LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-209903

A genetic variation map for chicken with 2.8 million single nucleotide polymorphisms

G. K-S. Wong, L. Hillier, M. Brandstrom, R. Croojmans, I. Ovcharenko, L. Gordon, L. Stubbs, S. Lucas, T. Glavina, P. Kaiser, U. Gunnarsson, C. Webber, I. Overton

February 23, 2005

Nature

A genetic variation map for chicken with 2.8 million single nucleotide polymorphisms

International Chicken Polymorphism Map Consortium

Summary

We describe a genetic variation map for the chicken genome containing 2.8 million single nucleotide polymorphisms (SNPs), based on a comparison of the sequences of 3 domestic chickens (broiler, layer, Silkie) to their wild ancestor Red Jungle Fowl (RJF). Subsequent experiments indicate that at least 90% are true SNPs, and at least 70% are common SNPs that segregate in many domestic breeds. Mean nucleotide diversity is about 5 SNP/kb for almost every possible comparison between RJF and domestic lines, between two different domestic lines, and within domestic lines – contrary to the idea that domestic animals are highly inbred relative to their wild ancestors. In fact, most of the SNPs originated prior to domestication, and there is little to no evidence of selective sweeps for adaptive alleles on length scales of greater than 100 kb.

Keywords: chicken, polymorphism, domestication, selection

Introduction

The generation of a high quality draft sequence for the genome of chicken (*Gallus gallus*) is an important advance[1]. Chickens are good models for studying the genetic basis of phenotypic traits, because of the extensive diversity among domestic chickens selected for different purposes. Monogenic traits are well-studied[2-4], but many interesting traits are complex and determined by an unknown number of genes. Quantitative trait loci (QTLs) have been mapped for a range of traits, including ones for growth, body composition, egg production, antibody response, disease resistance, and behaviour[5]. Determining causative genes is difficult, since each locus controls only a fraction of the phenotypic variance. We will describe a survey of the genetic variation between 3 domestic chickens and their wild ancestor. The 2.8 million single nucleotide polymorphisms (SNPs) that we identified will facilitate mapping of complex traits in many ways. First, improved marker density allows researchers to take advantage of the higher recombination rates in chicken[1], which are 2.5 to 21 cM/Mbp depending on the chromosome, compared to 1 cM/Mbp in human, and 0.5 cM/Mbp in mouse. The previous linkage map used 2000 markers[6,7], but only 800 of these were microsatellites or SNPs, which are the most useful[8]. More importantly, our new data allow researchers to construct detailed haplotypes that segregate in different QTL crosses. Because any mutation underlying a QTL must once have originated from a single founder animal, haplotype comparisons will facilitate the fine mapping of QTLs[9]. To this end, we conduct a genomewide search for evidence of selection due to domestication, and provide an initial characterization of the expected magnitude of these effects.

Genetic variation and utility

Our experiment is outlined in Figure 1. SNPs are generated by partial sequencing at ¼ coverage for each of 3 domestic breeds (a male layer, a female broiler, and a female Silkie), and comparison of the resultant reads to the 6.6x genome for the wild ancestor of domestic chickens, Red Jungle Fowl (RJF). We expect marked heterozygosity within the

3 domestic lines, but not within RJF because the sequenced bird for the genome project is from a highly inbred line that is essentially homozygous.

Comparing the sequence reads for broiler, layer, and Silkie to the genome of RJF, we identified nearly a million SNPs in each instance, at mean rates of about 5 SNP/kb, as shown in Table 1. Notice that all of the "SNP rates" quoted in this paper are computed as *nucleotide diversities* (π), and given in units of $\pi \times 10^3$. After correcting for SNPs detected in more than one line, there are 2,833,578 variant sites, or one potential marker every 374 bp along the 1.06 Gb genome. To assess the reliability of these data, we resequenced 295 SNPs in the exact bird in which it was detected (Table S1). As many as 94% of the SNPs were confirmed. However, confirmation rates are sensitive to the functional context (*e.g.,* coding versus non-coding) and SNPs in rare categories are less likely to be confirmed. In fact, only 83% of the non-synonymous SNPs were confirmed. Small indels of a few base pairs in length (mean of 2.3 and median of 1) are detected at rates that are well correlated with the corresponding SNP rates, but smaller by about a factor of 10.

Chicken autosomes are sorted by size into 5 large macrochromosomes (GGA1-5), 5 intermediate chromosomes (GGA6-10), and 28 microchromosomes (GGA11-38). SNP and indel rates are independent of chromosome size, as shown in Figure 2. GGA16 is the sole exception, because it contains the highly variable MHC[10]. This result is surprising, as recombination rates on microchromosomes are much higher than on macrochromosomes[1] and studies in other organisms exhibit a positive correlation between recombination rates and polymorphism rates[11-12]. We expect that higher gene densities on microchromosomes likely counteract the effect of higher recombination rates.

SNP rates between and within chicken lines can be determined from the overlaps between reads. Table 1 demonstrates that almost every pairwise combination gives a SNP rate of just over 5 SNP/kb, except for broiler-broiler and layer-layer, which show about 4 SNP/kb, as expected since the sequenced broiler and layer are from closed breeding lines. To ensure that there are no confounding factors from the single read nature of our data, or the complexities of the overlap analysis, we used comparisons to 3.8 Mb of finished BAC

sequence of a different White Leghorn[13] from the same breed but not the same line as the layer sequenced herein. 15 chromosomes were sampled, and the results confirm our rates of 5 SNP/kb. In another study of 15 kb of introns in 25 birds from 10 divergent breeds of domestic chickens[14], an autosomal rate of 6.5 SNP/kb was reported.

To quantify SNP and indel rate variation versus functional context, we considered three gene sets representing 3868 confirmed mRNA transcripts, 995 chicken orthologs of human disease genes, and 17,709 Ensembl annotations from the RJF analysis[1]. Complete details for all 3 lines are tabulated in the supplements (Table S2). An excerpt for broiler is shown in Table 2. Within genes defined by mRNA transcripts, the SNP rates are 3.5, 2.1, 5.7, and 3.4 SNP/kb in 5'-UTR, coding exon, intron, and 3'-UTR regions respectively. In coding regions, indel rates are 43 times smaller than SNP rates. Ka/Ks is 0.098, similar to what is typically seen in vertebrate comparisons. We also studied "conserved non-coding regions" from the RJF analysis[1]. SNP rates are similar to those of coding exons, but indel rates are intermediate to those of coding exons and UTRs, which supports the notion that these regions are functional, but may not encode proteins.

Utility of these SNPs depends on their frequency of occurrence in commonly used chicken populations. Hence, we typed 125 SNPs (including coding and non-coding SNPs, randomly distributed across the chicken genome) in 10 unrelated individuals from each of 9 divergent lines representing an assortment of European breeds. This collection includes commercial broiler and layer breeds, standardized breeds selected for their morphological traits, and an unselected breed from Iceland (Table S3). Both alleles segregated in 73% of 1113 successful marker-line combinations (out of 1125 possible combinations). Averaged minor allele frequency is 27%, but it decreases to 20% if marker-line combinations where one of the two alleles is fixed are included. This indicates that a majority of the SNPs are common variants that predate the divergence of modern breeds. Only 12% of the markers had a minor allele frequency of less than 10% in the 90 animals tested.

We demonstrate by example how these data can be used to target specific genome regions. Details of our experiments are in Supplement E (Examples). First, we consider a

body weight related QTL on GGA4 that was previously mapped to a 150 cM interval[15,16]. After a year of effort, where every known microsatellite (>50) was tested, 26 informative markers were developed. Further progress would have required the laborious sequencing of multiple chickens to find additional polymorphisms in this target region. With the SNP map, we selected 47 random broiler-layer SNPs, and ABI SNPlex assays were developed to type an experimental F_2 cross (n = 466). 28 (60%) of these SNPs were informative, but none had breed specific alleles, confirming that most variations predate domestication. In just one month, we doubled the number of markers, and resolved the initial QTL into two QTLs that affect body weight at 3 and 9 weeks of age.

In addition to providing markers for fine mapping, these SNPs are a rich source of polymorphisms that are candidates for the causative differences in important traits. Some of the candidate genes in disease resistance might include TGF-β[17,18], cytokines[19], and the major histocompatibility complex (MHC). As an example, 40 SNPs were identified from the SNP map in the coding and promoter regions of 12 cytokine genes. Typed on 8 inbred layer lines, 32 of these SNPs were informative. Cytokine genes on GGA13, including *IL4* and *IL13*, two genes that are expressed in T helper-2 (Th2) cells, drive antibody response. Four of the six SNPs that were polymorphic among lines were in *IL4* and *IL13*, and these SNPs were fixed for different alleles in lines N and 15I, which show differential antibody response to vaccination[20]. Thus, these SNPs will now allow us to test whether the *IL4* and *IL13* loci directly determine the observed differential antibody response.

