The duck genome and transcriptome provide insight into an avian influenza virus reservoir species

Yinhua Huang[1,2*], YingRui Li[3*], David W. Burt[2*], Hualan Chen[4], Yong Zhang[3], Wubin Qian[3], Heebal Kim[5], Shangquan Gan[1], Yiqiang Zhao[1], Jianwen Li[3], Kang Yi[3], Huapeng Feng[4], Pengyang Zhu[4], Bo Li[3], Qiuyue Liu[1], Suan Fairley[6], Katharine E. Magor[7], Zhenlin Du[1], Xiaoxiang Hu[1], Laurie Goodman[3], Hakim Tafer[8,9], Alain Vignal[10], Taeheon Lee[5], Kyu-Won Kim[11], Zheya Sheng[1], Yang An[1], Steve Searle[6], Javier Herrero[12], Martien A.M. Groenen[13], Richard P.M.A. Crooijmans[13], Thomas Faraut[10], Qingle Cai[3], Robert G. Webster[14], Jerry R. Aldridge[14], Wesley C. Warren[15], Sebastian Bartschat[8], Stephanie Kehr[8], Manja Marz[8], Peter F. Stadler[8,9], Jacqueline Smith[2], Robert H.S. Kraus[13,16], Yaofeng Zhao[1], Liming Ren[1], Jing Fei[1], Mireille Morisson[10], Pete Kaiser[17], Darren K. Griffin[18], Man Rao[1], Frederique Pitel[10], Jun Wang[3,19], Ning Li[1]

[1]State Key Laboratory for Agrobiotechnology, China Agricultural University, Beijing, China. [2]Division of Genetics and Genomics, University of Edinburgh, Edinburgh, United Kingdom. [3]BGI-Shenzhen, Shenzhen, China. [4]National Key Laboratory of Veterinary Biotechnology, Harbin Veterinary Research Institute, Harbin, China. [5]Department of Agricultural Biotechnology, Seoul National University, Seoul, Korea. [6]Wellcome Trust Sanger Institute, Wellcome Tust Genome Campus, Hinxton, Cambridge, United Kingdom. [7]Department of Biological Sciences, University of Alberta, Edmonton, Canada. [8]Department of Computer Science, University of Leipzig, Leipzig, Germany. [9]Department of Theoretical Chemistry University of Vienna, Vienna, Austria. [10]Laboratoire de Génétique Cellulaire, INRA, France. [11]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea. [12]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom.

