Polymorphism detection

We compare the sequence of 3 domestic chickens to the genome assembly of Red Jungle Fowl (RJF). About one million SNP reads are generated for each of a male broiler (Cornish) from Roslin Institute, a female layer (White Leghorn) from Swedish University of Agricultural Sciences, and a female Silkie from Chinese Agricultural University. DNA for libraries is extracted from the erythrocytes of a single bird, sheared by sonication, and size fractionated using agarose gels. Fragments of 3-kb size are ligated to SmaI-cut blunt-ended pUC18 plasmid vectors. Single colonies are grown overnight, and plasmid DNA is extracted by an alkaline lysis protocol. All sequences are read from both insert ends using vector primers and Amersham MegaBACE 1000 capillary sequencers. For broiler, layer, and Silkie, we get 841,790, 841,555, and 870,556 successful reads with total Q20 lengths of 380,729,199-bp, 372,263,344-bp, and 397,831,117-bp.

The two main issues in polymorphism detection are sequencing errors and paralog confusion. To guard against sequencing errors, we rely on the *Phred* quality Q[1,2]. This is related to the error rate by $-10\times\log_{10}(Q)$. In the first large-scale SNP discovery project[3], a 95% confirmation rate was reported, on a detection rule that required Q>20 at the variant site and Q>15 for the two flanking 5-bp regions. We use more conservative thresholds of 25 and 20 for substitutional polymorphisms, and raise them even further to 30 and 25 for insertion-deletions (indels). The higher thresholds are required to reduce the incidence of artifactual deletions relative to RJF, which upon closer examination of the sequence reads are due to doublet peaks that get called as singlet peaks. Notice that, in a deletion, there is no *Phred* quality for the missing bases in the shorter allele, and the Q30 threshold applies to the mean quality of the two flanking bases. Even so, we still find an excess of deletions over insertions. It is an intrinsic flaw of the base caller software, despite its near universal use in labs worldwide, and there is no easy fix. For our data files, we flag indels in simple repeats, and in our summaries, we do not count them at all. We further advise the users to treat all indels, in this or any other similar project, with due caution.

Paralog confusion is detected in the course of the genome level *BlastN* search that determines where the read is supposed to go. Once this is known, the detailed alignments are done in *CrossMatch*[4], not *BlastN*, because it is more accurate. We require every such alignment to incorporate 80% of the read. A small fraction of the reads will align to more than one region, and in almost every case, the best and second best hits differ by less than 2%, consistent with the finding in the genome paper that segmental duplications are more than 98% identical. Since we know where the duplications are, we use the following rule. If the best and second best hits are more than 2% apart, and the best hit is not to a known segmental duplication, we simply take the best hit. If either of these two rules is violated, we use the clone end pairing information to resolve the ambiguity. In practice, this allows us to salvage about 1.9% of the reads, after which, 7.7, 9.3, and 8.4% of the broiler, layer, and Silkie reads are rejected. The number of bases in the RJF genome that are covered by high quality reads is then 190,513,980-bp, 165,154,746-bp, and 210,214,479-bp. Most of the unusable bases result from our rule that flanking 5-bp regions must be of high quality, since even one low quality base will disqualify 9-bp of data.

Polymorphisms are confirmed by resequencing of PCR amplicons from the line in which they are initially detected. In most cases, we only resequence the domestic chicken (broiler, layer, Silkie) because their alleles are sampled by a single read, where as the RJF allele is on average represented by 6.6 reads. If the resequenced read is heterozygous, it is taken as a confirmation when one of the two alleles concurs with the *in silico* analysis. In those instances where the resequencing confirmation failed, the correct allele was always the RJF allele. An important consideration in all of these resequencing experiments is that the sample for each functional category must be statistically representative. For example, 75% and 25% of coding SNPs are synonymous and non-synonymous. Because we expect different confirmation rates in the two subcategories, we must explicitly check and ensure that the resequenced sample has the same proportions as the full data set.

