islands. A genomic window search starts with an arbitrary segment of sequence, and from there it can expand, contract or shift in ways that will incrementally improve the score(s) of the window. This genomic window method finds most islands but not as many as possible because incremental adjustments will be carried out only so far as they bring the score of the current window above or below the thresholds for length, GC-content and obs/exp CpG ratio. In some cases, if incremental improvements were made on one side of an island, the other side could extend even farther. Using CpG locations as

the boundaries for each window is more favorable for the obs/exp CpG score because including other neighboring C or G bases can only dilute the obs/exp ratio: including other A or T bases would raise the obs/exp ratio slightly but would dilute the GC percentage. Finally, using CpG locations as window boundaries also has an intuitive appeal because CpG's are the feature of interest in CpG islands.

Detail on variation in CpG islands and gene densities

In chicken, most CpG-islands are short, with a mean size of 828±1629 bp (range 200-80,868), 50-87% GC-content and O/E CpG ratios of 0.60-2.5. The GC content and O/E CpG ratios show little relationship with each other: macrochromosomes $r_s = -0.08$, $P < 0.001$; intermediates $r_s = -0.08$, $P < 0.001$; microchromosomes $r_s = -0.08$, $P < 0.01$). None of these basic characteristics differ between the chromosome classes. Overall, the frequency distribution of CpG-island lengths is similar between chromosome classes but is very highly skewed. This distribution has an excess of small (200-800 bp) and large (10,000+ bp) CpG-islands when transformed to the inverse length scale to remove the skewness. The longest CpG-island is on GGA2 (position 31812231-31893099) at 80,868 bp. The top-level chain in this region corresponds to the human *HOXA* cluster. The next largest island is on GGA7 (position 17076201-17140391) at 64,190bp, which is the *HOXD* cluster. Third largest is on GGA9 (position 7927467-7983179) at 55,712 bp, corresponding to the human FOXL2 region.

Forty-eight percent of predicted CpG-islands overlap a gene and the proportion decreases with chromosome length. This is partly due to a decrease in the density of genes with chromosome length ($r_s = -0.292$, $P < 0.0013$). We find 38% of chicken CpG-islands conserved with islands found in the human genome and the proportion decreases

with chromosome length (r_s = -0.360, P < 0.001). CpG-islands were classified further into conserved/not conserved and near/not near gene (Table S5). 10% of conserved CpG-islands do not overlap with an Ensembl gene prediction. More detailed analysis shows that within this set, 685 are located within 5-kb of an Ensembl gene, 2286 overlap existing chicken ESTs and 791 overlap a non-chicken RefSeq sequence. Taking these sequence overlaps (3762) into account, suggests that only 1889 of the conserved CpG-islands are not near a gene. These may represent distant regulatory regions or may serve another function. Overall, 26% of all predicted CpG-islands are not conserved and do not overlap any genes, including Ensembl, non-chicken RefSeq genes, or chicken ESTs. Of the nonconserved, non-genic CpG islands, 27% are found within 5-kb of an Ensembl gene (upstream or downstream), but when ESTs and non-Chicken RefSeqs are considered, nearly 50% of the non-conserved, non-genic CpG islands fall within 5-kb of a gene or putative gene. The rest may represent sequences of unknown function or may be a product of the prediction algorithm with no biological function at all. Mean exon length (macrochromosomes 166.4±2.7, intermediates 160.6±3.5, microchromosomes 163.9±7.9) and mean exon number per gene (macrochromosomes 17.3±0.5, intermediates 19.2±1.6, microchromosomes 17.0±1.9) vary little from chromosome to chromosome (exon number r_s = 0.467, P = 0.015). In contrast, there is a strong dependency between mean intron length and chromosome size (Fig. 10).

Housekeeping genes do not appear to be enriched on any size class of chromosomes

The higher CpG-island content and gene-density of microchromosomes prompted the hypothesis [44] that these encode most of the "housekeeping" genes [43,45] . Such genes are expected to be expressed in all or most tissues, and so to test this hypothesis we examined

the range of "tissue specificity" or "gene expression breadth" of chicken genes. Chicken ESTs were mapped to the genome assembly and to Ensembl genes (± 1kb), using BLAT[46] at a 95% identity threshold. The mapped ESTs were partitioned into 10 distinct tissue types (brain; fat & skin; bone & connective; heart; kidney & adrenal; immune; liver; female reproduction; alimentary; testis). Percentage amino acid sequence identities of 1:1 chicken-human orthologues were calculated as previously described. Genes were partitioned according to the number of these 10 tissues in which they are expressed: Partition A: 1-3 tissues, Partition B: 4-6 tissues, Partition C: 7+ tissues. Orthologous percentage amino acid identity distributions were compared using the Kalmogorov-Smirnov test. Distributions for partitions A and C were significantly different ($P < 0.001$). Genes expressed in few (1-3), or else many (7+), tissues are evenly distributed among the three classes of chicken chromosomes (macrochromosomes, intermediates, microchromosomes). Therefore there is no evidence to support the idea that "housekeeping" genes are clustered on microchromosomes.

Characterization of exon number, exon, intron and gene lengths

Exon numbers, exon, intron and gene lengths were all abstracted from the Ensembl gene prediction (chicken – ensemble v22.1.1) set. As all lengths followed positively skewed distributions, the graphs in figures in the text show geometric mean lengths which are less influenced by extreme values in the tails of distributions. Exons and introns are averaged as individual values, ignoring their gene designation.

Correlation, regressions and statistical analyses

We used nonparametric Spearman correlation coefficients (r_s, [47]) to assess covariation between recombination rate, chromosome length and other sequence variables.

Stability of bird chromosomes seen in chicken-turkey comparisons

Chromosome banding experiments have shown extensive homology in the general morphology and in the specific banding patterns between avian chromosomes [48] . The latter presumably reflects a high degree of conserved synteny between birds; this contrasts with the wide diversity found in mammalian karyotypes [49]. This was explored more thoroughly by comparative FISH experiments between chicken (*Gallus domesticus)* and turkey (*Meleagris gallopavo*) (D. Griffin, data not shown). Results reveal a conservation of synteny that is remarkable considering the 20-50 million years since the two species diverged. Results from a single BAC, MCW0275, gave clear evidence of an intrachromosomal rearrangement mapping near the middle of turkey chromosome 10 (a telocentric chromosome) and towards the telomere of chicken chromosome 8 (a metacentric chromosome. However, chromosome painting experiments revealed no inter-chromosome rearrangements between chicken chromosomes 1, 3, 5, 6, 7, 8, 9, 10 and Z and their turkey orthologues. In contrast, chromosome paints for chicken chromosomes 2 and 4 each hybridized to two turkey chromosomes. Banding comparisons resolved the homologies as being between chicken chromosome 2q and turkey chromosome 3, chicken 2p and turkey 6, chicken 4q and turkey 4, and chicken 4p and turkey 9. Comparisons with outgroups such as greylag goose (*Anser anser*), quail (*Coturnix coturnix*), blackbird (*Turdus merula*), emu (*Dromais novaehollandiae*) and Rhea (*Rhea Americana*) suggest that chicken chromosome 2, which is common to all these birds, is the ancestral type [50,51] . The inference is, therefore, that chromosomes 3 and 6 in turkey (and indeed in 3 pheasant species) arose by fission of the ancestral chromosome 2. The

derivation of chicken chromosome 4 is less clear. Chicken chromosome 4 paint hybridizes to both a larger and a smaller chromosome in turkey, blackbird, emu and Rhea, but not in goose or quail [50,51]. If one assumes that the "two chromosome pattern" is the ancestral type, then the differing patterns in chicken and turkey chromosomes 4 could have arisen by one of two different scenarios, both involving a chromosome fusion. The first hypothesis involves a fusion event in the Anseriforme-Galliforme ancestor with a subsequent fission occurring during the evolution of turkey and pheasants. In the second hypothesis, independent fusion events occurred in the evolution of goose and chicken/quail.