Domestication and selection

Domestic animals are useful models of phenotypic evolution under selection. The challenge is to find not only those loci that determine phenotypic differences, but also the causative alleles. We adopted two approaches, searching for evidence of selective sweeps and for non-synonymous amino acid substitutions at highly conserved sites. One example of a selective sweep is the *IGF2* locus in pigs[22]. Given the available data, determining the exact haplotype structure is difficult, because blocks of shared alleles can be erroneously disrupted by heterozygosity of the domestic lines and by sequencing errors. However, we

can search for the local reductions in heterozygosity that accompany selective sweeps[9,21], as long as we are mindful of the sequencing error rate.

We did 3-way comparisons of RJF and all possible combinations of two domestic lines. Given the limited coverage of the latter, we only examined 100 kb segments with at least 10 SNPs. In practice, each had an average of 25 to 28 SNPs. We then computed how often 80% or more of the SNP sites are identical in the two domestic lines but different in RJF. In Table S4, we show that 0.4 to 1.5% of the segments qualified. However, when we searched for shared alleles between RJF and one of the domestic lines, 1.2 to 2.6% of the segments qualified. Heterozygosity of the domestic lines is more of a confounding factor in searching for blocks of shared alleles between two domestic lines, versus between RJF and one domestic line. This could explain the difference, but if so, then heterozygosity of the domestic lines is the dominant factor in determining such blocks of shared alleles, not selective sweeps. Hence, selective sweeps that occurred before the divergence of modern domestic breeds must have left behind footprints that are much smaller than 100 kb. This would however be entirely consistent with the historically large effective population size of domestic chickens, and the reported high recombination rates.

For a glimpse of the true haplotype patterns, one can compare the aforementioned 3.8 Mb of finished BAC sequence, from the second layer line (L2), to the genome of RJF. These results are overlaid alongside the primary SNP data set in Figure 3. Short RJF-type fragments can be seen in all 4 lines. Shared domestic-type fragments can also be seen, but at sizes of 5 to 15 kb. This is consistent with our inability to detect footprints of selective sweeps at length scales of 100 kb and suggests that a better choice is 10 kb. However, our data are insufficient for such a genomewide analysis.

It has been hypothesized that loss-of-function (LOF) mutations have accumulated in domestic animals, as the result of relaxed purifying selection and selection for adaptive benefits[23]. An example of the latter is the deletion in the myostatin gene in cattle selected for muscularity[24]. Such deletions are rare, and so we looked for non-synonymous SNPs at highly conserved sites using *SIFT*[25]. Every substitution is thus classified as being likely to

affect function (intolerant) or not (tolerant). For genes defined by mRNA transcripts, 26% of testable SNPs are intolerant, although only 11% are intolerant if we restrict this to high confidence assessments (Table S5). Usually, it is the domestic allele that is intolerant, but we would emphasize that intolerant SNPs are rare, and only 59% were confirmed by PCR resequencing. Given that the domestic allele is represented by a single read, as opposed to 6.6 for the wild allele, much of this effect is likely due to sequencing errors. However, we noticed the same effect in 424 non-synonymous SNPs that we identified from an analysis of 330,000 ESTs, where every allele was seen in two or more ESTs. We conclude that the LOF hypothesis remains intriguing, but any effect is likely to be small.

Some of the experimentally confirmed *SIFT* intolerant SNPs could be functionally important. We show one example in Figure 4, from the ornithine transcarbamylase (*OTC*) gene. It substitutes glycine in RJF to arginine in layer and broiler. This SNP is identical to the G188R substitution associated with hyperammonaemia in humans[26]. Resequencing of additional domestic birds revealed a high frequency for the intolerant allele in both White Leghorns (p=0.65, n=20) and in broilers (p=0.75, n=6). In mammals, *OTC* is expressed in the liver and catalyzes the second step of the urea cycle. Chicken *OTC* is expressed in the kidney and exhibits a low enzymatic activity, with substantial variability among breeds[27]. Preservation and sequence conservation of *OTC*, along with all other enzymes in the urea cycle[1], was unexpected because avian species excrete uric acid (not urea) as their primary component of nitrogenous waste, and were believed to be lacking a functional urea cycle. The deleterious nature of human G188R makes this an attractive candidate for phenotypic studies of avian-specific adaptations in the urea cycle.

Discussion

This analysis has provided the first global assessment of nucleotide diversity for a domestic animal in comparison to a representative of its wild ancestor. The small number of birds sequenced is compensated for by the vast number of sites examined. We detected surprisingly little difference in diversity in comparisons between RJF and domestic lines, between different domestic lines, and within domestic lines. The total rates are typically 5

SNP/kb, with the only exception being a slight reduction to 4 SNP/kb in broiler and layer lines that are maintained as closed breeding populations. In comparison, 5 SNP/kb is 6 to 7-fold larger than humans[28] and domestic dogs[29], 3-fold larger than gorillas[30], but similar to the diversity between different mouse subspecies[31].

Most of the nucleotide diversity observed between and within domestic lines must have originated prior to the domestication of chickens 5,000 to 10,000 years ago. Given a neutral substitution rate of 1.8×10^{-9} sites per year for galliform birds[32], we estimate that a coalescence time of 1.4 million years would be required to account for the observed rates of 5 SNP/kb. Considering that the rates observed between RJF and domestic lines are not much higher than those between domestic lines, it would seem that domestication has not resulted in a substantial genomewide loss of diversity, as would be expected had a severe population bottleneck occurred. This is important, because it contradicts the assertion that animal domestication began from a small number of individuals in a restricted geographic region[33]. That is still a possible scenario for the very earliest phases of domestication, but if so, our data imply that subsequent crossing with the wild ancestor (in the first thousand years until more developed breeds were established) restored this diversity. Nevertheless, extensive diversity is consistent with the ongoing improvements in agricultural traits that have been achieved over the last 80 years, in layer and broiler lines[34].

These 2.8 million SNPs, of which 70% or more are polymorphic in widely studied chicken populations, are already being used to analyze specific QTL regions or candidate genes, to greatly improve efficiency. However, their ultimate value will be realized when detailed haplotypes that segregate in any particular QTL cross are constructed. The small haplotype blocks detected by this study underscore the need for a larger number of SNPs, from which informative markers can be selected. These data can also be used for *in silico* detection of functional SNPs, although one must be more cautious of sequencing errors in this instance. Finally, despite our failure to detect evidence of selective sweeps on the 100 kb scale, we firmly believe that they exist at smaller scales, and that they will be found by the resequencing of target regions for major trait loci.

Materials and methods

Our broiler and layer lines are from European breeds with dramatic differences in meat and egg production traits. This specialization started only in the first half of the 20th century[35]. The sequenced male White Cornish-type broiler is from a closed line breeding population commonly used in the production of commercial meat-type hybrids (Aviagen, Newbridge, Scotland); effective population size is about 800. The female White Leghorn layer is from a closed line developed at Swedish University of Agricultural Sciences[36]; its effective population size has been 60 to 80 birds for the past 30 years. The Chinese Silkie is used in meat/egg production and traditional Chinese medicine[37]. Selection intensity has been low, and the sequenced female is from a large outbred population.

DNA was extracted from erythrocytes of a single bird, sheared by sonication, and size fractionated via agarose gels. Fragments of 3-kb size were ligated to SmaI-cut blunt-ended pUC18 plasmid vectors. Single colonies were grown overnight, and plasmids were extracted by an alkaline lysis protocol. Sequences were read from both ends of the insert, with vector primers and Amersham MegaBACE 1000 capillary sequencers. Roughly one million reads were generated for each bird. For broiler, layer, and Silkie, we got a total of 841,790, 841,555, and 870,556 successful reads, whose Q20 lengths add to 380,729,199-bp, 372,263,344-bp, and 397,831,117-bp respectively.

To minimize sequencing errors, we use the *Phred* quality, Q[38,39]. This is related to the single base error rate by the equation: $-10\times\log_{10}(Q)$. We use more stringent thresholds than normal[40], with Q>25 for the variant site and Q>20 in both flanking 5-bp regions. For an insertion-deletion (indel), the variant site in the shorter allele is given the quality of its two flanking bases. We originally found many artifactual deletions relative to RJF, which upon a closer examination of the sequence reads were due to doublet peaks that got called as singlet peaks. This is an unavoidable flaw of the base caller software. Hence, we raised the indel thresholds to Q30 and Q25. We must still advise caution, and to that end, indels in simple repeats are flagged and none are counted in our summary tables.

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 within *CrossMatch*[41]. Analysis of the RJF genome[1] shows that recent segmental duplications typically agree to 2%. When the best and second best *BlastN* hits were more than 2% apart, and the best hit was not to a known segmental duplication, the best hit was taken. When either rule was violated, clone-end pairs information was used to resolve the ambiguity. Every alignment had to incorporate 80% of the read. Mapped back to the RJF genome, the amount of usable data for broiler, layer, and Silkie covered 190,513,980-bp, 165,154,746-bp, and 210,214,479-bp respectively.

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[42]: $\pi = K \Big/ \sum_{i=1}^{n-1} \frac{L}{i}$, 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 $\frac{L}{2^{m-1}} \cdot \left(2^{m-1} - 1\right)$.

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[43]. 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[44], which are full-length using a *TBlastX* mapping to vertebrate Refseq and *BlastX* mapping to SWALL. Merging all 3 data sets, we have 3868 non-redundant genes. For the

detailed gene models, we do a genome level search in *BLAT*[45] and use *SIM4*[46] to compute the exon-intron boundaries. The last two data sets are for 995 chicken orthologs of human disease genes and 17,709 non-redundant Ensembl genes.