Polymorphism rates are normalized to the length of the sequence on which we can detect SNPs. To correct for heterozygosity within a line, we compute nucleotide diversity using the approximation[5]: $p = K \Big/ \sum_{i=1}^{n-1} \frac{L}{i}$, where K is the number of variant sites found by sequencing n chromosomes in a region of length L. When comparing RJF to one of the 3 domestic lines, n can only be 2 or 3, and it is a stochastic variable, because there is a 50% chance that any two overlapping reads are from the same chromosome. When there are m overlapping reads, the denominator is $\frac{L}{2^{m-1}} \cdot \left(1 + \left(2^{m-1} - 1\right) \cdot \left(1 + \frac{1}{2}\right)\right)$. We then sum over all possible regions, with different L and m for each region, to get what we call the "effective length". Similar considerations are used to compute SNP rates within a line, except that n is 1 or 2, and as a result, the denominator becomes $\frac{L}{2^{m-1}} \cdot \left(2^{m-1} - 1\right)$.

To be fair, these SNP rates are only meaningful if the shotgun reads are uniformly distributed across the genome. We have already removed reads that align ambiguously to multiple loci. The only other source of potential bias is the library itself, which may over-represent certain classes of sequence, like interspersed repeats. We find that 15.9% of the sequence is identified by *RepeatMasker* as being of transposon origins, but only 14.8% of the usable reads are aligned to these regions. Hence, any bias is negligible.

Functional assessment

We compute gene context relative to 5 different data sets. The first 3 are based on experimentally derived genes and the last 2 are based on computer annotations. Riken1 is a set of 1758 full-length cDNAs taken from bursal B-cells of a two week old CB inbred[6]. GenBank refers to 1178 chicken genes with "complete CDS" designation, downloaded as version 2003-12-15. BBSRC is a set of 1184 cDNAs, taken from a larger group of 18,034 cDNAs[7], which are full-length using a *TBlastX* mapping to vertebrate Refseq and *BlastX* mapping to SWALL. The criterion is that the cDNAs must span the start and stop codons, with E-values below 10^{-25} (*TBlastX* to Refseq) and 10^{-12} (*BlastX* to SWALL). Because we find such similar results in all 3 experimentally derived gene sets, we collapse them into a single non-redundant set, based on where they map to the genome and keeping the largest transcripts. The combined set has 1707 Riken1, 1087 GenBank, and 1074 BBSRC genes. Our last two data sets are 995 chicken orthologs of human disease genes and 17,709 non-redundant Ensembl annotations, from the genome paper.

For the cDNA-to-genome alignments, the initial genome level search is done with *BLAT*[8], but the detailed exon-intron boundaries are determined by *SIM4*[9]. Some fraction of the cDNAs, for example 16.9% of Riken1, will disagree with the genome sequence by a length difference in the coding regions. To define the reading frames, we always use the cDNAs, because cDNA sequencers can easily detect and correct frame shift errors, while genome sequencers cannot. However, we do not accept SNPs and indels on the particular codons where we detect such length differences. In contrast, for substitutional differences between cDNA and genome, we always rely on the genome, because of the expected high error rate from reverse transcriptase used for library construction. Essentially, exon-intron boundaries and reading frames are defined through cDNA sequences, but gene sequences themselves are defined through the reference RJF genome assembly.

Coding regions SNPs are divided into non-synonymous or synonymous, for those that do or do not change the protein. We determine the likelihood that a non-synonymous SNP is functional based on the degree of conservation over all available homologs, using the program *SIFT*[10-12], which has been shown to detect 69% of disease causing mutations, with 20% false positive rates. Homologs are selected from UniProt[13], version dated 2004-02-16, which combines SwissProt, the highly curated protein database, and TrEMBL, the computer translation of the EMBL nucleotide entries not yet in SwissProt. We tested both alleles explicitly, by running *SIFT* twice, and using X as the amino acid at the variant site in the query sequence. The latter step is required to expunge a subtle bias arising from the fact that *SIFT* assumes the query is functional. We also remove homologs more than 95% identical to the query, to prevent the alignment from being contaminated by pseudogenes or chicken sequences with the polymorphism in question.