Segmental Duplications

Recent segmental duplication content was assessed using two different methods: the whole-genome assembly comparison (WGAC) method and the whole genome shotgun sequence read detection (WSSD) method. WGAC was performed as described previously [52]. All alignments >1 kb and >90% identical were analyzed. Common repeats were excluded (RepeatMasker) and initial seed alignments were set at 500 bp. The "unknown" chromosome, which contains unmapped chicken sequence, was considered as an independent chromosome and statistics were computed with and without unmapped sequence (Table S11). The average length of alignments detected on the unknown chromosome (1712 +/- 1388 bp) was shorter than mapped genomic sequence (3060 +/- 2631). A modified version of the WSSD method [53] was implemented during the analysis of the chicken genome. Each chicken read was searched by BLAT against the Gal2 assembly. In our first pass, we required at least 400 bp of the read and >90% of the read length to align. We subsequently filtered all alignments where the degree of sequence identity was >94% and which contained at least 300 bp of unique sequence. Read-depth

across the chicken genome was measured in 10 kb windows sliding every 1 kb. Repeats and sequence gaps were excluded.

Using these methods, we analyzed the chicken genome assembly for the presence of pairwise alignments larger than 1kb with more than 90% sequence identity (Tables S11, S12, S13). More than half of the duplicated segments (62.75 Mb) are not within sequences assigned unambiguously to chromosomes, a finding consistent with other recently published genome assemblies[54 4] . Our analysis of the chicken genome predicts an abundance of short duplications with an average pair wise alignment length of 1799 bp. Almost all mapped duplications (93 %) within the chicken genome are intrachromosomal (excluding random). Macrochromosomes show fairly uniform duplication content (~5%), while microchromosomes are much more variable. GGA11 appears to be the most enriched, with as much as 25% of the chromosome predicted to be duplicated (Table S11).

As described in the main paper, the full set of duplicated segments show a high degree of sequence identity: 91% of the alignments consist of duplications with near perfect sequence identity (>98%) (see main paper; Fig. 13). To investigate whether this is a biological property or whether it represents assembly artifacts, we used a second method to predict duplications based on an excess of underlying whole-genome shotgun sequence reads. This method has been used previously [53,55] to confirm nearly identical duplications within a genome sequence. Only 26% of the WGAC segmental duplications (32.3/122.7 Mb) could be confirmed by this approach. While 65% of the WGAC duplications sharing 90-97% identity could be confirmed, only 22% of duplications with near perfect sequence identity (>98%) were validated.

Of the 3.7% of Ensembl predicted genes showing evidence of being recently duplicated (see main text), we analyzed a subset (n = 249) that showed evidence of two or more duplicated exons and likely arose exclusively by duplication as opposed to retroposition. The majority of these genes were not mapped within the genome assembly (n = 135) or had no assigned annotation (n = 148). Annotated genes included a large number of immunity-related genes such as the immunoglobulins, T-cell receptors and various MHC-related proteins. Potential lineage-specific duplicated genes or pseudogenes include basic-helix loop helix transcription factors, otokeratin, ribonuclease A, an alcohol dehydrogenase subunit and a putative microtubule associated protein (Nau).

Methods – Evolution of vertebrate genomes

Conserved Synteny

Data sets used:

Chicken – Ensembl v22.1.1 (official set supporting this manuscript)

Human – Ensembl v19.34a

Mouse – Ensembl V13.30.1

Fugu – Ensembl v21.2c.1

Tetraodon – Genoscope June 2004

Estimation of the number of rearrangement events and the reconstruction of mammalian ancestor genome architecture.

In the current study, we have chosen to use the set of orthologous genes rather than the "similarity anchors" as described before for human-mouse-rat comparison [56] since the chicken genome is rather distant from mammalian genomes and using similarity anchors (at lower similarity thresholds, due to the distance) leads to many spurious hits, which can be difficult to unambiguously assemble into synteny blocks. Moreover, highly diverged genes may not generate any four-way anchors (based on exact nucleotide *l*-mer matches) but still can be reliably detected with amino acid scoring matrices.

To generate synteny blocks using genes as anchors, we started from a set of 6447 four-way orthologous genes, pre-filtered for evidence of conserved pairwise synteny using SyntQL (Zdobnov, unpublished) as described earlier[28]. We then applied GRIMM-Synteny[57] to determine the synteny blocks, imposing a threshold of at least 3 genes per synteny block and allowing a tolerance for micro-rearrangements and for up to 2 intervening genes per species. This gave a set of 586 synteny blocks containing 6140 genes. Some of the three-way synteny blocks previously described [56,58] have fewer than 3 genes, and thus these blocks were not included in our set of four-way synteny blocks. Furthermore, not every mammalian gene has an ortholog in chicken, and any of the previously defined human-mouse-rat synteny blocks that had "lost" some of their genes in chicken were excluded in the four-way comparison. As a result, our 586 four-way synteny blocks correspond to only 299 three-way synteny blocks after "projecting" to the human, mouse, and rat genomes. The average size of synteny blocks varies: 3.2 Mb in human, 2.9 Mb in rat, 2.8 Mb in mouse and only 1.2 Mb in chicken. These synteny blocks were used as input to the GRIMM[59,60] and MGR[61] algorithms to reconstruct the

likely rearrangement evolutionary scenarios, considering inversions, translocations, fusions, and fissions.

Methods for Fig. 14.

To estimate relative branch lengths of the phylogenetic tree using genome structure divergence we used two conceptually different approaches: 1) counting synteny breaks where ancestral state is supported by synteny to an outgroup species, and 2) reconstruction of ancestral genomes through a combinatorial search for a most parsimonious rearrangement scenario (as described above), with details shown in the supporting table.

The synteny blocks referred to below were identified using the SyntQL algorithm (Zdobnov, unpublished) as described earlier [28], looking for a conserved neighborhood of each orthologous gene pair, but allowing for up to 4 intervening genes and micro-rearrangements inside otherwise orthologous chromosomal loci. To simplify the synteny map construction, we identified orthologous genes as reciprocally best BLASTP matches in inter-proteome comparisons, the subset of which found in synteny is identical to that of the set provided using the more stringent gene orthology detection method described in the protein section. The synteny maps are available from http://azra.embl-heidelberg.de/~zdobnov/Chicken/.

The fraction of orthologous genes that retained their genomic neighborhood shown as pie-charts were calculated as the ratio of orthologous genes that form the synteny blocks to the total number of identified orthologous genes. It varied from 87% to 84% for the 9679 chicken/human and 9313 chicken/mouse orthologs in synteny, and it is

95% for the 14707 human/mouse orthologs in synteny, and about 48% for the 4608 chicken/fugu and 4224 chicken/tetraodon orthologs in synteny.

To estimate the relative frequencies of synteny breaks (without discriminating the type of chromosomal rearrangements) along the branches of the phylogenetic tree, we counted the number of synteny splits where the ancestral state is supported by synteny to an outgroup species, e.g. if there are two neighboring genes in synteny between chicken and human that are found in different synteny blocks in mouse we counted a genome break on the rodent lineage. Thus we measured the length of the MA-Human (h) and MA-Mouse (m) branches by using chicken as the outgroup and counting all events when adjacent chicken genes belong to the same synteny block in one species and to different blocks in the other species. Using Tetraodon fish genome as the outgroup we measured the length of the AA-Chicken (c), AA-MA-Human (a+h) and AA-MA-Mouse (a+m) branches, yielding numbers in a different scale due to more sparse synteny conservation. The numbers shown in the table are derived from solving the following simple equation system:

$$\begin{cases} h/m=43/109, \\ c=35, \\ a+h=56, \\ a+m=69 \end{cases}$$

Since synteny maps involving the Tetraodon genome are quite sparse, the estimates were also done using the Fugu genome as the outgroup, resulting in practically the same estimates for the MA-Human (h) and MA-Mouse (m) branches and significantly larger AA-Chicken (c=62), AA-MA (a=68) branches, keeping a similar a/c ratio (1.3 vs 1.1).

The lower estimates of inter-chromosomal rearrangements were done by counting the number of inter-chromosomal relations having at least one synteny block with more than three genes and subtracting the number of putative ancestral chromosomal correspondences (see Fig. S14 and all maps available from http://azra.embl-heidelberg.de/~zdobnov/Chicken/index.html#macro).

Associations among human-chicken length ratio, gene density and GC content

The 6727 non-overlapping intervals used to analyze the ratio of chicken to human genome lengths were computed by partitioning each of the 1000 highest scoring chains of blastZ alignments (see below) as follows. Call a gap-free segment within an alignment a "block". Our chaining procedure works with blocks, rather than indivisible local alignments. We looked for blocks B and C such that (1) C is the first block in B's chain whose human start-position is at least 100 kb after B's and (2) neither species has any gaps of over 40 kb between adjacent blocks between B and C in the chain. For any such pair of blocks we assumed that the human and chicken intervals from the start of B to one position before the start of C are "unbroken orthologs".