Additional details are in Supplement M (Methods).

Corresponding authors: Gane Ka-Shu Wong gksw@genomics.org.cn, Leif Andersson leif.andersson@imbim.uu.se, HuanMing Yang hyang@genomics.org.cn.

The individual SNPs were deposited at GenBank/dbSNP with submitted SNP (ss) number ranges: 24821291 to 24922086, 24922088 to 26161960, 26161962 to 28446123, and 28452569 to 28452598. They are also available from http://chicken.genomics.org.cn, the UCSC genome browser, and the Ensembl genome browser. Access to raw sequencing traces is being provided through the NCBI Trace Archive.

Acknowledgments

Beijing Institute of Genomics of Chinese Academy of Sciences Gallus gallus SNP discovery and analysis was supported by Chinese Academy of Sciences (KSCX2-SW-223), State Development Planning Commission, Ministry of Science and Technology (2002AA104250; 2004AA231050; 2001AA231061; 2001AA231101), National Natural Science Foundation of China (30200163; 90208019), Beijing Municipal Government, Zhejiang Provincial Government, Hangzhou Municipal Government, Zhejiang University, and China National Grid. Some equipment and reagents were provided by Wellcome Trust and Sanger Institute of the UK. Recent segmental duplications were analyzed by G. Cheng and E.E. Eichler. Riken1 cDNAs were provided by R. Caldwell and J.M. Buerstedde. Noncoding conserved motifs were analyzed by J. Taylor and W. Miller. ***Washington University School of Medicine*** Gallus gallus sequence generation was supported by National Human Genome Research Institute. ***Uppsala University*** HE was supported by Swedish Research Council, Knut and Alice Wallenberg Foundation, and Royal Academy of Sciences. LA was supported by Wallenberg Consortium North,

Foundation for Strategic Research, and Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning. ***Institute for Animal Health*** PK, NB, JRY, and JK were supported by BBSRC. ***Iowa State University*** SJL was supported by Hatch Act and State of Iowa. Skeletal data for ISU resource population was collected by C. Ashwell and A. Mitchell. ***Roslin Institute*** PMH, AL, DJK, and DWB were supported by BBSRC. SNP genotyping was partially funded by Cobb-Vantress. ***USDA-ARS Avian Disease and Oncology Laboratory*** J. Kenyon and N. Evenson provided technical assistance. ***University of Oxford*** CPP was supported by UK Medical Research Council. ***University of Manchester Institute of Science and Technology*** SJH was supported by BBSRC. ***University of Sheffield*** SAW was supported by BBSRC.

We dedicate this paper to Nat Bumstead, who died during preparation of the manuscript. Nat was recognised as a major figure in researching the genetics of disease resistance in poultry. He worked tirelessly to realise the sequence of the chicken genome, which led in part to this consortium.

This work was performed under the auspices of U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

References

1. The Chicken Genome Sequencing Consortium. The sequence of the chicken genome, *Gallus gallus*. *Nature* (companion paper).

2. Pisenti, J.M. et al. Avian genetic resources at risk: an assessment and proposal for conservation of genetic stocks in the USA and Canada. *Avian Poult. Biol. Rev.* **12**, 1-102 (2001).

3. Dodgson, J.B. & Romanov, M.N. Use of chicken models for the analysis of human disease. In *Current Protocols in Human Genetics* (eds. Dracopoli, N.C. et al.) 15.5.1-11 (John Wiley & Sons, Hoboken, 2004).

4. Nicholas, F.W. Online Mendelian Inheritance in Animals (OMIA): a comparative knowledgebase of genetic disorders and other familial traits in non-

laboratory animals. *Nucleic Acids Res.* **31**, 275-277 (2003). http://www.angis.org.au/Databases/BIRX/omia.

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

6. Groenen, M.A. et al. A consensus linkage map of the chicken genome. *Genome Res.* **10**, 137-147 (2000).

7. Groenen, M.A. & Crooijmans, R.P. Structural genomics: integrating linkage, physical and sequence maps. In *Poultry Genetics, Breeding and Biotechnology* (eds. Muir, W.M. & Aggrey, S.E.) 497-536 (CABI Publishing, Wallingford, 2003).

8. Vignal, A., Milan, D., SanCristobal, M. & Eggen, A. A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.* **34**, 275-305 (2002).

9. Andersson, L. & Georges, M. Domestic-animal genomics: deciphering the genetics of complex traits. *Nat. Rev. Genet.* **5**, 202-212 (2004).

10. There are only 20-kb of aligned sequence on GGA16, and if we were to remove it, the total SNP rate would only change by 0.02%.

11. Begun, D.J. & Aquadro, C.F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* **356**, 519-520 (1992).

12. Nachman, M.W. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.* **9**, 481-485 (2001).

13. Crooijmans, R.P., Vrebalov, J., Dijkhof, R.J., van der Poel, J.J. & Groenen, M.A. Two-dimensional screening of the Wageningen chicken BAC library. *Mamm. Genome* **11**, 360-363 (2000).

14. Sundstrom, H., Webster, M.T. & Ellegren, H. Reduced variation on the chicken Z chromosome. *Genetics* **167**, 377-385 (2004).

15. Ikeobi, C.O. et al. Quantitative trait loci affecting fatness in the chicken. *Anim. Genet.* **33**, 428-435 (2002).

16. Sewalem, A. et al. Mapping of quantitative trait loci for body weight at three, six, and nine weeks of age in a broiler layer cross. *Poult. Sci.* **81**, 1775-1781. (2002).

17. Li, H. et al. Chicken quantitative trait loci for growth and body composition associated with transforming growth factor-β genes. *Poult. Sci.* **82**, 347-356 (2003).

18. Zhou, H., Li, H. & Lamont, S.J. Genetic markers associated with antibody response kinetics in adult chickens. *Poult. Sci.* **82**, 699-708 (2003).

19. Gallagher, G., Eskdale, J. & Bidwell, J.L. Cytokine genetics - polymorphisms, functional variations and disease associations. In *The Cytokine Handbook, 4th Edition* (eds. Thomson, A.W. & Lotze, M.T.) 19-55 (Academic Press, London, 2003).

20. Bumstead, N. et al. *EU Project FAIR3 PL96-1502 New Molecular Approaches for Improved Poultry Vaccines* (Institute for Animal Health, Compton, 2000).

21. Maynard-Smith, J. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23-35 (1974).

22. Van Laere, A.S. et al. regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* **425**, 832-836 (2003).

23. Olson, M.V. When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.* **64**, 18-23 (1999).

24. Grobet, L. et al. A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. *Nat. Genet.* ***17***, 71-74 (1997).

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

26. Gilbert-Dussardier, B. et al. Partial duplication [dup. TCAC (178)] and novel point mutations (T125M, G188R, A209V, and H302L) of the ornithine transcarbamylase gene in congenital hyperammonemia. *Hum. Mutat.* **8**, 74-76 (1996).

27. Tamir, H. & Ratner, S. Enzymes of arginine metabolism in chicks. *Arch. Biochem. Biophys.* **102**, 249-258 (1963).

28. Sachidanandam, R. et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933 (2001).

29. Parker, H.G. et al. Genetic structure of the purebred domestic dog. *Science* **304**, 1160-1164 (2004).

30. Yu, N., Jensen-Seaman, M.I., Chemnick, L., Ryder, O. & Li, W.H. Nucleotide diversity in gorillas. *Genetics* **166**, 1375-1383 (2004).

31. Lindblad-Toh, K. et al. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat. Genet.* **24**, 381-386 (2000).

32. Axelsson, E., Smith, N.G., Sundstrom, H., Berlin, S. & Ellegren, H. Male-biased mutation rate and divergence in autosomal, Z-linked and W-linked introns of chicken and turkey. *Mol. Biol. Evol.* **21**, 1538-1547 (2004).

33. Mason, I.L. (ed.) *Evolution of Domesticated Animals* (Longman, Inc., New York, 1984).

34. Arthur, J.A. & Albers, G.A. Industrial perspective on problems and issues associated with poultry breeding. In *Poultry Genetics, Breeding and Biotechnology* (eds. Muir, W.M. & Aggrey, S.E.) 1-12 (CABI Publishing, Wallingford, 2003).

35. Crawford, R.D. (ed.) *Poultry Breeding and Genetics* (Elsevier Science, New York, 1990).

36. Liljedahl, L.E., Kolstad, N., Sorensen, P. & Maijala, K. Scandinavian selection and cross-breeding experiment with laying hens. 1. Background and general outline. *Acta Agricult. Scand.* **29**, 273-285 (1979).

37. Niu, D. et al. The origin and genetic diversity of Chinese native chicken breeds. *Biochem. Genet.* **40**, 163-174 (2002).

38. 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).

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

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

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

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

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

44. 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).

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

46. 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.

Figure captions

Figure 1: SNP discovery experiment. We sampled 3 domestic chickens at 1/4 coverage each and compared the resultant sequence to the 6.6x draft genome of Red Jungle Fowl (RJF). Chicken photographs shown here are provided by Bill Payne (RJF), Paul Hocking (broiler), Leif Andersson (layer), and Ning Yang (Silkie).