Additional gene-based SNPs were derived from a 20,067 subset of 85,486 contigs assembled from 330,000 EST reads of chicken cDNAs selected from 21 tissue libraries as previously described[14]. Each of these contigs contains 4 or more ESTs and putative SNPs are identified by *PolyBayes*[15]. From this, we select a high quality subset where each allele is represented by at least 2 ESTs, the *PolyBayes* p-value is less than 0.01, and the *Phrap* quality is more than 30. We also add an indel-filtering step to remove SNPs from regions with alignment gaps. This identified 10,572 high quality SNPs, of which 2,277 map to the Ensembl annotations on a reciprocal top-hits criterion, based on *BlastN* with at least 98% sequence identity over a minimum of 100-bp. Although these ESTs derive from domestic chickens, 2,103 (92%) of them share an allele with RJF. Of these 2,103 SNPs, 424 (20%) are non-synonymous changes that alter the encoded protein sequence.

Genotyping in populations

In order to assess the polymorphism of chicken SNPs across a selection of diverse breeds, 125 SNPs were tested in a selection of 9 different breeds, derived from a previous population study[16] aimed at the characterization of diversity for a wide range of European breeds, including both commercial and fancy breeds. 96 SNPs were previously identified as segregating in a particular breed or cross, based on the sequencing of 8 individuals (32 detected in a layer breed and 64 detected in a broiler breed). There may be a small bias in this data set because markers with very low allele frequencies (0.05-0.10) would not have been selected. However, 29 of these SNPs were unbiased because they were derived from a comparison of two sets of finished BAC sequences, one from the same RJF bird as used in the genome assembly and another from a single White Leghorn bird (Lisa Stubbs, Ivan Ovcharenko, Laurie Gordon, Richard Crooijmans, and Martien Groenen, unpublished). In this unbiased subset, both alleles segregated in 76% of all marker-line combinations. This is comparable to the 73% rate observed for the complete sample set, and it argues that our SNP selection process was not severely biased. Additional details on the markers are kept in the ChickAce database maintained at Wageningen[17].

PCR primers were designed with *Primer3*[18]. SBE primers were designed to have a specific 3'-end 18-25 bp in length. A non-specific 5'-tail was used to create primers 25 to 120 bp in length, at 5 bp intervals to assist multiplexing of 12-16 markers simultaneously.

We used AccuPrime (Invitrogen) kits in the PCR amplification. Multiplex PCR reactions containing 3-6 amplicons were performed in 20 µl containing 60 ng template DNA, 10 µl AccuPrime SuperMix II, and 0.2 µM of each primer. PCR conditions were set to 94°C for 10 min, 41 cycli of 94°C for 30 sec, annealing temperature for 30 sec and 68°C for 3 min, followed by 68°C for 2 min. PCR products were pooled based on SBE primer lengths into 6 super-pools containing 14-16 different fragments. Genotyping was performed using the standard SNaPshot Multiplex Kit (Applied Biosystems) with the following modifications. For the Exo1 treatment, 0.4 µl Exo1 was used, as opposed to 0.2 µl. For the SBE reaction, we used 4 µl Half Big Dye Buffer (GenPak) and 1 µl SNaPshot Ready Reaction Mix. The SBE reaction involved 40 cycli. Genotype detection was performed on a ABI Prism 3100 Genetic Analyzer. The sample preparation protocols used 2 µl SNaPshot product, 8 µl Hi-Di formamide and 0.25 µl GeneScan-120 LIZ size standard. Scoring in Genemapper v3.0 (Applied Biosystems) was confirmed by two independent persons.