Many parameters were tested singly and in combination for their ability to explain the variation in length ratio (hL/cL) in these unbroken orthologs. These parameters were the ratio of the densities of interspersed repeats (or masked bases, hMS/cMS), the ratio of gaps in the alignments (a proxy for indels, hGap/cGap), distance from human (dhTel) and chicken (dcTel) telomeres, ratio of GC content in the non-repetitive sequence ((hGC/cGC)(um), "um" for unmasked), ratio of GC content in the repetitive sequence ((hGC/cGC)(ms), "ms" for masked). Multivariate regressions were conducted

considering all the 6727 unbroken orthologous intervals, and also separating such intervals by chicken chromosome classification into macro, intermediate and micro. Log scales were used for all variables to regularize their distribution. Regression results for the 218 intervals on chicken sex chromosomes are not reported separately here, but those intervals were included when fitting the overall regression.

The results reported in Table S6 show that, among the genomic parameters considered in our analysis, the ratios of repeat density (positive association), GC content (negative association) and gap frequencies (positive association) are the major contributors to explaining the variation in length ratio, whereas the contribution of distance from telomeres is ambiguous. From the multivariate regression fits, the share of explained variability is approximately 22% when considering all chromosomes, slightly lower for macro and intermediate chromosomes, and higher for microchromosomes. Also the relative importance of GC content ratios vis-a-vis repeat density ratios increases for smaller chromosomes.

Methods for Fig. 15

Fig. 15 shows the variation in the ratio of lengths of human and chicken DNA in aligning segments. (A) An example of length ratio variation, HSA4 and GGA4. Dotplot comparing orthologous regions of human and chicken, showing variable slope. In the upper left, the human genomic interval is shorter than the orthologous chicken interval, and the human GC content is much higher than in chicken (54% for the first megabase in human compared to 42% in the orthologous chicken region). In the lower right, the human sequence is about 2.5 times longer than the orthologous chicken region, and the GC contents are very similar (both around 39.5%). (B) Genome-wide variation in length

ratios of aligned segments, comparing human-chicken and human-mouse. Distributions of log length ratios in human-chicken and human-mouse alignments. Log transformations were used to regularize the distributions. For human-chicken alignments, log length ratio distributions are shown for all DNA (green) and for non-repetitive DNA (red). For human-mouse alignments, only the distribution for all DNA is shown (black.)

Methods – Illuminating the human genome; non-coding alignments

Human/chicken whole-genome alignments were obtained by using the program blastZ[62] to produce short (typically 100-1000 bases) 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[63]. These alignments, which were used to generate data for Figs. 2,15-17, S1,S15 and Tables 4,S6,S7, can be obtained from the U.C. Santa Cruz Browser (http://genome.ucsc.edu/).

BlastZ program was run with the parameters L=10000, K=2200, H=2000, Y=3400 and the substitution scores:

A	C	G	T
91	-90	-25	-100
-90	100	-100	-25

-25 -100 100 -90

-100 -25 -90 91

We used the following piecewise linear gap-scoring function:

bases	1	2	3	11	111	2111	12111	32111	72111	152111	252111
cost	325	360	400	450	600	1100	3600	7600	15600	31600	56600

Where the gap score is linear, and very small, past 152111 bases. Gaps in both sequences simultaneously cost 300 more than single-sequence gaps. The gap scoring function needs to obey some constraints, but is ultimately chosen empirically. Small gaps are scored similarly to blastZ. Penalties for large gaps need to be small enough to permit chaining to occur across syntenic regions, but large enough to prevent similar-by-chance rather than similar-by-homology alignments from coalescing into chains. The convexity in the scoring function is also helpful in preventing inversions-within-inversions from breaking up the insertion caused by the main inversion.

Methods for Table 4

The following percentages of human RefSeq genes were determined to have a non-coding match of at least 100 bp with chicken: 5' flank: 2.1%; 5' UTR: 3.6%; an intron: 32.0%; 3' UTR 18.3%; 3' flank: 4.5%. Percentages for flanking regions are low because we considered only RefSeq genes that are at least 5 kb from any RefSeq or ENSEMBL transcript annotation; this filtering eliminated 52% of RefSeq-annotated 5′UTRs and 47% of the 3′UTRs. Also, it is quite plausible that 5' UTRs are on average substantially less well documented than 3' UTRs, which would explain some of the discrepancy

between their frequencies of alignment. For each of the five classes of genes, we looked for enrichment of Gene Ontology categories. For instance, of the human gene products having GO annotations, 2.6% (305 out of 11590) were annotated as having ion-channel activity, whereas 4.1% (150 out of 3634) of the ones showing intronic conservation have ion-channel activity. We calculated the probability of an enrichment as strong or stronger occurring by chance alone; using an hypergeometic distribution, this results in a p-value of 3.8×10^{-11}. The Table contains p-values for the five gene classes across fourteen GO categories, but no correction for multiple testing (e.g. Bonferroni) is implemented; the small p-values would remain significant after correction. For introns of genes with ATP-binding products, the software we used (the R software for statistical computing) computed a negative number, so we report 0.0.

Conservation patterns in proximal *cis*-regulatory regions

Conservation in functional elements was determined using the GALA database[64] (www.bx.psu.edu). Datasets for cis-acting functional elements were compared using the intersection feature with alignments from human and chicken that fell within second level nets[63]. Known regulatory regions[65], functional and predicted promoters[66], CpG islands (A. Heinrichs and A. Law, unpublished method) and conserved matches to transcription factor binding sites are all available as tracks in the GALA database.

Conserved non-coding fraction

For analysis of the relationship between conserved non-coding regions and coding regions (Fig. S15), we considered only alignments scoring at least 10,000. This was done

to avoid circularity in the argument, since (because of the L=10000 parameter for blastZ), alignments scoring less than 10,000 contained in the output are, by definition, close to an alignment scoring at least 10,000 (i.e., generated by the "inner alignment process" – see the H parameter). These alignments were also used to define the high-CNF intervals. Another step that we took to make the CNF logically independent of the coding fraction was to eliminate from the definition of CNF all non-coding regions that are contained in an alignment that intersects a coding region. The need for this precaution arises because alignments that start in a coding region can extend into adjacent non-coding segments having only very modest similarity; once an alignment achieves a certain minimum threshold score, regions of score 0 or more are added to either side. This property of the Smith-Waterman approach to local alignments biases the non-coding aligned positions to be adjacent to coding regions. Thus for the "conserved non-coding fraction", we started with non-repetitive regions that are not in an alignment that intersects a coding region, and asked for the fraction that is aligned.

Our method of determining high-CNF segments of the human genome is as follows. Let X denote a CNF threshold (in the paper, we use X = 0.08). An interval in the human genome is "full" if any initial or terminal subinterval (including the interval itself) has CNF at least X. It is straightforward though not trivial to see that if two full intervals intersect, then their union is full. This implies that a chromosome can be decomposed into maximal full intervals, separated by segments that contain no non-coding alignment. It is possible to compute this decomposition in time proportional to the chromosome's length. Specifically, for each position in the chromosome, define a score s by: s = 1-X if the position is non-repetitive and in a non-coding alignment, s = -X for non-repetitive

positions that are not in any alignment, and s = 0 otherwise (i.e., positions in repeats, an exon or an alignment that intersects an exon). The score of an interval is the sum of the s-values over all positions in the interval; an interval's score is positive if and only if its CNF is at least X. The decomposition can be found with the algorithm in Fig. 6 of "Parametric recomputing in alignment graphs"[67]. In practice, we remove positions from either end of the interval for which the score is 0. Our high-CNF intervals are just those maximal full intervals exceeding a length threshold (500 kb in the paper).

References

1. Lee, M. K. et al. Construction and characterization of three BAC libraries for analysis of the chicken genome. *Anim Genet* **34**, 151-2 (2003).

2. Khatib, H., Genislav, E., Crittenden, L. B., Bumstead, N. & Soller, M. Sequence-tagged microsatellite sites as markers in chicken reference and resource populations. *Anim Genet* **24**, 355-62 (1993).

3. Abplanalp, H., Sato, K., Napolitano, D. & Reid, J. Reproductive performance of inbred congenic Leghorns carrying different haplotypes for the major histocompatibility complex. *Poult Sci* **71**, 9-17 (1992).

4. The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

5. Huang, X., Wang, J., Aluru, S., Yang, S. P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Res* **13**, 2164-70 (2003).

6. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175-185 (1998).

7. Flint, J. et al. Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Hum Mol Genet* **10**, 371-382. (2001).

8. Bulger, M. et al. Conservation of sequence and structure flanking the mouse and human beta-globin loci: the beta-globin genes are embedded within an array of odorant receptor genes. *Proc Natl Acad Sci U S A* **96**, 5129-5134 (1999).

9. Matzke, M. A. et al. A 41-42 bp tandemly repeated sequence isolated from nuclear envelopes of chicken erythrocytes is located predominantly on microchromosomes. *Chromosoma* **99**, 131-7 (1990).

10. Ovcharenko, I. et al. Evolution and functional classification of vertebrate gene deserts. (in the press).

11. Kaufman, J. et al. The chicken B locus is a minimal essential major histocompatibility complex. *Nature* **401**, 923-5 (1999).

12. Afanassieff, M. et al. At least one class I gene in restriction fragment pattern-Y (Rfp-Y), the second MHC gene cluster in the chicken, is transcribed, polymorphic, and shows divergent specialization in antigen binding region. *J Immunol* **166**, 3324-3333 (2001).

13. Miller, M. M. et al. 2004 Nomenclature for the chicken major histocompatibility (B and Y) complex. *Immunogenetics* **56**, 261-279 (2004).

14. 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**, 19440-51 (2001).

15. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-64 (1997).

16. Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Res* **32 Database issue**, D109-11 (2004).

17. Hofacker, I. L. Vienna RNA secondary structure server. *Nucleic Acids Res* **31**, 3429-31 (2003).

18. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. *Nucleic Acids Res* **31**, 439-41 (2003).

19. Eddy, S. R. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* **3**, 18 (2002).

20. Korf, I., Flicek, P., Duan, D. & Brent, M. R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1**, S140-8 (2001).

21. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78-94 (1997).

22. 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).

23. Curwen, V. et al. The Ensembl automatic gene annotation system. *Genome Res* **14**, 942-50 (2004).

24. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12**, 1269-76 (2002).

25. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res* **32 Database issue**, D138-41 (2004).

26. Letunic, I. et al. SMART 4.0: towards genomic data integration. *Nucleic Acids Res* **32 Database issue**, D142-4 (2004).

27. Koonin, E. V. A Non-Adaptationist Perspective on Evolution of Genomic Complexity or the Continued Dethroning of Man. *Cell Cycle* **3**, 280-285 (2004).

28. Zdobnov, E. M. et al. Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. *Science* **298**, 149-59 (2002).

29. von Mering, C. et al. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* **31**, 258-61 (2003).

30. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85**, 2444-8 (1988).

31. Emerson, J. J., Kaessmann, H., Betran, E. & Long, M. Extensive gene traffic on the mammalian X chromosome. *Science* **303**, 537-40 (2004).

32. Betran, E., Thornton, K. & Long, M. Retroposed new genes out of the X in Drosophila. *Genome Res* **12**, 1854-9 (2002).

33. Wu, C. I. & Xu, E. Y. Sexual antagonism and X inactivation--the SAXI hypothesis. *Trends Genet* **19**, 243-7 (2003).

34. Handley, L. J., Ceplitis, H. & Ellegren, H. Evolutionary strata on the chicken Z chromosome: implications for sex chromosome evolution. *Genetics* **167**, 367-76 (2004).

35. Masabanda, J. S. et al. Molecular cytogenetic definition of the chicken genome: the first complete avian karyotype. *Genetics* **166**, 1367-73 (2004).

36. Groenen, M. A. et al. A consensus linkage map of the chicken genome. *Genome Res* **10**, 137-47 (2000).

37. Schmid, M. et al. First report on chicken genes and chromosomes 2000. *Cytogenet Cell Genet* **90**, 169-218 (2000).

38. Groenen, M. A. M., Crooijmans, R.P.M.A. in *Poultry Breeding and Biotechnology* (ed. Muir, W. M. a. A., S.E.) 497-536 (CABI International, 2003).

39. Donis-Keller, H. et al. A genetic linkage map of the human genome. *Cell* **51**, 319-37 (1987).

40. Haldane, J. B. S. Sex ratio and unisexual sterility in hybrid animals. *J. Genet.* **12**, 101-109 (1922).

41. Takai, D. & Jones, P. A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* **99**, 3740-5 (2002).

42. Smit, A. a. G., P. (1999). RepeatMasker. http://www.repeatmasker.org

43. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J Mol Biol* **196**, 261-82 (1987).

44. McQueen, H. A. et al. CpG islands of chicken are concentrated on microchromosomes. *Nat Genet* **12**, 321-4 (1996).

45. Holmquist, G. P. Evolution of chromosome bands: molecular ecology of noncoding DNA. *J Mol Evol* **28**, 469-86 (1989).

46. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).

47. Siegel, S. *Nonparametric Statistics* (McGraw-Hill:London, 1956).

48. Auer, H., Mayr, B., Lambrou, M. & Schleger, W. An extended chicken karyotype, including the NOR chromosome. *Cytogenet Cell Genet* **45**, 218-21 (1987).

49. Andersson, L. et al. Comparative genome organization of vertebrates. The First International Workshop on Comparative Genome Organization. *Mamm Genome* **7**, 717-34 (1996).

50. Shetty, S., Griffin, D. K. & Graves, J. A. Comparative painting reveals strong chromosome homology over 80 million years of bird evolution. *Chromosome Res* **7**, 289-95 (1999).

51. Guttenbach, M. et al. Comparative chromosome painting of chicken autosomal paints 1-9 in nine different bird species. *Cytogenet Genome Res* **103**, 173-84 (2003).

52. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11**, 1005-17. (2001).

53. Bailey, J. A. et al. Recent segmental duplications in the human genome. *Science* **297**, 1003-7 (2002).

54. Waterston, R. H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).

55. 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).

56. Gibbs, R. A. et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521 (2004).

57. Pevzner, P. & Tesler, G. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res* **13**, 37-45 (2003).

58. Bourque, G., Pevzner, P. A. & Tesler, G. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res* **14**, 507-16 (2004).

59. Tesler, G. Efficient algorithms for multichromosomal genome rearrangements. *J. Comp. Sys. Sci.* **65**, 587-609 (2002).

60. Tesler, G. GRIMM: genome rearrangements web server. *Bioinformatics* **18**, 492-493 (2002).

61. Bourque, G. & Pevzner, P. A. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res* **12**, 26-36 (2002).

62. Schwartz, S. et al. Human-mouse alignments with *Blastz*. *Genome Res.* **13**, 103-105 (2003).

63. 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 U S A* **100**, 11484-9 (2003).

64. Giardine, B. et al. GALA, a database for genomic sequence alignments and annotations. *Genome Res* **13**, 732-741 (2003).

65. Elnitski, L. et al. Distinguishing regulatory DNA from neutral sites. *Genome Res* **13**, 64-72 (2003).

66. Trinklein, N. D., Aldred, S. J., Saldanha, A. J. & Myers, R. M. Identification and functional analysis of human transcriptional promoters. *Genome Res* **13**, 308-12 (2003).

67. Huang, X., Pevzner, P. & Miller, W. in *Springer Lecture Notes in Computer Science, 807* 87-101 (Springer, 2000).

68. Ostertag, E. M. & Kazazian, H. H., Jr. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res* **11**, 2059-65 (2001).

69. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-25 (1987).

70. Page, R. D. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12**, 357-8 (1996).

71. Goodstadt, L. & Ponting, C. P. CHROMA: consensus-based colouring of multiple alignments for publication. *Bioinformatics* **17**, 845-846 (2001).

Table S1 Distribution of reads

Insert size (kb)	3-8	40	180	Total
Vector	plasmid	fosmid	BAC	
Total reads	10.44	0.64	0.15	11.23
Assembled reads	9.96	0.59	0.13	10.68
Paired reads	9.12	0.49	0.1	9.71
Total bases	9.2	0.46	0.11	9.77
>Phred20 bases	6.63	0.34	0.07	7.04
Total seq coverage	8.68	0.43	0.1	9.2
>Phred20 seq coverage	6.25	0.32	0.07	6.64
Physical coverage	17.2	9.2	8.5	34.9

Reads in millions; bases in billions. Coverage estimate based on 1.06Gb genome.

Table S2 Comparison of genetic and physical distances on genome assembly, and number of genetic markers assigned to a chromosome, before and after application of 10% error window.