Figure 2: SNP and indel rates versus chromosome number. We excluded all sequences with “random” chromosome positions. Because of the assembly problems on W, it is not shown. The rates are computed as an average of all 3 domestic lines.

Figure 3: Detailed haplotype patterns in 3 regions, each covered by 2 overlapping BACs from the second layer line (L2). The primary SNP data are labeled B (broiler), L1 (layer), and S (Silkie). All comparisons are to RJF, and we show only those sites where a SNP is identified in at least one of the 4 lines. Hence, the horizontal scale is linear in the number of SNP sites, but non-linear for size. BLUE colors indicate where a particular line agrees with RJF, while RED colors indicate where it does not. Overlapping BACs on GGA1 and GGA7, but not GGA14, are clearly from different haplotypes.

Figure 4: Multi-species alignments for ornithine transcarbamylase (*OTC*), indicating non-synonymous substitutions relative to human protein. *SIFT* intolerant position is indicated by site number and bold-faced lettering. WT=wild type. MUT=mutant.

Table captions

Table 1: Frequency of SNPs in different comparisons of RJF and the 3 domestic chicken lines. In addition, we show comparisons involving 3.8-Mb of finished BAC sequence from another line of the layer (White Leghorn) breed. SNP rates are an estimate of nucleotide diversity (π), as embodied by the effective length, which considers how much of the data is of sufficiently good quality to actually detect SNPs and the probability that overlapping reads might be derived from homologous chromosomes.

Table 2: Frequency of sequence polymorphisms between RJF and broiler, decomposed by functional context based on three non-redundant gene sets of 3868 confirmed mRNA transcripts, 995 chicken orthologs of known human disease genes, and 17,709 Ensembl annotations. Human-chicken motifs are conserved sequences that exhibit no evidence of being genic in origin. Gene regions are subdivided into 5’-UTR, coding exon, intron, and 3’-UTR. Ka and Ks indicate non-synonymous and synonymous rates.

International Chicken Polymorphism Map Consortium

(Group contributions are listed by their order of appearance in the manuscript)

Polymorphism discovery and analysis: ***Beijing Institute of Genomics of Chinese Academy of Sciences*** Gane Ka-Shu Wong[1-3,*‡], Bin Liu[1,*], Jun Wang[1,2,*], Yong Zhang[1,4,*], Xu Yang[1,*], Zengjin Zhang[1], Qingshun Meng[1], Jun Zhou[1], Dawei Li[1], Jingjing Zhang[1], Peixiang Ni[1], Songgang Li[1,4], Longhua Ran[5], Heng Li[1,6], Jianguo Zhang[1], Ruiqiang Li[1], Shengting Li[1], Hongkun Zheng[1], Wei Lin[1], Guangyuan Li[1], Xiaoling Wang[1], Wenming Zhao[1], Jun Li[1], Chen Ye[1], Mingtao Dai[1], Jue Ruan[1], Yan Zhou[2], Yuanzhe Li[1], Ximiao He[1], Yunze Zhang[1], Jing Wang[1,4], Xiangang Huang[1], Wei Tong[1], Jie Chen[1], Jia Ye[1,2], Chen Chen[1], Ning Wei[1], Guoqing Li[1], Le Dong[1], Fengdi Lan[1], Yongqiao Sun[1], Zhenpeng Zhang[1], Zheng Yang[1], Yingpu Yu[2], Yanqing Huang[1], Dandan He[1], Yan Xi[1], Dong Wei[1], Qiuhui Qi[1], Wenjie Li[1], Jianping Shi[1], Miaoheng Wang[1], Fei Xie[1], Jianjun Wang[1], Xiaowei Zhang[1], Pei Wang[1], Yiqiang Zhao[7], Ning Li[7], Ning Yang[7], Wei Dong[1], Songnian Hu[1], Changqing Zeng[1], Weimou Zheng[1,6], Bailin Hao[1,6].

[1]Beijing Institute of Genomics of Chinese Academy of Sciences, Beijing Genomics Institute, Beijing Proteomics Institute, Beijing 101300, China. [2]James D. Watson Institute of Genome Sciences of Zhejiang University, Hangzhou Genomics Institute, Key Laboratory of Bioinformatics of Zhejiang Province, Hangzhou 310007, China. [3]UW Genome Center, Department of Medicine, University of Washington, Seattle, WA 98195, USA. [4]College of Life Sciences, Peking University, Beijing 100871, China. [5]Beijing North Computation Center, Beijing 100091, China. [6]The Institute of Theoretical Physics Chinese Academy of Sciences, Beijing 100080, China. [7]China Agricultural University, Beijing 100094, China.

Genome sequence of Red Jungle Fowl: ***Washington University School of Medicine*** LaDeana W. Hillier[8], Shiaw-Pyng Yang[8], Wesley C. Warren[8], Richard K. Wilson[8].

[8]*Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, St. Louis, MO 63108, USA.*

Molecular evolution: ***Uppsala University*** Mikael Brandström[9], Hans Ellegren[9].

[9]*Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 34 Uppsala, Sweden.*

Genotyping in populations, BAC sequences and haplotypes: ***Wageningen University*** Richard P.M.A. Crooijmans[10], Jan J. van der Poel[10], Henk Bovenhuis[10], Martien A.M. Groenen[10]; ***Lawrence Livermore National Laboratory*** Ivan Ovcharenko[11,12], Laurie Gordon[11,13], Lisa Stubbs[11]; ***DOE Joint Genome Institute***, Susan Lucas[13], Tijana Glavin[13], Andrea Aerts[13].

[10]*Animal Breeding and Genetics Group, Wageningen University, Marijkewg 40, 6709 PG Wageningen, The Netherlands.* [11]*Genome Biology Division, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA.* [12]*Energy, Environment, Biology and Institutional Computing, Lawrence Livermore National Laboratory.* [13]*DOE Joint Genome Institute, Walnut Creek, CA 94598, USA.*

Examples of application to complex traits: ***Institute for Animal Health*** Pete Kaiser[14], Lisa Rothwell[14], John R. Young[14], Sally Rogers[14], Brian A. Walker[14], Andy van Hateren[14], Jim Kaufman[14], Nat Bumstead[14]; ***Iowa State University*** Susan J. Lamont[15], Huaijun Zhou[15]; ***Roslin Institute*** Paul M. Hocking[16], David Morrice[16], Dirk-Jan de Koning[16], Andy Law[16], Neil Bartley[16], David W. Burt[16]; ***USDA-ARS Avian Disease and Oncology Laboratory*** Henry Hunt[17], Hans H. Cheng[17].

[14]*Institute for Animal Health, Compton, Berkshire RG20 7NN, UK.* [15]*Department of Animal Science, Iowa State Univeristy, Ames, IA 50011, USA.* [16]*Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, UK.* [17]*USDA-ARS Avian Disease and Oncology Laboratory, 3606 E. Mount Hope Rd., East Lansing, MI 48823, USA.*

Domestication and selection: ***Uppsala University*** Ulrika Gunnarsson[18], Per Wahlberg[18], Leif Andersson[18,19,‡]; ***Karolinska Institutet*** Ellen Kindlund[20], Martti T. Tammi[20,21], Björn Andersson[20].

[18]Department of Medical Biochemistry and Microbiology, Uppsala University, Box 597, SE-751 24 Uppsala, Sweden. [19]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, SE-751 24 Uppsala, Sweden. [20]Center for Genomics and Bioinformatics, Karolinska Institutet, SE-171 77 Stockholm, Sweden. [21]Departments of Biological Sciences and Biochemistry, National University of Singapore, Singapore.

Human disease genes: ***University of Oxford*** Caleb Webber[22], Chris P. Ponting[22].

[22]MRC Functional Genetics Unit, University of Oxford, Department of Human Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, UK.

EST-based SNP data: ***University of Manchester Institute of Science and Technology*** Ian M. Overton[23], Paul E Boardman[23], Haizhou Tang[23], Simon J. Hubbard[23]; ***University of Sheffield*** Stuart A Wilson[24].

[23]Department of Biomolecular Sciences, University of Manchester Institute of Science and Technology, PO Box 88, Manchester M60 1QD, UK. [24]Department of Molecular Biology and Biotechnology, University of Sheffield, Firth Court, Western Bank, Sheffield S10 2TN, UK.

Scientific management: ***Beijing Institute of Genomics of Chinese Academy of Sciences*** Jun Yu[1,2,*], Jian Wang[1,2], HuanMing Yang[1,2,‡].

[*]These authors contributed equally to this work.

Corresponding authors: Gane Ka-Shu Wong gksw@genomics.org.cn, Leif Andersson leif.andersson@imbim.uu.se, HuanMing Yang hyang@genomics.org.cn.