Only 12 of the 1125 possible marker-line combinations failed, which means 1113 combinations were analyzed. The failures were entirely due to poor or no amplification of the PCR fragments in the SNP multiplex-assay. Generally, markers failed in only a single population. The sole exception was marker SCW0261, which could only amplify 4 of the 9 lines. We believe that failure of certain markers to amplify particular lines might be due to additional polymorphisms at the primer binding sites.

Domestication analysis

A slightly different SNP set was used to search for selective sweeps, based on the same underlying experimental data, but with relaxed *Phred* quality thresholds. Other than the expected increase in the number of SNPs, the overall rates and characteristics for this second data set were comparable to primary data set. Since this second analysis was done independently, it validates our computational methods.

Sequence reads from the domestic lines were aligned to the RJF assembly using a tool developed at Karolinska Institutet, Stockholm, which allows for rapid and sensitive analysis of extremely large data sets. This tool, named RAT (Rapid Alignment Tool), will be presented elsewhere (Kindlund et al. unpublished). After vector screening and quality trimming, we found 781,638 broiler reads, 770,867 layer reads, and 824,895 Silkie reads. 84% of these sequence reads were aligned to the 111,864 quality trimmed RJF contigs. A best match algorithm resolved any duplicated regions or repeats. This required less than a week to run on our desktop computer. All differences between the sequence reads and the RJF assembly were recorded. Only differences of high quality were considered SNPs and used in further studies. The cutoffs required a *Phred* quality over 20 in the domestic reads and a consensus quality over 20 in the RJF assembly. Overall, 3,924,329 such differences were found. More errors are expected in this SNP set, but it was a necessary compromise as we felt that this analysis would require as many SNPs as possible.

In the following analysis, we considered every possible trio consisting of RJF and 2 of the 3 domestic lines. The chromosomes were traversed with a 100 kb window, which was adjusted in 25 kb steps. Two domestic lines were compared to the RJF assembly at a time, so as to keep the coverage relatively high. Only the SNPs covered by all three lines were considered. Every SNP could be assigned to one of three categories. For example, when comparing broiler and layer with RJF, these categories were: broiler-specific, layer-specific, and RJF-specific. A SNP was broiler-specific when one allele was found only in broiler while the other allele was found both in layer and in RJF. When two or more reads from one domestic line overlapped, only those SNPs where all bases within that line were identical were considered. In such instances, we also relaxed the quality constraint so that only one of the overlapping bases had to exceed the *Phred* threshold of 20, so as to retain as many SNPs as possible. Counts for each category were tallied, and windows with over 10 SNPs and more than 80% in one category were recorded.

References

1. 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).

2. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186-194 (1998).

3. Altshuler, D. et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513-516 (2000).

4. Green, P. *CrossMatch* is the underlying alignment tool for the *Phrap* assembly software at http://www.phrap.org.

5. Cargill, M. et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231-238 (1999).

6. Caldwell, R. et al. A large collection of bursal full-length cDNA sequences to facilitate gene function analysis. *Genome Biol.* (companion issue).

7. Hubbard, S.J. et al. Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags. *Genome Res.* (companion issue).

8. Kent, W.J. BLAT – the BLAST-like alignment tool. *Genome Res.* **12**, 656-664 (2002). http://www.genome.ucsc.edu/cgi-bin/hgBlat.

9. Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M. & Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**, 967-974 (1998). http://globin.cse.psu.edu/html/docs/sim4.html.

10. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863-874 (2001). http://blocks.fhcrc.org/sift/SIFT.html.

11. Ng, P.C. & Henikoff, S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* **12**, 436-446 (2002).

12. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812-3814 (2003).

13. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365-370 (2003).

14. Boardman, P.E. et al. A comprehensive collection of chicken cDNAs. *Curr. Biol.* **12**, 1965-1969 (2002).

15. Marth, G.T. et al. A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23**, 452-456 (1999).

16. Hillel, J., et al. Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. *Genet. Sel. Evol.* **35**, 533-557 (2003).

17. ChickAce database from the Animal Science Group of the Wageningen University and Research Center. https://acedb.asg.wur.nl.