Chr.	Assembly Length (Mb, include N's)	Assembly Length (Mb, exclude N's)	Genetic Length (cM)	Recombination Rate (cM/Mb)	No Markers Before	No Markers After
1	188.2	183.7	553	2.94	267	253
2	147.6	143.8	474	2.76	199	195
3	108.6	105.9	317	2.46	123	120
4	90.6	88.0	270	2.60	114	109
5	56.3	54.0	199	3.27	98	94
6	33.9	33.4	116	2.76	48	29
7	37.3	35.4	165	3.92	43	40
8	30.0	28.2	105	3.55	60	48
9	23.4	23.1	132	5.17	42	31
10	20.9	19.0	120	5.05	251	215
11	19.0	18.0	90	4.28	27	21
12	19.8	19.0	90	4.92	17	9
13	17.3	16.8	74	4.24	67	49
14	20.6	20.2	77	3.02	21	10
15	12.4	12.2	60	5.09	53	43
16	0.2	0.2	60	Na	4	0
17	10.6	9.9	70	6.17	27	22
18	8.9	8.8	48	7.93	15	8
19	9.5	9.3	41	6.54	13	5

20	13.5	13.3	62	5.84	16	13
21	6.2	6.0	70	12.44	18	13
22	2.2	2.2	21	Na	4	2
23	5.7	5.0	11	Na	15	5
24	5.9	5.8	60	11.82	67	46
26	4.3	3.7	54	11.95	20	14
27	2.7	2.5	60	21.07	23	14
28	4.7	4.0	74	15.75	28	23
32	1.02	0.99	Na	Na	Na	Na
33 (E22C19W28)	0.07	0.05	Na	Na	5	3
34 (E26C13)	0.2	0.2	52	Na	5	1
35 (E50C23)	0.02	0.01	Na	Na	2	2
36 (E64)	0.002	0.002	Na	Na	Na	Na
W	0.2	0.1	Na	Na	6	Na
Z	33.7	30.8	100.5	5.45	41	34
Total	935.5	903.6	3626	3.88	1739	1471
Autosomes	901.6	872.7	3525	3.91	1692	1437

Table S3 Statistics of chicken genes

Set	Region	Footprint Length		CDS length		Exon number		Exon length	
		Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
Ensembl	Macro	27,068	11,354	1,435	1,059	8.9	6	167	127
Twinscan		20,084	10,689	1187	744	7.0	4	169	130
SGP2		29,804	15,344	1,050	537	7.2	4	146	116
Ensembl	Intermed.	21,756	9,753	1,463	1,106	9.4	7	162	127
Twinscan		15,654	7,408	1,277	780	7.5	4	171	133
SGP2		21,976	10,286	1,062	525	7.3	4	145	118
Ensembl	Micro	15,726	6,362	1,377	1,014	8.8	6.0	165	129
Twinscan		10,683	5,336	1,274	804	7.0	4	182	139
SGP-2		18,815	9,004	1,184	735	8.1	5	145	119
Ensembl	Sex	22,968	12,122	1,360	1,020	8.8	6.5	159	127
Twinscan		23,688	13,028	1,294	843	7.6	5	170	131
SGP2		30,153	14,742	1,047	528	7.2	4	145	118
Ensembl	Other	4,905	1,676	803	579	4.8	4	172	129
Twinscan		13,769	8,042	988	744	5.2	4	192	139
SGP2		18,566	10,384	711	492	5.1	4	140	109
Human		51,078	17,868	1,531	1,173	9.4	7	238	130

Mean and median values are provided for different chromosome types for all three prediction sets. Footprint means maximal extent of the gene; CDS length is the length of coding sequence; exon number is the number of exons per transcript, and exon length is the average length of exons.

Table S4. Chicken-specific retrotransposed regions

	Location				Detected truncation	*d*N/*d*S	Selective Constraint	Parental gene
	Chromosome	Strand	start	end				
GG_retro.1	chr14	+	6374163	6375089		0.0045	Yes	ENSGALT00000020301
GG_retro.2	chr18	-	3898950	3899429		0.006	Yes	ENSGALT00000027292
GG_retro.3	chr5	+	2561934	2562899		0.0044	Yes	ENSGALT00000004965
GG_retro.4	chrZ_random	-	4484728	4485478	+	0.133	Yes	ENSGALT00000011069
GG_retro.5	chr10	+	3530261	3530611		0.0052	Yes	ENSGALT00000015138
GG_retro.6	chrUn	+	147180227	147180754		0.0118	Yes	ENSGALT00000002218
GG_retro.7	chrUn	-	73987437	73987988	+	0.0405	Yes	ENSGALT00000018477
GG_retro.8	chrUn	+	66228829	66229290		0.0026	Yes	ENSGALT00000010093
GG_retro.9	chrUn	+	101288945	101289250		0.1017	Yes	ENSGALT00000021612
GG_retro.10	chr21	-	1156969	1158066		0.0272	Yes	ENSGALT00000023569
GG_retro.11	chr1	-	151237509	151237995	+	1.5658	No	ENSGALT00000014979
GG_retro.12	chrUn	+	155287368	155288051	+	1.0374	No	ENSGALT00000008111
GG_retro.13	chr1	+	173592854	173593225	+	0.756	No	ENSGALT00000000935
GG_retro.14	chr2	-	107892175	107892564	+	0.5215	No	ENSGALT00000000584
GG_retro.15	chr27	+	2029632	2030129	+	0.6738	No	ENSGALT00000021756
GG_retro.16	chr3	+	10484962	10485282	+	1.0136	No	ENSGALT00000012551
GG_retro.17	chr4	-	61957574	61958133	+	0.88	No	ENSGALT00000006839
GG_retro.18	chr5	-	30048292	30048817	+	0.9875	No	ENSGALT00000006839
GG_retro.19	chrUn	-	11909527	11910127	+	99	No	ENSGALT00000006839
GG_retro.20	chrUn	+	149875267	149875479		6.2205	No	ENSGALT00000017188
GG_retro.21	chr1	-	20505024	20505336	+	0.4755	No	ENSGALT00000002864
GG_retro.22	chrZ	+	5911718	5912411	+	0.7359	No	ENSGALT00000005588
GG_retro.23	chrZ_random	-	5620127	5621826	+	0.6583	No	ENSGALT00000006245
GG_retro.24	chr14	-	13775206	13775651	+	0.5094	No	ENSGALT00000007631
GG_retro.25	chr12	-	4647973	4648735	+	0.3055	No	ENSGALT00000009842
GG_retro.26	chr10_random	-	1793161	1793748	+	0.6563	No	ENSGALT00000011117
GG_retro.27	chr1	+	174330282	174330488		0.1063	No	ENSGALT00000011618
GG_retro.28	chr1	-	157083460	157083588	+	0.2954	No	ENSGALT00000011817

GG_retro.29	chr1	+	78261683	78262573	+	0.4844	No	ENSGALT00000014681
GG_retro.30	chr4	-	49726902	49727084		0.5133	No	ENSGALT00000016284
GG_retro.31	chr1	+	143160799	143161170	+	0.8267	No	ENSGALT00000018588
GG_retro.32	chr6	+	15836310	15836855		0.7635	No	ENSGALT00000019737
GG_retro.33	chr2	+	54603296	54603460	+	0.5285	No	ENSGALT00000023608
GG_retro.34	chr23	-	1426378	1426731	+	0.531	No	ENSGALT00000024218
GG_retro.35	chr23	-	273461	273628		0.4103	No	ENSGALT00000024459
GG_retro.36	chrZ_random	+	8977203	8977340		0.599	No	ENSGALT00000024906
GG_retro.37	chr2	+	3258592	3258979	+	0.8831	No	ENSGALT00000026540
GG_retro.38	chr2	+	32345490	32345887	+	0.4717	No	ENSGALT00000027951
GG_retro.39	chr1	+	115993604	115994118	+	0.4172	No	ENSGALT00000005962
GG_retro.40	chr5	-	11696993	11697172		0.5828	No	ENSGALT00000022285
GG_retro.41	chrUn	+	2609116	2609496	+	0.8273	No	ENSGALT00000006376
GG_retro.42	chrUn	+	10297087	10297776	+	0.4081	No	ENSGALT00000009018
GG_retro.43	chrUn	+	11306022	11306476	+	0.6413	No	ENSGALT00000014221
GG_retro.44	chrUn	-	4172075	4173084	+	0.8178	No	ENSGALT00000018455
GG_retro.45	chrUn	-	108441238	108441624	+	0.795	No	ENSGALT00000013263
GG_retro.46	chr14	+	11706517	11706708	+	0.3812	No	ENSGALT00000003816
GG_retro.47	chr1	+	152327678	152328806	+	0.4955	No	ENSGALT00000004589
GG_retro.48	chrZ	-	31494996	31496671	+	0.4967	No	ENSGALT00000004696
GG_retro.49	chr8	+	8703149	8703352	+	0.8400	No	ENSGALT00000006431
GG_retro.50	chr2	+	57663975	57664559	+	0.6895	No	ENSGALT00000011828
GG_retro.51	chr2	-	71836366	71836961	+	0.3597	No	ENSGALT00000022863