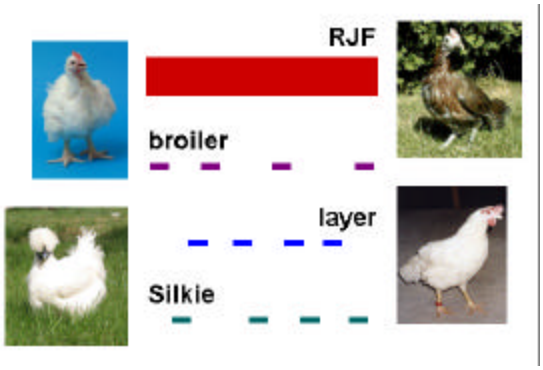

Figure 1:

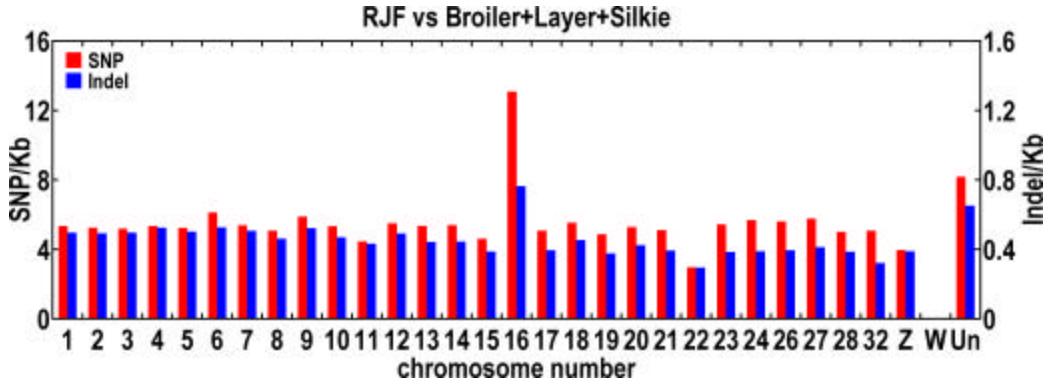
RJF vs Broiler+Layer+Silkie
SNP
Indel
SNP/Kb
Indel/Kb
0
4
8
12
16
0
0.4
0.8
1.2
1.6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 26 27 28 32 Z W Un
chromosome number

Figure 2:

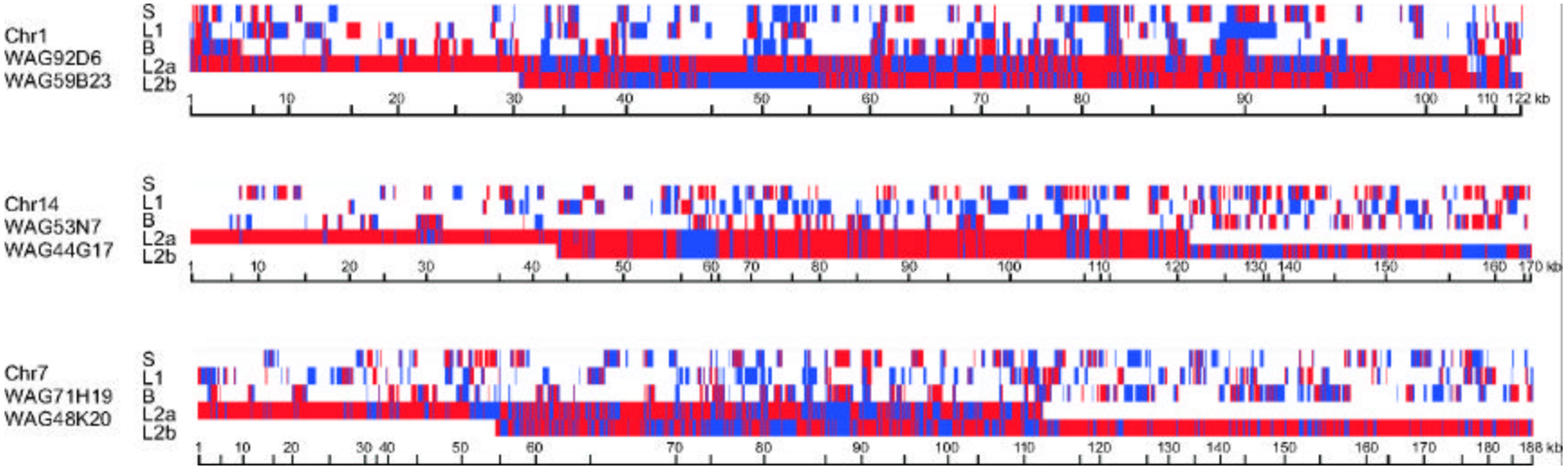
Chr1
WAG92D6
WAG59B23
S
L1
B
L2a
L2b
1 10 20 30 40 50 60 70 80 90 100 110 122 kb
Chr14
WAG53N7
WAG44G17
S
L1
B
L2a
L2b
1 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 kb
Chr7
WAG71H19
WAG48K20
S
L1
B
L2a
L2b
1 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 188 kb

Figure 3:

Ornithine transcarbamylase (*OTC*)

	188
Human	HYSSLK**G**LTLSWIGDGN
Pig	--GA----------
Mouse	--G----------
Rat	--G----------
Chicken, WT	--GG-N---IA------
Chicken, MUT	--GG-N**R**--IA------

Figure 4:

Table 1:

	# of SNPs	L(effective)	SNP/Kb
Wild versus domestic			
RJF-Broiler	1,041,948	197,431,517	5.28
RJF-Layer	889,377	170,586,544	5.21
RJF-Silkie	1,217,817	217,841,171	5.59
Between domestic lines			
Broiler-Layer	194,605	37,506,800	5.19
Broiler-Silkie	257,849	47,554,311	5.42
Layer-Silkie	246,954	42,682,304	5.79
Within domestic lines			
Broiler-Broiler	59,227	13,835,075	4.28
Layer-Layer	40,412	10,863,595	3.72
Silkie-Silkie	83,630	15,253,383	5.48
Compare to layer BACs			
RJF-to-BAC	20,925	3,809,567	5.49
BAC-Broiler	4,404	847,456	5.20
BAC-Layer	3,904	740,392	5.27
BAC-Silkie	5,089	925,738	5.50

Table 2: RJF-Broiler polymorphisms

	SNP/Kb	Indel/Kb	# of SNP	# of Indel
Confirmed mRNA transcripts				
5'-UTR	3.45	0.46	203	27
coding region	2.11	0.05	1,772	41
non-synonymous Ka	0.73			
synonymous Ks	7.44			
introns	5.70	0.52	86,586	7,915
3'-UTR	3.40	0.42	1,946	243
Human disease genes				
coding region	2.74	0.04	1,005	15
non-synonymous Ka	1.10			
synonymous Ks	9.40			
introns	5.36	0.49	27,768	2,553
Ensembl (final version 040427)				
5'-UTR	4.22	0.37	616	54
coding region	2.71	0.06	12,229	276
non-synonymous Ka	1.17			
synonymous Ks	8.28			
introns	5.64	0.52	367,361	33,869
3'-UTR	3.92	0.43	2,130	236
Human-chicken motifs	2.41	0.25	3,636	379
Genomewide average	**5.28**	**0.48**	**1,041,948**	**94,578**

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. The final set of 10,572 high quality SNPs is mapped to the Ensembl annotations on a reciprocal top-hits criterion, and matches are screened for 98% sequence identity over a minimum of 100-bp. Although these ESTs are from domestic chickens, in every instance, one of the two alleles matches the RJF allele. A subset of 2103 SNPs map to the Ensembl annotations, and 424 of these are non-synonymous.

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

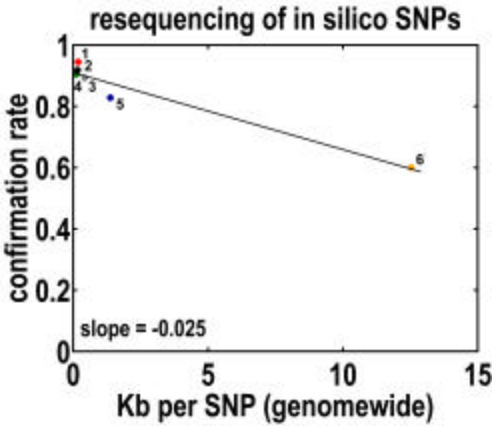

Table S2: RJF-Broiler polymorphisms

	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 100 kb segment	>80% shared alleles 1=2, not 3	>80% shared alleles 1=3, not 2	>80% shared alleles 2=3, not 1	>80% shared alleles 1=2, not 3	>80% shared alleles 1=3, not 2	>80% shared alleles 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			
	TOTAL	w/NS SNPs	TOTAL	SIFT done	intolerant wild RJF	intolerant domestic	intolerant both way	tolerant high conf	intolerant wild RJF	intolerant domestic	intolerant both way	tolerant low 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	372	372	424	359	7	43	1	165	9	33	25	76
		100.0%		**84.7%**	**1.9%**	**12.0%**	**0.3%**	**46.0%**	**2.5%**	**9.2%**	**7.0%**	**21.2%**

QTL fine mapping of entire chromosome: F2 broiler x layer cross identifying a single QTL on GGA4 affecting body weight

A previous microsatellite QTL analysis in an F2 broiler x layer chicken cross identified a single QTL on GGA4 affecting body weight[1]. The male-line broiler was from the same line as used in the SNP project. Starting with 256 randomly selected SNPs on GGA4 that are polymorphic between the layer and broiler lines used for the SNP project, 47 assays were designed using the SNPlex™ Genotyping System v2.0 (Applied Biosystems). Our F2 experimental cross (n = 466) was typed for these SNPs, and any informative SNPs were merged with the genotype data from 26 polymorphic microsatellite markers to give a higher density linkage map of the QTL region on GGA4. Genetic linkage maps were estimated for both sexes using *CriMap*[2]. QTL analysis was done in *QTL Express*[3] using the sex-averaged linkage map of 54 markers.