18. Rozen, S. & Skaletsky, H.J. Primer3 (1996,1997,1998). http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

Table captions

Table S1: SNP confirmation rates based on resequencing PCR amplicons in the specific bird where it was detected. We tested 295 SNPs, but the table adds up to more than that because some SNPs appear in multiple rows. L and S indicate if the SNP was from layer or Silkie. If the observed rate (R) is the sum of an actual rate (A) and a noise rate (N), we would expect the confirmation rate A/R to follow the equation 1-N/R.

Table S2: Detailed version of SNP and indel rates given in Table 2. Our data set of 3868 confirmed mRNA transcripts is decomposed into its constituent subsets of 1758 Riken1,

1178 GenBank, and 1184 BBSRC genes. In addition, we show 995 chicken orthologs of human disease genes and 17,709 Ensembl annotations.

Table S3: Poultry breeds used to characterize the SNPs. The observed major and minor allele frequencies, for each of 1113 successfully analyzed marker-line combinations, are given in a separate Excel spreadsheet 1113_marker_population.xls.

Table S4: 3-way comparisons of RJF and all possible combinations of 2 domestic birds from broiler (B), layer (L), and Silkie (S). We use 100 kb segments with at least 10 SNPs (covered by reads from every bird) and count segments where at least 80% of the alleles are shared between two birds but different in the third.

Table S5: *SIFT* analysis for non-synonymous SNPs. The gene totals are non-redundant, insofar as we do not count alternative transcripts. When summing over all 3 lines, we do not count a gene more than once. In contrast, the SNP totals do count SNPs detected in more than one line. All SNPs that change stop codons are assumed to be intolerant, and we explicitly indicate if the intolerant allele is domestic, wild, or both.

Table S1:

		breed	SNPs	confirm	SNP/kb
1	genomewide	S+L	145	94.5%	5.34
2	intron DNA	S	49	91.8%	6.00
3	protein coding	S	56	89.3%	2.34
4	coding SY	S	42	90.5%	8.53
5	coding NS	S	64	82.8%	0.73
6	SIFT intolerant	S+L	70	58.6%	0.08