Table S5 CpG-islands in chicken and mammalian genomes

Species	CpG-islands	Coverage (Mb)	% Coverage
Chicken (full)	73,381	60.8	5.77
Chicken (masked)	70,655	47.9	4.55
Human (full-Hg16)	260,761	135.2	4.72
Human (masked)	80,350	56.2	1.96
Mouse (full-Mm3)	114,386	61.6	2.46
Mouse (masked)	66,504	37.2	1.48
Rat (full-Rn3)	152,003	70.4	2.74
Rat (masked)	91,015	43.9	1.71

Table S6 Explaining the variability in length ratio of human-chicken orthologous intervals through multivariate regressions on genomic parameters.

Windows considered (number)	Genomic Parameters	Pairwise relations to Log(hL/cL) (including ms bases)		Regression of Log(hL/cL) (including ms bases) on all Genomic Parameters			
		correl	**p-value**	**coeff**	**T (c/se)**	**p-value**	**R-sq**
All chromosomes	Log(hMS/cMS)	0.326	0.000	0.145105	23.76	0.000	22.1%
(6727)	Log(hGap/cGap)	0.156	0.000	0.12107	11.89	0.000	
	Log(dhTel)	0.105	0.000	0.00129	0.28	0.781	
	Log(dcTel)	-0.210	0.000	-0.026194	-8.09	0.000	
	Log(hGC/cGC)(um)	-0.320	0.000	-0.87332	-20.12	0.000	
	Log(hGC/cGC)(ms)	-0.035	0.000	-0.17862	-8.11	0.000	
Macrochromosomes	Log(hMS/cMS)	0.303	0.000	0.175038	19.41	0.000	16.7%
(3781)	Log(hGap/cGap)	0.127	0.000	0.09178	7.26	0.000	
	Log(dhTel)	0.137	0.000	-0.002509	-0.39	0.693	
	Log(dcTel)	-0.051	0.002	-0.003337	-0.63	0.529	
	Log(hGC/cGC)(um)	-0.262	0.000	-0.55983	-8.91	0.000	
	Log(hGC/cGC)(ms)	-0.070	0.000	-0.33526	-8.37	0.000	
Intermediate chromosomes	Log(hMS/cMS)	0.148	0.000	0.11470	6.76	0.000	19.0%
(1227)	Log(hGap/cGap)	0.175	0.000	0.14301	5.67	0.000	
	Log(dhTel)	0.147	0.000	0.04169	3.43	0.001	

	Log(dcTel)	0.150	0.000	-0.03681	-3.39	0.001	
	Log(hGC/cGC)(um)	-0.347	0.000	-0.8715	-8.47	0.000	
	Log(hGC/cGC)(ms)	-0.159	0.000	-0.25410	-5.17	0.000	
Microchromosomes (1501)	Log(hMS/cMS)	0.245	0.000	0.11503	8.60	0.000	25.0%
	Log(hGap/cGap)	0.179	0.000	0.14419	6.15	0.000	
	Log(dhTel)	0.108	0.000	-0.004254	-0.43	0.668	
	Log(dcTel)	-0.102	0.000	-0.02335	-2.32	0.021	
	Log(hGC/cGC)(um)	-0.436	0.000	-1.30069	-15.50	0.000	
	Log(hGC/cGC)(ms)	-0.025	0.341	-0.06394	-1.80	0.072	

Note: The third and fourth columns report pair-wise correlations between the length ratio and each genomic parameter, with corresponding p-values. The following columns summarize linear regressions of the length ratio on all genomics parameters. T ratios (regression coefficients to their standard errors) and corresponding p-values illustrate the importance of each genomic parameter when all parameters are considered in explaining the length ratio. The last column contains the coefficients of determination (i.e. shares of explained variability). The 0.326 correlation between log(hL/cL) (including masked bases) and log(hMS/cMS) is the one reported in the main text. The second correlation reported there (-0.256) is the one between log(hL/cL) computed *on unmasked bases only*, and log(hGC/cGC)(um), which is the GC ratio for unmasked bases – thus, this correlation does not appear in the Table.

Table S7. 57 high-CNF regions.

Human chromosome	Human start	Human end	Human size	NFC	Chicken chromosom
HSA1	60694048	61298482	604435	0.135	GGA8
HSA1	81113474	82151932	1038459	0.141	GGA8
HSA1	87605999	88395498	789500	0.107	GGA8
HSA1	89999508	90878437	878930	0.134	GGA8
HSA1	212880854	215486877	2606024	0.131	GGA3
HSA2	57935365	60781669	2846305	0.116	GGA3
HSA2	66441885	67931055	1489171	0.125	GGA3
HSA2	103295273	104945023	1649751	0.112	GGA1
HSA2	144058338	147593622	3535285	0.131	GGA7
HSA2	156602788	157225116	622329	0.121	GGA7
HSA2	164306081	165539153	1233073	0.128	GGA7
HSA3	17354807	18459444	1104638	0.113	GGA2
HSA3	70046540	71548708	1502169	0.142	GGA12
HSA3	77010147	77619848	609702	0.097	GGA1
HSA3	148207149	149132049	924901	0.120	GGA9
HSA4	23525138	24028705	503568	0.138	GGA4
HSA4	151763241	152359020	595780	0.156	GGA4
HSA4	182967672	183478588	510917	0.108	GGA4
HSA5	3155898	4092375	936478	0.104	GGA2
HSA5	76939886	77484234	544349	0.146	GGA13_random
HSA5	91877971	94313887	2435917	0.122	GGAW
HSA5	157961556	158500224	538669	0.185	GGA13
HSA6	9506498	10047421	540924	0.104	GGA2
HSA6	98120108	99175642	1055535	0.122	GGA3
HSA7	26174051	27393084	1219034	0.105	GGA2
HSA7	68941152	69588865	647714	0.139	GGA19
HSA7	113489625	115130202	1640578	0.128	GGA1
HSA8	76572553	79225945	2653393	0.135	GGA2
HSA8	92925282	93892423	967142	0.124	GGA2
HSA8	106058235	106773378	715144	0.139	GGA2
HSA9	16257430	16859424	601995	0.166	GGAZ
HSA9	123599572	124673395	1073824	0.131	GGA17
HSA10	76554667	78113196	1558530	0.147	GGA6
HSA10	113909660	114686494	776835	0.174	GGA6
HSA10	119143402	119939296	795895	0.102	GGA6
HSA10	129708473	131245417	1536945	0.159	GGA6
HSA11	15951983	16704060	752078	0.182	GGA5
HSA11	30968167	31812716	844550	0.120	GGA3
HSA13	69508342	71050452	1542111	0.118	GGA1
HSA13	76611706	77417924	806219	0.114	GGA1
HSA14	31084271	32263955	1179685	0.171	GGA5
HSA14	55104946	55664543	559598	0.146	GGA5
HSA14	95330236	95952228	621993	0.113	GGA5
HSA15	33966166	35893675	1927510	0.145	GGA5
HSA15	57705539	58235688	530150	0.099	GGA10
HSA15	93004068	96112248	3108181	0.125	GGA10
HSA16	50626650	52571828	1940179	0.096	GGA11

HSA16	53411312	54627661	1216350	0.163	GGA11
HSA16	78090235	79022279	932045	0.114	GGA11
HSA17	35194323	35723334	529012	0.106	GGA19
HSA18	20869286	21543103	673818	0.218	GGA2
HSA18	28495264	29016028	520765	0.098	GGA2
HSA18	70475661	71497943	1022283	0.160	GGA2
HSA18	73893202	74859103	965902	0.092	GGA2
HSA19	35228592	37425166	2196575	0.162	GGA11
HSA20	51561077	52747896	1186820	0.103	GGA20
HSAX	84162709	84857596	694888	0.138	GGA4