Of the 47 SNPlex assays, 7 failed, 11 were monomorphic, 1 was heterozygous in all F2, and the remaining 28 were informative. None of these 28 informative SNPs were line specific (i.e. both lines fixed for alternative alleles), and only four SNPs had line specific genotypes (e.g. one line homozygous and the other line partly heterozygous and partly homozygous for the other allele). The joint linkage map for GGA4 contained 54 markers spanning a total of 276 cM (sex-averaged map, Figure S1), with the female map longer than the male map by about 19%, contrary to expectations from the whole genome map[4].

These analyses provided evidence of two QTLs affecting body weight (Table S6). Their combined additive genetic effect of 230 g was similar to the previous estimate[1] from a single QTL of 249 g, at an average body weight of 2.0 kg. Together, these QTLs account for about one-third of the difference between broiler and layer lines at 6 weeks of age. The benefits of this new data are reflected in the improved genetic information content in areas of GGA4 with gaps in the microsatellite map (Figure S2). In the past, PIC values exceeding 0.5 to 0.6 were rare, but using the additional SNP data, they no longer are. The average marker interval is 5.2 cM for GGA4 in its entirety, but 4.3 cM for the q-arm that

has both microsatellite and SNP markers (6.9 and 3.7 cM for microsatellite and SNPs individually). Further benefits are expected when characterizing an Advanced Intercross Line[5], since identification of all recombinations in the 8-10th generation from the F2 should contribute to fine mapping of the QTL.

References

1. Sewalem, A. et al. Mapping of Quantitative Trait Loci (QTL) for body weight at 3, 6 and 9 weeks of age in a broiler layer cross. *Poult. Sci.* **81**, 1775-1781 (2002).
2. Green, P., Falls, K. & Crooks, S. *CriMap version 2.4* (Washington University School of Medicine, Saint Louis, 1990).
3. Seaton G., Haley C.S., Knott S.A., Kearsey M. & Visscher P.M. QTL Express: mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics* **18:** 339-340 (2002).
4. Groenen, M.A. & Crooijmans, R.P. Structural genomics: integrating linkage, physical and sequence maps. In *Poultry Genetics, Breeding and Biotechnology* (eds. Muir, W.M. & Aggrey, S.E.) 497-536 (CABI Publishing, Wallingford, 2003).
5. Song, J.Z., Soller, M. & Genezi, A. The full-sib intercross line (FSIL): a QTL mapping design for outcrossing species. *Genet. Res.* **73**, 61-73 (1999).

Table S6. QTL analysis for 6-week body weight of chicken chromosome 4, based on 26 microsatellites, with and without 28 novel SNP markers.

	26 microsatellites	**26 microsatellites + 28 SNPs**
Test statistic 2 vs. 0 QTL	21.5***	22.6***
Test statistic 2 vs. 1 QTL	9.6***	10.2***
Position of QTL 1	98	100
Position of QTL 2	240	237
Additive effect of QTL1, g	79±18	81±15
Dominance effect of QTL 1, g	17±26	14±26
Additive effect of QTL2, g	161±20	152±19
Dominance effect of QTL2, g	-42±35	-20±31

*** $P<0.001$

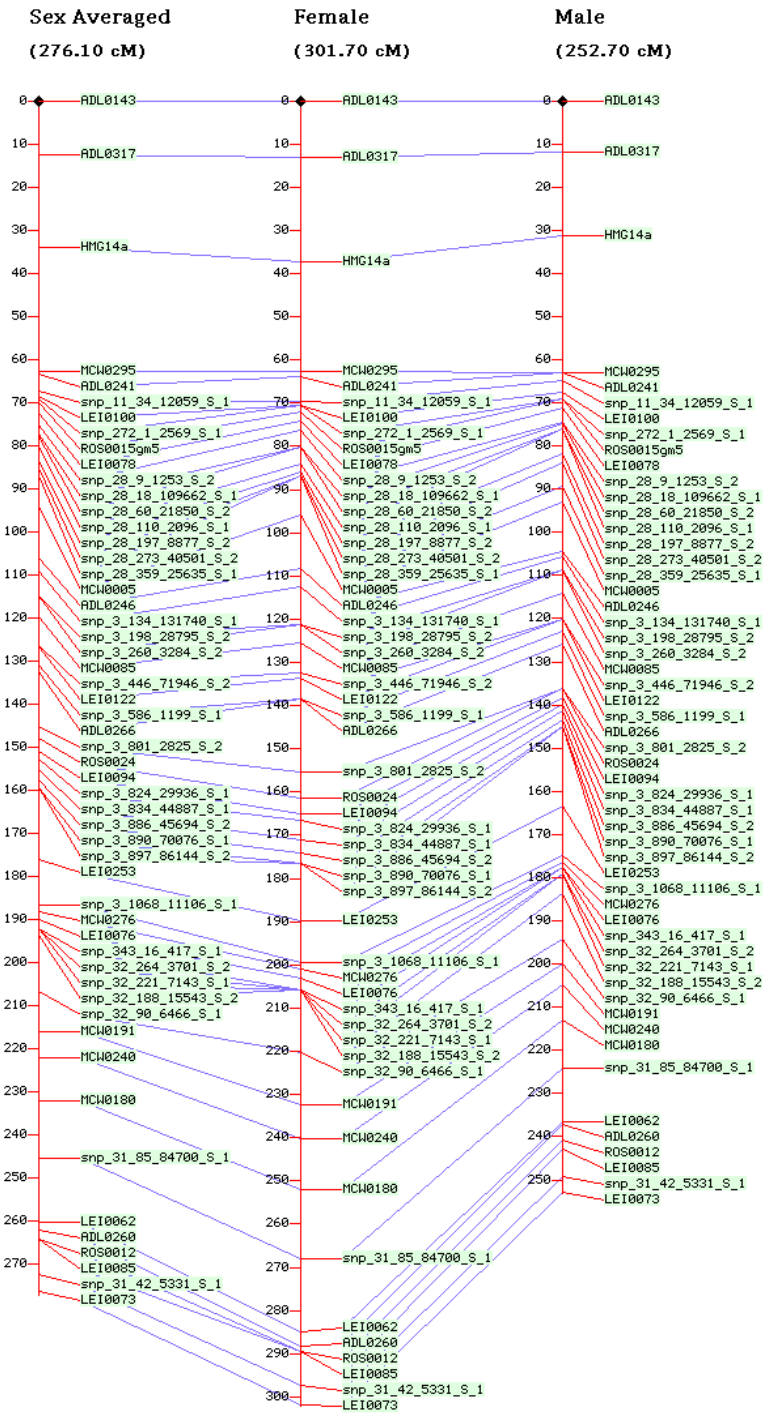

Figure S1. Sex-averaged and sex-specific linkage maps for chicken chromosome 4 using 26 microsatellites and 28 novel SNP markers.

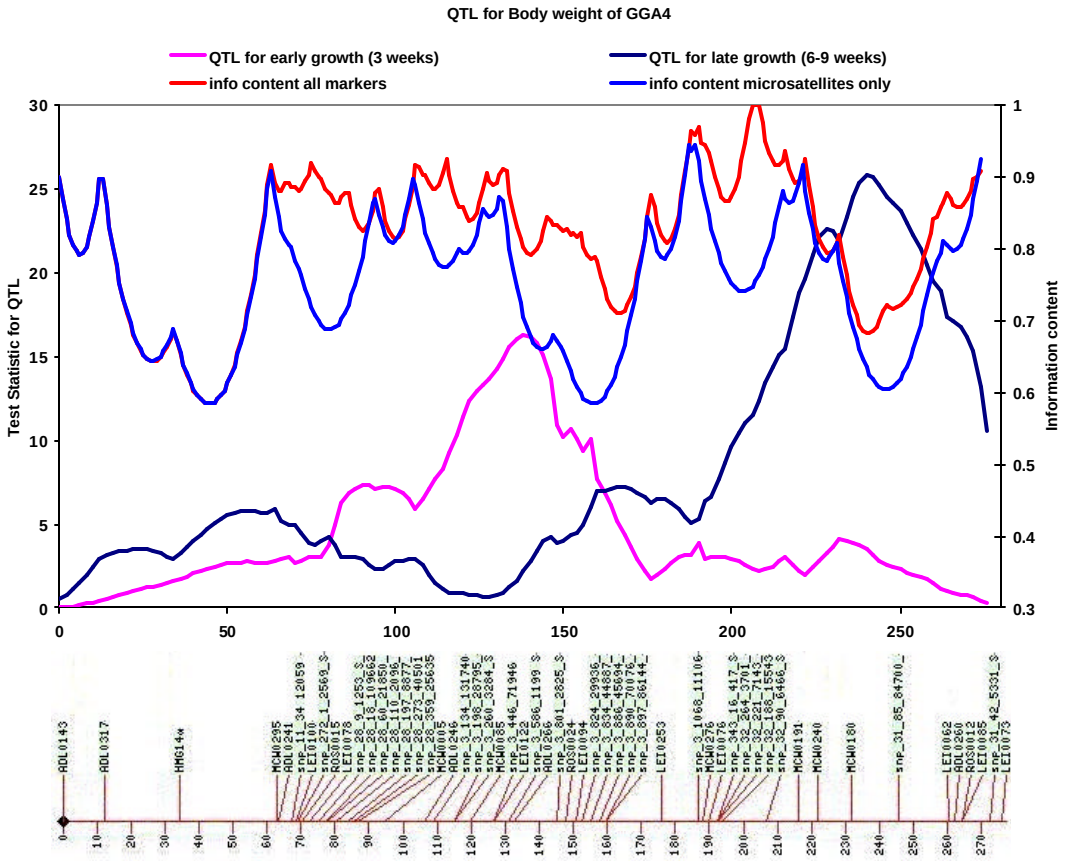

Figure S2. Information content from scan along GGA4 for QTL affecting body weight. The blue line shows information content using 26 microsatellite markers, and the red line shows information content using an additional 28 novel SNP markers. Test statistic for QTL affecting early growth (dark blue) and late growth (purple) are also shown. Marker positions are indicated under the x-axis.