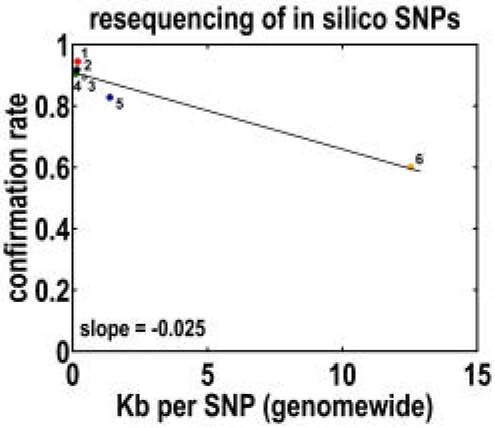

Table S2: RJF-Broiler polymorphisms

	SNP/kb	Indel/kb	SNP/Indel	# of SNP	# of Indel	Aligned Bp	Effective Bp
Riken1 full length cDNAs							
5'-UTR	2.97	0.39	7.6	68	9	22,239	22,868
coding region	2.07	0.07	31.6	884	28	412,647	427,725
non-synonymous (Ka)	0.73						
synonymous (Ks)	7.24						
introns	6.07	0.57	10.7	39,962	3,724	6,344,074	6,580,715
3'-UTR	3.44	0.43	7.9	1,396	176	390,456	406,385
							7,437,693
GenBank with "complete cds"							
5'-UTR	4.52	0.53	8.5	68	8	14,466	15,046
coding region	2.42	0.04	62.1	745	12	297,886	308,147
non-synonymous (Ka)	0.82						
synonymous (Ks)	8.19						
introns	5.19	0.47	11.1	30,774	2,766	5,716,657	5,927,794
3'-UTR	3.10	0.41	7.5	329	44	101,860	106,143
							6,357,129
BBSRC gene collection							
5'-UTR	3.49	0.42	8.3	91	11	25,235	26,098
coding region	1.98	0.03	57.4	287	5	139,828	144,701
non-synonymous (Ka)	0.75						
synonymous (Ks)	7.06						
introns	6.08	0.54	11.2	20,139	1,800	3,197,931	3,309,867
3'-UTR	3.44	0.44	7.8	295	38	82,946	85,723
							3,566,389
Confirmed mRNA transcripts							
5'-UTR	3.45	0.46	7.5	203	27	56,847	58,784
coding region	2.11	0.05	43.2	1,772	41	809,652	838,636
non-synonymous (Ka)	0.73						
synonymous (Ks)	7.44						
introns	5.70	0.52	10.9	86,586	7,915	14,640,319	15,178,325
3'-UTR	3.40	0.42	8.0	1,946	243	550,695	572,391
							16,648,135
Human disease genes							
coding region	2.74	0.04	67.0	1,005	15	354,213	367,226
non-synonymous (Ka)	1.10						
synonymous (Ks)	9.40						
introns	5.36	0.49	10.9	27,768	2,553	5,005,217	5,179,950
							5,547,176
Ensembl (final version 040427)							
5'-UTR	4.22	0.37	11.4	616	54	140,758	146,111
coding region	2.71	0.06	44.3	12,229	276	4,357,256	4,518,133
non-synonymous (Ka)	1.17						
synonymous (Ks)	8.28						
introns	5.64	0.52	10.8	367,361	33,869	62,870,171	65,174,120
3'-UTR	3.92	0.43	9.0	2,130	236	523,238	543,777
							70,382,141
Human-chicken motifs	2.41	0.25	9.6	3,636	379	1,457,199	1,510,505
Genomewide average	**5.28**	**0.48**	**11.0**	**1,041,948**	**94,578**	**190,513,980**	**197,431,517**

Table S2: RJF-Layer polymorphisms

	SNP/kb	Indel/kb	SNP/Indel	# of SNP	# of Indel	Aligned Bp	Effective Bp
Riken1 full length cDNAs							
5'-UTR	3.51	0.34	10.3	82	8	22,668	23,346
coding region	2.31	0.05	49.2	837	17	350,886	362,615
non-synonymous (Ka)	0.79						
synonymous (Ks)	8.31						
introns	6.02	0.52	11.6	32,677	2,805	5,253,151	5,424,176
3'-UTR	3.81	0.41	9.3	1,263	136	321,885	331,549
							6,141,685
GenBank with "complete cds"							
5'-UTR	7.54	0.67	11.3	102	9	13,192	13,519
coding region	2.20	0.04	51.6	619	12	271,987	281,118
non-synonymous (Ka)	0.72						
synonymous (Ks)	7.48						
introns	5.18	0.44	11.8	27,052	2,289	5,053,536	5,222,674
3'-UTR	3.92	0.57	6.9	359	52	88,694	91,530
							5,608,841
BBSRC gene collection							
5'-UTR	3.86	0.45	8.6	77	9	19,368	19,937
coding region	2.28	0.05	47.0	282	6	119,949	123,494
non-synonymous (Ka)	0.66						
synonymous (Ks)	9.10						
introns	5.93	0.47	12.7	17,231	1,362	2,818,899	2,906,830
3'-UTR	3.34	0.41	8.1	235	29	68,278	70,432
							3,120,692
Confirmed mRNA transcripts							
5'-UTR	4.67	0.43	10.9	250	23	52,062	53,545
coding region	2.23	0.05	48.2	1,639	34	711,186	734,283
non-synonymous (Ka)	0.73						
synonymous (Ks)	8.01						
introns	5.67	0.48	11.9	73,431	6,166	12,539,092	12,947,095
3'-UTR	3.77	0.43	8.7	1,793	207	461,652	475,885
							14,210,809
Human disease genes							
coding region	2.33	0.04	56.9	796	14	330,984	342,204
non-synonymous (Ka)	0.74						
synonymous (Ks)	8.33						
introns	5.27	0.47	11.3	24,326	2,145	4,463,554	4,611,630
							4,953,834
Ensembl (final version 040427)							
5'-UTR	4.67	0.31	15.3	626	41	129,661	134,059
coding region	2.58	0.07	36.1	10,373	287	3,882,965	4,012,939
non-synonymous (Ka)	1.10						
synonymous (Ks)	8.05						
introns	5.54	0.48	11.5	312,527	27,122	54,564,092	56,372,291
3'-UTR	4.07	0.45	9.0	1,881	208	447,689	462,368
							60,981,656
Human-chicken motifs	2.23	0.23	9.5	2,901	305	1,257,689	1,298,035
Genomewide average	**5.21**	**0.45**	**11.6**	**889,377**	**76,723**	**165,154,746**	**170,586,544**