Table S8 Telomeres

GGA Chromosome (Mb sequenced)	Search parameters: penta $(TTAGGG)_5$ or di-repeats $(TTAGGG)_2$	No. #	Sequence Position	Location Relative to Human Syntenic Blocks	Num of repe (nt
1 (188.24 Mb)	penta	6	15,214,083	between HSA7 & 22 syntenic blocks	16 (96
			40,274,551	within block of HSA12	11 (66
			69.46 MB CEN		
			145,288,604	within HSA13, but near a single HSAX gene	56 (33
			165,854,078	within HSA13, but adjacent to a single HSA17 gene	10 (60
			184,527,927	between HSA11 & one HSA17 gene, and near 12 block	14 (84
			187,509,463	within HSA11 block	17 (10
2 (147.59 Mb)	di	2	1,098,056	between HSA7 & 3 blocks	3 (18)
			51.10 MB CEN		
			138,597,030	within HSA8	3 (18)
3 (108.64 Mb)	di	4	3,278,889	within HSA20, adjacent to single HSA19 gene	3 (18)
			11.52 MB CEN		
			93,552,018	within HSA2	31 (18
			106,161,627	within HSA6	3 (18)

			106,214,185	within HSA6 (within 40kb of above telo-repeat)	2 (12)
4 (90.63 Mb)	di	2	**18.44 MB CEN**		
			90,477,457	toward end (no syntenic region apparent, at the end?)	9 (54)
			90,519,613	toward end (within 40kb of above telo-repeat)	10 (60
5 (56.31 Mb)	di	0	**3.10 MB CEN**		
Z (33.65 Mb)	di	1	5,836,638	within HSA5	7 (42)
			19.88 MB CEN		
Chr Un (165 Mb)	penta	11	12,663,366		23 (13
			24,902,293		99 (59
			34,393,779		31 (18
			46,286,700		69 (41
			52,847,739		23 (13
			57,516,340		16 (96
			59,138,544		104 (6
			76,211,757		19 (11
			94,312,033		11 (66
			144,220,351		70 (42
			152,443,343		46 (27

Table S9 Sensitivity and specificity for various combination of gene prediction methods

Type	Exons	Sp	Sn	Predicted total
(S&T)!E	13470	46%	3%	206495
(E&S)!T	13818	74%	10%	102253
(T&E)!S	4555	50%	3%	75613
E!(T|S)	41638	89%	14%	266483
(E&S&T)	60721	98%	40%	148766

Table S10 Properties of CpG-islands

	Macrochromosomes	Intermediates	Microchromosomes
Conserved - Overlap gene	4943	2704	6017
Conserved - No gene overlap	2210	1231	2210
Not Conserved - Overlap gene	3756	2218	4766
Not Conserved - No gene overlap	9221	3891	7986
Minimum length (bp)	200	200	200
Median length (bp)	318	291	288
Mean length (bp)	1022	918	804
Maximim length (bp)	80868	64190	42441

Table S11 Chromosomal Distribution of Chicken Segmental Duplications (>90%; >1kb)

		Inter		Intra		Total	
Chrom	Chrom Size	Duplication %	Dup Bp	Duplication %	Dup Bp	Duplication %	Dup Bp
chr1	183734174	3.900407771	7166382	3.050896781	5605540	6.819932148	1253054
chr2	143790626	4.375897911	6292131	4.020471404	5781061	8.286549222	1191528
chr3	105886973	1.958117171	2073391	1.56884549	1661203	3.456657506	366015
chr4	87960023	3.444764902	3030016	2.364479827	2079797	5.703116972	501646
chr5	54035908	3.558174686	1922692	2.863762371	1547460	6.311497532	341047
chr6	33396359	3.693187033	1233390	3.597125663	1201309	6.994651722	233595
chr7	35403612	2.518788196	891742	2.132785208	755083	4.590003415	162502
chr8	28178085	1.992356826	561408	1.488234562	419356	3.429853377	96646
chr9	23053195	3.700363442	853052	3.497588946	806306	7.047938474	162477
chr10	18952620	9.665914264	1831944	7.622297076	1444625	17.08454029	323796
chr11	17998723	15.15822539	2728287	12.35074844	2222977	27.19913518	489549
chr12	19040516	2.222991226	423269	2.473231293	470916	4.437547806	84493
chr13	16795884	7.447086441	1250804	5.762507052	967864	13.07272663	219568
chr14	20156004	5.006850564	1009181	3.577192186	721019	8.48435037	171010
chr15	12220198	2.301722116	281275	2.568624502	313891	4.831026469	59036
chr16	190217	13.93408581	26505	13.09504408	24909	17.16513245	3265
chr17	9892545	6.470357224	640083	6.051182987	598616	12.47456544	123405
chr18	8796964	0.297409424	26163	0.574323141	50523	0.749099348	6589
chr19	9317106	0.085970901	8010	0.159094466	14823	0.236446811	2203
chr20	13294266	1.830398158	243338	1.776216904	236135	3.391281625	45084
chr21	6044380	5.071636793	306549	3.675860882	222183	8.664577674	52372
chr22	2187216	1.356473252	29669	1.040180759	22751	1.475482988	3227
chr23	5031919	0.16246287	8175	0.66058297	33240	0.82304584	4141
chr24	5779840	1.778232615	102779	1.317804645	76167	3.000013841	17339
chr26	3666228	0.083791843	3072	0.081964351	3005	0.165756194	607
chr27	2501417	5.685177641	142210	4.187786363	104754	9.653328493	24147
chr28	4040210	0.662713077	26775	0.599498541	24221	1.262211618	5099
chr32	990147	0.507298411	5023	0	0	0.507298411	502
chrW	4135031	0.39162947	16194	0.416514411	17223	0.785314548	3247

chrZ	30827149	0.174213321	53705	0.550650986	169750	0.654685907	20182
chrUn	121150786	28.55987496	34600513	24.98074012	30264363	51.79586949	6275110
chrE22C19W28	47138	2.796045653	1318	0	0	2.796045653	131
chrE26C13	213449	9.106625002	19438	6.025326893	12861	15.09540921	3222
chrE50C23	10171	76.90492577	7822	22.59364861	2298	76.90492577	782
chrE64	1525	0	0	0	0	0	
chrM	16775	0	0	0	0	0	

Table S12 Proportion of chicken segmental duplications: length vs. %identity (includes random)

% Identity	90-100	90-98	90-99	90-99.5	98-100	99-100	99.5-100
>=1kb	11.051	2.313	8.410	10.784	9.112	2.876	0.357
>=5kb	2.871	0.324	2.419	2.868	2.566	0.460	0.007
>=10kb	0.577	0.090	0.523	0.576	0.488	0.054	0.000
>=20kb	0.037	0.019	0.037	0.037	0.018	0.000	0.000
>=50kb	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table S13 Proportion of chicken segmental duplications: length vs. %identity (excludes random)

% Identity	90-100	90-98	90-99	90-99.5	98-100	99-100	99.5-100
>=1kb	6.349	1.018	4.744	6.194	5.426	1.676	0.179
>=5kb	1.866	0.212	1.565	1.864	1.670	0.306	0.007
>=10kb	0.397	0.059	0.357	0.397	0.340	0.040	0.000
>=20kb	0.029	0.008	0.029	0.029	0.021	0.000	0.000
>=50kb	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Legends for Supplementary Figures

Figure S1. Relative conservation over splice site consensi at intron/exon boundaries. The x-axis shows idealized base position from intron through exon to intron. The gray areas show the regions where expected conservation from the presence of splice site consensi was removed. Unlike inter-mammal comparisons, the chicken-mammal comparison shows a higher relative conservation rate in the splice sites than in the introns.

Figure S2. The Jukes-Cantor substitution level of all interspersed repeats was calculated from the divergence level in the RepeatMasker annotation. No adjustments have been made for CpG content (they are rare in the common elements). The paucity of young repeats probably reflects a lack of source gene definition (if more source genes/consensus sequences were reconstructed, more repeats would very closely match them) and perhaps an under-representation of young CR1 copies in the current assembly. Note the burst of DNA transposon copies with a 16-19% substitution level. They represent invasions of two unrelated elements that probably were active simultaneously and, as is the rule for DNA transposons, for a short period of time. The tapering off of repeats above an 22% substitution level is probably caused by complication in defining these repeat families and by random deletion events. This pattern is similar to that of human and other mammalian genomes,[Lander, 2001 #28;Waterston, 2002 #26] a relatively constant contribution of repeats of different ages. It differs markedly from genomes with a high transposable element activity and high deletion rate of non-functional DNA, like the Fugu genome[Aparicio, 2002 #29], which contain a large number of repeat families with low divergence level and few older repeats.