QTL mapping to a specific region: SNP analysis of QTL in the TGFB2 region on GGA3 in broiler-Leghorn F2 cross

Some TGF-ß gene SNPs are associated with QTLs for important agronomic traits like antibody kinetics and body composition[1,2]. To refine the QTL regions within or near the TGFB2 gene on GGA3, we tested some additional SNPs from the SNP project on F2 resource populations[3] that were generated by crossing sires from a broiler breeder male line with dams from two genetically distinct highly inbred (> 99%) chicken lines[4], the Leghorn G-B2 and the Fayoumi M15.2. Fayoumi was imported to the U.S. from Egypt in 1954 because of reported resistance to avian leucosis. The F1 birds were intercrossed, within dam line, to produce two F2 populations.

We measured body weight at two-week intervals up to 8 weeks of age, as well as length, bone mineral content and bone mineral density of the tibia at 8 weeks of age[1]. To refine the QTL, we selected four regions (two on each side of TGFB2), spaced 10 cM apart and with 4 SNPs per region (16 SNPs in total). Ten of the 16 SNPs were informative in the F2 resource population, five in the broiler-Leghorn cross, and five in the broiler-Fayoumi cross (Table S7). In addition, five novel SNPs were identified. One SNP in each region was selected for PCR-RFLP typing of 386 F2 individuals from the broiler-Leghorn cross. An analysis of these SNP-trait associations (Table S8) indicate that the skeletal QTL is most likely to be between SNP2 and TGFB2, a region containing the following candidate genes: usherin isoform A, estrogen-related receptor gamma, lysophospholipase-like 1, zinc transporter 8, bifunctional amino acyl-tRNA synthetase. A separate, growth-related QTL may be present between SNP1 and SNP2.

References

1. Li, H., et al. Chicken quantitative trait loci for growth and body composition associated with transforming growth factor-ß genes. *Poult. Sci.* **82,** 347-356 (2003).

2. Zhou, H. & Lamont, S.J. Association of transforming growth factor ß genes with quantitative trait loci for antibody response kinetics in hens. *Anim. Genet.* **34,** 275-282 (2003).
3. Deeb, N. & Lamont, S. J. Genetic architecture of growth and body composition in unique chicken populations. *J. Heredity* **93,** 107-118 (2002).
4. Zhou, H. & Lamont, S. J. Genetic characterization of biodiversity in highly inbred chicken lines by microsatellite markers. *Anim. Genet.* **30**, 256-264 (1999).

Table S7: Primer and SNP information from the SNP project and the Iowa State University (ISU) F2 resource population.

Primer Set	Primer Sequences	Position on GGA3	Predicted in SNP dataset	ISU Resource Population (F/B/L)
Ch3IL	F: 5'-ATCTTCCTGAGTGGAGTAGTTCT-3'	10708457	G>A (JF/L)	G>G>G
	R: 5'-CGTAACCTAACCAAAAAGTAAAA-3'	10708476	C>T	T>T>T
		10708495	14 bp del in JF	no deletion
		10708551	A>C (JF/L)	A>C>C
		10708573		G>G/A>A
		10708769		**G>G/A>A**
Ch3IIB	F: 5'-GCAAGGTAGCAAGGTTTATAGTA-3'	13210964		G>A>A
	R: 5'-TTGCATTGTATTTCATATGATTC-3'	13211017		**7 bp deletion in Broiler**
		13211080	T>C (JF/B)	C>T>T
		13211096	C>T (JF/B)	T>C>C
		13211116		A>G>G
		13211180	7 bp del in B	7 bp del in Fayoumi
		13211190	1 bp del in B	1 bp del in Fayoumi
Ch3IIIL	F: 5'-ACAGTCTGCATATCCAACACTAC-3'	18263896	T>G (JF/L)	T>T>G
	R: 5'-GTGAAAGCCATGTTAGAGATAAG-3'	18263990	A>G (JF/L)	**A>A>G**
		18264025	A>G (JF/L)	A>A>G
		18264067	C>T (JF/L)	C>C>T
Ch3IVL2	F: 5'-TTGTAGGTAACAAATGACAGGAT-3'	20576320	1 bp del in L C>T>T	no deletion
	R: 5'-AAGCAATGCTGTATCAGAGAGTA-3'	20576375	(JF/L/S)	C>C>C
		20576509	G>A (JF/L)	**G>G>A**
		20576531	G>A (JF/L)	G>G>G

Notes: JF/L = Jungle Fowl to Leghorn nucleotide change; JF/B = Jungle Fowl to broiler nucleotide change; S = Silkie; F/B/L = nucleotide change from Fayoumi to broiler sire to Leghorn in the Iowa State University resource population; the **three SNPs and an indel in bold** (one per amplicon; referred to as SNP1, 2, 3, and 4 in the text and Table S8) were used for F2 genotyping of the ISU broiler-Leghorn resource population.

Table S8: Associations (*P* value) of SNPs with chicken skeletal and growth traits in a broiler-Leghorn F2 cross.

Trait	*P* value				
	SNP1	**SNP2**	**TGFB2**	**SNP3**	**SNP4**
Skeletal Traits					
BMC (g)	NS	0.03	0.02	NS	NS
BMD (g/cm2)	NS	0.1	0.05	NS	NS
TBL (mm)	NS	NS	0.05	0.17	NS
Growth					
BW (g) 2 wk	0.19	NS	NS	NS	NS
BW (g) 4 wk	0.13	NS	NS	NS	NS
BW (g) 6 wk	0.14	0.07	NS	NS	NS
BW (g) 8 wk	0.14	0.04	NS	NS	NS

Location of SNPs noted in Table S7; NS = $P > 0.20$; BMC = bone mineral content; BMD = bone mineral density; TBL = tibia length; BW = body weight; TGFB2 data comes from Li et al. 2003.

Application of SNPs for candidate gene association: Cytokines

Cytokines control the immune response, and in mammals, polymorphisms in cytokine genes are associated with disease resistance or susceptibility[1]. We identified 326 SNPs in some 12 pro-inflammatory, Th1, Th2 and Treg cytokine genes previously characterized in our laboratory. Forty such SNPs that mapped to coding sequences or known regulatory regions were amplified by PCR of genomic DNA from each of 8 inbred White Leghorn (Layer) lines. SNPs were identified by direct sequencing of the PCR products, and 32 of them were informative (Table S9). Six segregated between eight inbred layer lines (Table S10), and they mapped correctly in the genome when their segregation was analyzed in backcross mapping populations (Compton reference populations line 6_1 x line 7_2 and line 15I x line N - PMID 1353476)[2]. Four of the SNPs, in the Th2 cytokine genes IL-4 and IL-13 that drive antibody responses, segregated between the inbred Layer lines N and 15I that show differential antibody responses to vaccination[3]. They are therefore candidate SNPs for the differential responses between these two lines.

References

1. Gallagher, G., Eskdale, J. & Bidwell, J.L. Cytokine genetics - polymorphisms, functional variations and disease associations. In *The Cytokine Handbook, 4th Ed.* (eds. Thomson, A.W. & Lotze, M.T.) 19-55 (Academic Press, London, 2003).
2. Bumstead, N. Genomic mapping of resistance to Marek's disease. *Avian Pathol.* **27**, S78-S81 (1998).
3. Bumstead, N. et al. *EU Project FAIR3 PL96-1502 New molecular approaches for improved poultry vaccines* (Institute for Animal Health, Compton, 2000).

Table S9. Details of SNPs identified within cytokine genes. The cytokines are grouped according to function. B-L-S = broiler-layer-Silkie, i.e. the number of SNPs identified in a particular line for each cytokine gene. Forty of these SNPs were in coding or regulatory regions. Of these, 32 were informative. Of these, 6 segregated between our inbred lines, and their id numbers are given.