Table S2: RJF-Silkie polymorphisms

	SNP/kb	Indel/kb	SNP/Indel	# of SNP	# of Indel	Aligned Bp	Effective Bp
Riken1 full length cDNAs							
5'-UTR	2.82	0.21	13.7	96	7	33,032	34,033
coding region	2.10	0.04	56.8	1,023	18	469,766	486,480
non-synonymous (Ka)	0.63						
synonymous (Ks)	7.69						
introns	6.50	0.63	10.4	46,835	4,504	6,952,043	7,203,446
3'-UTR	3.62	0.51	7.1	1,480	208	394,321	408,396
							8,132,355
GenBank with "complete cds"							
5'-UTR	4.86	0.56	8.6	86	10	17,033	17,701
coding region	2.68	0.05	54.5	927	17	333,653	345,408
non-synonymous (Ka)	0.92						
synonymous (Ks)	9.56						
introns	5.43	0.51	10.7	36,432	3,415	6,477,938	6,713,561
3'-UTR	3.53	0.56	6.3	414	66	112,998	117,386
							7,194,056
BBSRC gene collection							
5'-UTR	4.59	0.32	14.3	143	10	30,080	31,143
coding region	2.45	0.05	50.0	400	8	157,868	163,516
non-synonymous (Ka)	0.80						
synonymous (Ks)	8.75						
introns	6.09	0.57	10.7	22,674	2,127	3,584,360	3,722,140
3'-UTR	3.86	0.58	6.7	368	55	91,816	95,254
							4,012,052
Confirmed mRNA transcripts							
5'-UTR	3.79	0.35	10.8	292	27	74,414	76,944
coding region	2.34	0.05	51.8	2,229	43	919,508	951,926
non-synonymous (Ka)	0.73						
synonymous (Ks)	8.53						
introns	6.00	0.57	10.5	101,164	9,663	16,261,271	16,858,360
3'-UTR	3.66	0.52	7.0	2,195	315	579,035	600,177
							18,487,406
Human disease genes							
coding region	2.41	0.04	55.6	1,000	18	401,637	414,997
non-synonymous (Ka)	0.75						
synonymous (Ks)	8.71						
introns	5.58	0.51	10.9	32,367	2,958	5,600,870	5,800,641
							6,215,638
Ensembl (final version 040427)							
5'-UTR	4.71	0.29	16.5	856	52	174,978	181,730
coding region	2.83	0.07	39.9	14,326	359	4,889,618	5,067,010
non-synonymous (Ka)	1.23						
synonymous (Ks)	8.52						
introns	5.95	0.56	10.6	429,826	40,530	69,716,307	72,274,205
3'-UTR	3.99	0.47	8.5	2,345	275	566,546	587,690
							78,110,636
Human-chicken motifs	2.54	0.30	8.5	4,110	481	1,556,347	1,615,482
Genomewide average	**5.59**	**0.53**	**10.6**	**1,217,817**	**114,822**	**210,214,479**	**217,841,171**

Table S3: Details are in the Excel file 1113_marker_population.xls.