These data suggest that the overall low interspersed repeat density in the chicken genome is the result of low transposable element activity rather than quick removal of junk DNA.

Figure S3. Distribution of CR1 and L1 insertion sizes. Comparison of the length distribution of the 75,669 LINE-1 copies in the human genome assembly HS16 and 34,067 CR1 copies in chicken assembly GGA6 that are less than 10% diverged from known consensus sequences. Full-length primate LINE-1 and chicken CR1 copies are 6.1-6.3 kb and 4.5-4.6 kb long, respectively. The measurements are based on the 5' position in the consensus of the most 5' fragment and not on the length of the fragment, so that insertions, deletions, rearrangements and sequence gap interruptions do not affect the result. The absence of (near) full-length CR1 copies is striking; only 0.6% of copies are less than 100 bp truncated, compared to 10% of the LINE-1 copies (and 30% of human specific LINE-1 copies are full-length). Assembly issues are unlikely to contribute significantly to this difference, as a similar ratio of full-length copies (8.5% of L1s, 0.5% of CR1s) is observed for copies that are 5-10% diverged and should not interfere with assembly. The data suggests a higher efficiency of the L1 reverse transcriptase or higher stability of the L1 transcript. On the other hand, over 20% of human-specific L1 copies show an inverted 5' end as a result of homology-based secondary priming from the target[68], whereas only 88 of the 200,000 CR1 copies show this pattern. This may be due to the fact that the L1 endonuclease creates long 3' overhangs at the break point, which give rise to the long target site duplications and can function as rogue primers, while CR1-like elements, which do not create target site duplications, presumably does not create such potential primers.

Figure S4. A likely domain accretion for the evolution of a new gene function: A fibrinogen related gene acquired a Scavenger receptor cysteine-rich (SRCR) domain from an exon cassette to form the chicken gene (ENSGALG00000000805 with predicted protein, ENSGALP00000001164).Also shown is a closely-related paralogue (ENSGALG00000000732, proteinENSGALP00000001042) that lacks a SRCR domain, and a Fugu protein(SINFRUP00000129054). The protein-based alignment generated using SMART[26] displays globular domains and introns. The latter are indicated by vertical lines, showing intron phase and amino acid position in the alignment. Black boxes represent gaps in the protein alignment. Coloured boxes correspond to properly aligned regions.

Figure S5. Phylogenetic relationships between 6 human olfactory receptor (OR) sequences (gene names boxed in blue) and 218 chicken ORs (lineages indicated in red). Distances were calculated using the Neighbour-Joining method [69]. The unrooted tree was displayed using TreeView [70]. The monophyly of the chicken OR sequences is supported by a bootstrap value of 100%.

Figure S6. Phylogenetic relationship among ADH sequences. In addition to chicken and human sequences, all dimeric ADHs identified in Fugu (SINFRUP prefixes taken from the ENSEMBL gene set and a fifth, Fr_4230 predicted from the genome) and ADH2 sequences from salamander and ostrich are shown, the latter to narrow the pseudogenisation event of chicken ADH2 (dotted line). This

tree (rooted using the Haemophilus ADH_HAEIN protein) shows that clear 1:1 orthologous relations between chicken and human are limited to class III proteins (ADHX) and also indicates that despite similar copy numbers probably all vertebrate lineages contain independent ADH expansions and losses.

Figure S7. Chicken *CPS1* Gene. **A:** Alignment of the Human *CPS1* gene against the predicted sequence of the Chicken *CPS1* gene (GGA7, + strand, coordinates 2,780,574 – 2,863,870.) The alignment was annotated by Chroma[71] **B:** Chicken-specific tissue expression pattern of *CPS1*. Using primers derived from exons 17 and 18 in the predicted Chicken *CPS1* gene, the expression pattern of *CPS1* across a range of tissues was investigated. It can be seen from the gel image that *CPS1* was found to be expressed in brain, leg muscle, bursal, spleen, and thymus tissue, but not in liver, ovary, or skin tissue

Figure S8. Aligned sequence lengths and chromosome numbers. The logarithms of chromosome length for chromosomes 1-28 decline approximately linearly with chromosome number. Labelled exceptions are chromosomes 16, 22 and 32-36, which may be incomplete. These and the sex chromosomes Z and W are excluded from chromosome size-related comparisons of genetic architecture. Symbols: red=macrochromosomes, black=Intermediates, green=microchromosomes and blue=sex chromosomes.

Figure S9. Small chromosomes tend to be GC-rich. (a) The distribution of GC content in 20 kb non-overlapping windows is plotted for all chromosomes

(yellow), and separately for the different size classes of chromosomes. (b) The distribution of GC content in 20 kb non-overlapping windows for each chicken chromosome is shown as a box plot. The horizontal line within a box indicates the median of the data. The bottom of a box is the first quartile (Q1), i.e. 25% of the data values are less than or equal to this value, and the top of the box is the third quartile (Q3). The whiskers extend to the adjacent value within the lower limit = Q1 - 1.5*(Q3-Q1) or the upper limit = Q3 + 1.5*(Q3-Q1). Outliers (small plus marks) are unusually large or small values beyond the whiskers. For panels a and b, the icons or lines are red for macrochromosomes, black for intermediate chromosomes, green for microchromosomes, and blue for the Z sex chromosome. (c) The distribution of GC content in 20 kb non-overlapping windows for each human chromosome is shown as a box plot. The distributions for X and Y sex chromosomes are in blue and those for autosomes are in black.

Figure S10. Repeat density vs. chromosomal position. The interspersed repeat density in 2 Mb windows displayed against the position of these windows on the four largest chicken chromosomes. Centromeres are located towards the 5' end and repeat density increases with increasing distance from the centromeres.

Figure S11. Telomere characteristics of GGA4. Illustrates that the p-arm possesses an interstitial telomere signal adjacent to the centromere, which may indicate a recent fusion event between a microchromosome and GGA4q in chicken.

Figure S12. Segmental duplications in the chicken genome. a) Distribution of pairwise alignments within the chicken genome. The patterns of large (>5 kb) intrachromosomal (blue) and interchromosomal (red) segmental duplications are shown for chicken chromosomes. b) Degree of sequence identity between segmental duplications**.** The position of the alignments and the % identity (y axis) are shown with respect to a scaled image of each chromosome.

Figure S13. Maps of conserved synteny between chicken chromosomes and human chromosomes. (a) Chicken compared to human. (b) Human compared to chicken.

Figure S14. Chromosome-level mapping of Human/Chicken and Human/Mouse sorted by significance. The headers show the corresponding chromosome names with the number of accommodated orthologous genes in brackets. Each cell shows the number of shared orthologs (with the random expectation in brackets) and the number of shared synteny blocks (with the random expectation in brackets) between each pair of chromosomes, statistically significant similarities are marked by green color and significant dissimilarity by red color.

Figure S15. Scatter plot of conserved non-coding fraction with chicken vs. coding fraction for 5663 non-overlapping 500 kb intervals of the human genome. A high density of non-coding conserved bases is coupled with a low density of coding bases. The correlation coefficient between the two fractions is -0.197 (p-value 0.000). Moreover, although a large majority of intervals in the analysis

presents very low levels of both fractions, the negative association induces a downward average trend (captured non-parametrically by a lowess with smoothing parameter 0.5 -- red curve superimposed to the plot).

Figure S16. GGA1 with terminal and interstitial telomere locations indicated. Telomere sequence positions (Mb) as identified in the draft sequence are aligned with the FISH signals identified at p-terminal and interstitial locations along with q-interstitial and terminal locations. The centromere is shown at 69.46 Mb. GGA1 has 188.24 Mb sequenced not including the centromere (assigned 1.5 Mb) and the telomere regions. The asterisk (*) identifies the most terminal telomeric DNA signal for which sequence data was not apparent.

Figure S17. Comparison of genetic and sequence length. Comparisons for chromosomes 1:28, excluding chromosomes 16, 22, 23, 25, which have insufficient genetic markers or sequence. Symbols: red = macrochromosomes, black = intermediates and green = microchromosomes.