Cytokine gene	No. of SNPs (B-L-S)	No. of informative SNPs	No. of segregating SNPs	SNP #
Pro-inflammatory				
IL-6	0-0-6	0	0	-
Th1				
IL-2	0-0-2	1	0	-
IL-12α	18-3-12	19	1	snp.43.100.1355.S.1
IL-12β	17-10-33		0	-
IL-18	25-0-19	0	0	-
Th2				
IL-4	3-13-12	5	4	snp.103.50.22506.S.3 snp.103.50.22726.S.3 snp.103.50.22795.S.3 snp.103.50.22884.S.3
IL-5	14-2-9	0	0	-
IL-13	0-2-15	4	1	snp.103.50.16122.S.3
Treg				
IL-10	2-0-3	0	0	-
Others				
IL-3	9-6-9	0	0	-
IL-15	15-14-24	0	0	-
GM-CSF	10-8-10	3	0	-

Table S10: SNPs in cytokine genes that are polymorphic between layers with different MHC haplotypes. SNPs are shown as nucleotide changes, with positions in the chicken genome indicated by chromosome and base number. BLS refers to the sequence in broiler, layer, and/or Silkie (BLS = change in broiler, layer or Silkie respectively, - = no change, x = not sequenced). The gene in which each SNP is located is indicated. Under MHC haplotype, - = not determined. The four SNPs in bold were used for Backcross genotyping of the Compton Mapping (Nx15I) and MDV Mapping (6x7) populations.

					Line (MHC B Haplotype)								
SNP Number	**Chr. - Base No.**	**SNP**	**BLS**	**Gene**	**6 (2)**	**7 (2)**	**15I (15)**	**N (21)**	**0 (21)**	**W (14)**	**B4**	**B12**	**Notes**
snp.43.100.1355.S.1	9-21724516	T>C	BXX	IL-12A	C	-	-	C	-	T	C	C	Promoter
snp.103.50.16122.S.3	13-15971216	**G>C**	XXS	IL-13	C	G	C	G	G	G	C	C	Promoter
snp.103.50.22506.S.3	13-15977600	T>C	X-S	IL-4	C	C	C	T	C	C	C	C	Promoter
snp.103.50.22726.S.3	13-15977820	**C>T**	X-S	IL-4	T	T	T	C	-	T	T	T	Promoter
snp.103.50.22795.S.3	13-15977889	**T>C**	XXS	IL-4	C	T	C	C	C	T	C	C	Met>Thr
snp.103.50.22884.S.3	13-15977978	**G>A**	XXS	IL-4	G	G	A	G	G	G	G	G	Intronic

Application of SNPs for candidate gene association: The MHC

DNA from eight 15-B congenic lines[1,2] was analyzed. The DNA was purified from whole blood cells using the QIAamp DNA Blood Minikit (QIAGEN, Valencia, CA), and then used as a template in a standard PCR reaction with the primers given in Table S11. When the SNP generated a restriction site, the PCR product was further analyzed by restriction fragment length polymorphism (RFLP). When the SNP produced no restriction site, the PCR product was directly sequenced with an ABI 3100 (both strands). We had previously sequenced numerous MHC-encoded genes from different haplotypes of White Leghorn (layer) chickens. We could therefore easily determine that some of the nucleotides in the MHC-encoded genes with SNPs from broiler, layer, and Silkie were also polymorphic between our haplotypes. Moreover, these SNPs can be used to distinguish between lines of White leghorn chickens that are resistant or susceptible to commercially important pathogens like Marek's Disease Virus. The combined results from both studies are shown in Table S12.

References

1. Shen, P.F., Smith, E.J. & Bacon, L.D. The ontogeny of blood cells, complement and immunoglobulins in 3- to 12-week-old 15I5-B congenic white Leghorn chickens. *Poult. Sci.* **63,** 1083-1093 (1984).
2. Bacon, L.D., Ismail, N. & Motta, J.V. Allograft and antibody responses of 15I5 B congenic chickens. *Prog. Clin. Biol. Res.* **238,** 219-233 (1987).

Table S11. Details of primers and methods used to analyze SNPs in the MHC.

SNP#	Chr. - Base No.	Forward Primer	Reverse primer	SNP Detection Method
snp.26856.S.1	MHC-26856	GCCTGAACCTTGATGTCCTTA	TTAGGGGACCGATGCTATG	RFLP (MnlI)
snp.36295.S.2	MHC-36295	ACAACGACAGCCCTAAGCACA	GGCAGCCGATGGAACCTAC	RFLP (MaeII)
snp.67126.S.2	MHC-67126	CACGTGGAGGGACAGCGGTCA	GGGACACTGAGCCGCACGCA	Sequencing
snp.67152.S.2	MHC-67152	CACGTGGAGGGACAGCGGTCA	GGGACACTGAGCCGCACGCA	Sequencing
snp.67164.S.2	MHC-67164	CACGTGGAGGGACAGCGGTCA	GGGACACTGAGCCGCACGCA	Sequencing
snp.67221.S.2	MHC-67221	CACGTGGAGGGACAGCGGTCA	GGGACACTGAGCCGCACGCA	Sequencing
snp.67272.S.2	MHC-67272	CACGTGGAGGGACAGCGGTCA	GGGACACTGAGCCGCACGCA	Sequencing
snp.64376.S.2	MHC-64376	CCCTTTGGCTGCGAGGATCTC	CGCTCACTCCACGCCAAC	RFLP (BstNI)
snp.69245.S.1	MHC-69245	TGGGGGCCGTTCTAAA	GCTCCAGGCAGACCTACATAG	RFLP (DsaI)

Table S12: SNPs in the chicken MHC that are polymorphic between layers with different MHC haplotypes. SNPs are shown as nucleotide changes, with position in the genome indicated by chromosome and base number, except for those labeled MHC, which are numbered according to EMBL Acc. No. AL023516. BLS refers to the sequence in broiler, layer, and/or Silkie (BLS = change in broiler, layer or Silkie respectively, - = no change, x = not sequenced). The gene in which each SNP is located is indicated and the amino acid residue encoded is shown in bold (where applicable). Under MHC haplotype, - = not determined.

					MHC B Haplotype									
SNP Number	**Chromosome - Base No.**	**SNP**	**BLS**	**Gene**	**2**	**4**	**5**	**12**	**13**	**14**	**15**	**19**	**21**	**Notes**
snp.7544.2.239.S.3	Un-151325532	C>T	XXS	TAP1 exon 10 (RD**P**RI)	C	C	-	C	C	C	C	C	C	non-coding
snp.7544.2.566.S.3	Un-151325859	A>G	XXS	TAP1 exon 11 (AE**R**VV)	G	A	-	G	A	G	G	G	A	non-coding
snp.7544.2.576.S.3	Un-151325869	T>C	XXS	TAP1 exon 11 (VV**L**EG)	T	T	-	C	T	T	C	C	T	non-coding
snp.368.11.10208.S.2	16-168065	A>G	-LX	BNK exon 6 (RL**H**P)	G	G	-	G	-	G	G	G	G	His>Tyr
snp.368.11.11881.S.2	16-169738	C>T	XLX	BNK intron 1	T	T	-	T	-	T	T	T	C	non-coding
snp.368.12.1112.S.2	16-171995	T>C	XLS	Blec 5'UTR	C	C	-	C	-	C	C	C	C	non-coding, possible NF-AT site
snp.368.12.1115.S.2	16-171998	G>C	XLS	Blec 5'UTR	C	C	-	C	-	C	C	C	G	non-coding
snp.368.14.2060.S.2	16-178072	T>C	XLX	Tapasin exon 5 (RV**S**VR)	C	C	-	C	-	C	T	C	C	non-coding
snp.368.14.2069.S.2	16-178081	G>A	XLX	Tapasin exon 5 (VR**L**LL)	G	G	-	G	-	G	G	G	A	non-coding
snp.26856.S.1	MHC-26856	G>A	BXX	B-NK exon 4 (AE**E**DH)	A	-	A	G	A	-	A	G	A	Glu>Lys
snp.36295.S.2	MHC-36295	A>G	XLX	Tapasin exon 5 (GD**I**YS)	G	-	G	A	G	-	G	A	G	Ile>Val
snp.64376.S.2	MHC-64376	A>G	XLX	TAP1 exon 9 (AR**Q**VG)	G	-	G	A	G	-	G	A	G	Gln>Arg
snp.69245.S.1	MHC-69245	G>A	BXX	TAP2 exon 1 (GP**R**GA)	G	-	G	G	G	-	G	G	G	Arg>His
snp.67126.S.2	MHC-67126	G>A	XLX	TAP1 exon 2 (QRF)	G	-	G	G	G	-	A	G	G	non-coding
snp.67152.S.2	MHC-67152	A>C	XLX	TAP1 exon 2	C	-	A	A	C	-	C	C	C	non-coding
snp.67164.S.2	MHC-67164	A>G	XLX	TAP1 exon 2	A	-	A	A	A	-	G	A	G	non-coding
snp.67221.S.2	MHC-67221	C>T	XLX	TAP1 exon 2	C	-	C	C	T	-	T	C	T	non-coding
snp.67272.S.2	MHC-67272	T>C	XLX	TAP1 exon 2	T	-	T	T	C	-	C	T	T	non-coding