Population		Pop. Type*	Country of origin	Founded **	Population size (range)	# of animals genotyped
Name	**No.**					
White Leghorn	00	C	The Netherlands	1980	500	10
Fayoumi	04	B	Egypt	1978	50-300	10
Marans	13	B	France	1988	200-350	10
Icelandic landrace	16	A	Iceland	900	2000-4000	10
Transsylv. Naked Neck	26	B	Hungary	1990	70-220	10
Green-legged Partridge	27	B	Poland	1950	1600	10
Broiler sire line B	42	D	France	1970	10,000-70,000	10
Brown-egg layer line D	45	C	The Netherlands	1962	1000	10
Broiler dam line D	50	D	Middle East	1970	5000-20,000	10

*Population Types: A - domesticated unselected breed, B - standardized breed selected on morphology, C - Layers, selected on quantitative traits, D - Broilers, selected on quantitative traits. **Estimated year that the sampled line was established.

Table S4:

	>10 SNPs within	>80% shared alleles			>80% shared alleles		
	100 kb segment	1=2, not 3	1=3, not 2	2=3, not 1	1=2, not 3	1=3, not 2	2=3, not 1
RJF(1)-B(2)-L(3)	34,089	637	600	497	1.9%	1.8%	1.5%
RJF(1)-B(2)-S(3)	36,098	419	610	310	1.2%	1.7%	0.9%
RJF(1)-L(2)-S(3)	34,907	916	457	139	2.6%	1.3%	0.4%

Table S5:

	# of genes		# of SNPs			HIGH confidence				LOW confidence		
		w/NS		SIFT	intolerant	intolerant	intolerant	tolerant	intolerant	intolerant	intolerant	tolerant low
	TOTAL	SNPs	TOTAL	done	wild RJF	domestic	both way	high conf	wild RJF	domestic	both way	conf
Confirmed mRNA transcripts												
Broiler	3,868	269	460	382	14	27	4	233	11	18	14	61
Layer	3,868	274	411	347	12	18	3	191	12	18	26	67
Silkie	3,868	321	535	445	10	37	4	227	22	21	29	95
TOTAL	3,868	657	1,406	1,174	36	82	11	651	45	57	69	223
		17.0%		**83.5%**	**3.1%**	**7.0%**	**0.9%**	**55.5%**	**3.8%**	**4.9%**	**5.9%**	**19.0%**
Human disease genes												
Broiler	995	127	286	255	5	11	1	192	4	8	9	25
Layer	995	129	196	175	5	16	2	115	4	8	6	19
Silkie	995	147	232	189	7	17	4	99	11	7	2	42
TOTAL	995	283	714	619	17	44	7	406	19	23	17	86
		28.4%		**86.7%**	**2.7%**	**7.1%**	**1.1%**	**65.6%**	**3.1%**	**3.7%**	**2.7%**	**13.9%**
Ensembl (final version 040427)												
Broiler	17,709	2,038	4,236	2,792	101	221	78	1,394	93	164	177	564
Layer	17,709	1,967	3,548	2,375	81	188	44	1,167	80	132	172	511
Silkie	17,709	2,498	5,035	3,353	107	255	62	1,676	129	166	228	730
TOTAL	17,709	4,820	12,819	8,520	289	664	184	4,237	302	462	577	1,805
		27.2%		**66.5%**	**3.4%**	**7.8%**	**2.2%**	**49.7%**	**3.5%**	**5.4%**	**6.8%**	**21.2%**
BBSRC coding SNPs												
TOTAL	2,095	372	424	359	7	43	1	165	9	33	25	76
		17.8%		**84.7%**	**1.9%**	**12.0%**	**0.3%**	**46.0%**	**2.5%**	**9.2%**	**7.0%**	**21.2%**