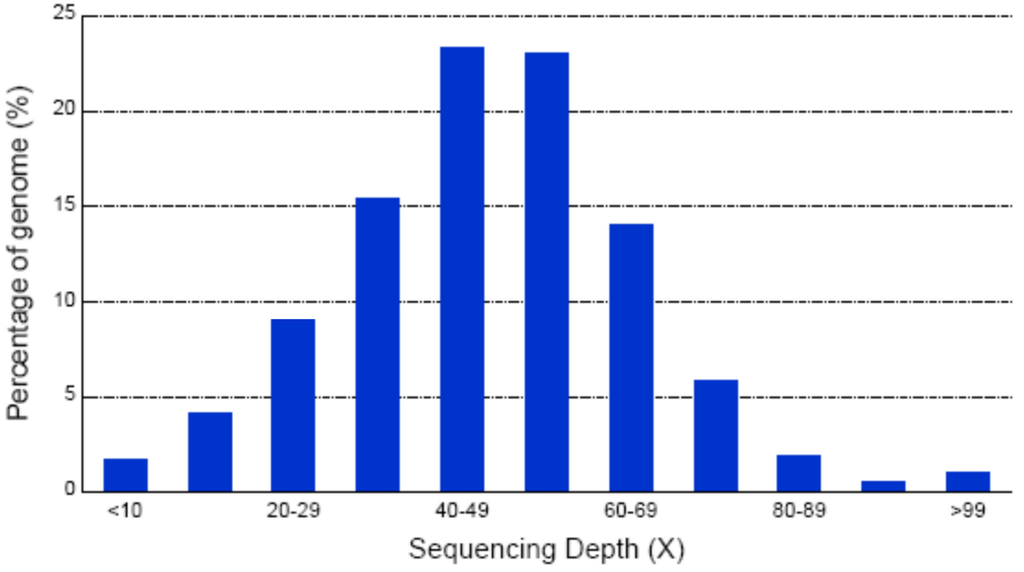
Nature Genetics: doi:10.1038/ng.2657

[13]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, The Netherlands. [14]Department of Infectious Diseases, St. Jude Children's Research Hospital, Memphis, USA. [15]The Genome Institute, Washington University School of Medicine, St Louis, USA. [16]Conservation Genetic Group, Senckenberg Research Institute and Natural History Museum, Gelnhausen, Germany. [17]Division of infection and immunity, University of Edinburgh, Edinburgh, United Kingdom. [18]Genetics School of Biosciences, University of Kent, Kent, United Kingdom. [19]Department of Biology, University of Copenhagen, Copenhagen, Denmark.

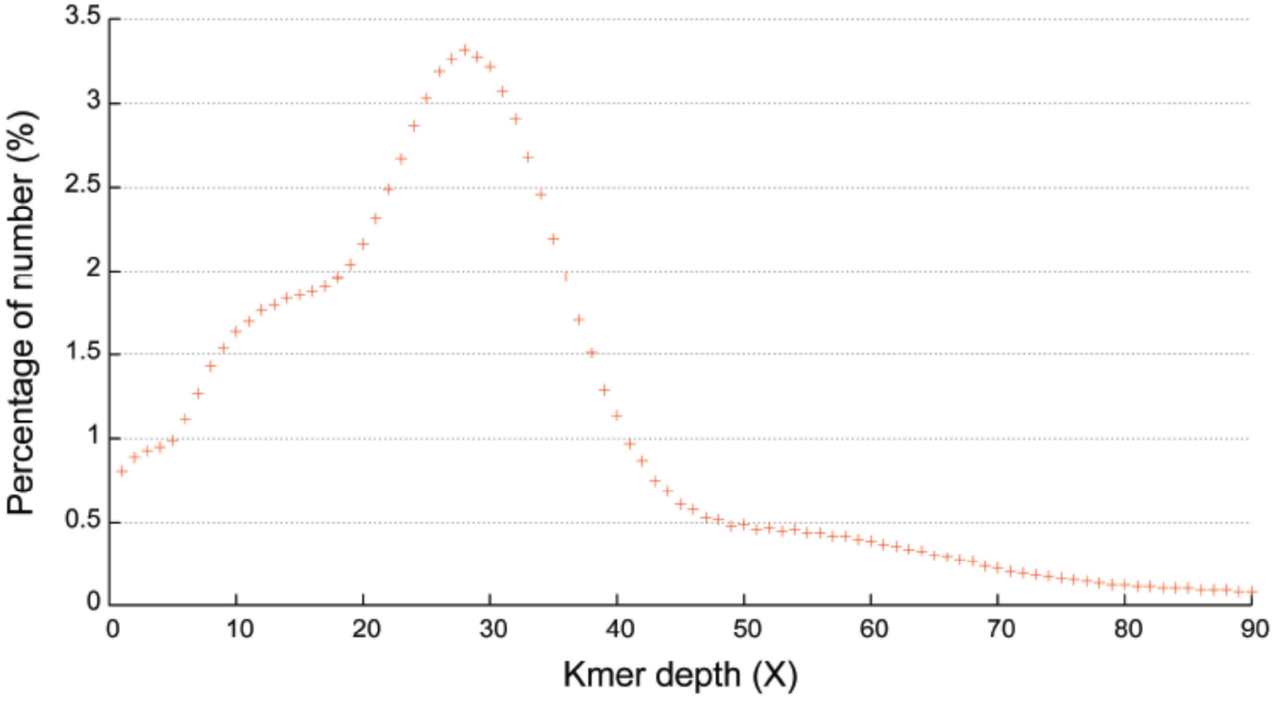
[*]These authors contributed equally to this work.

Correspondence and request for materials should be addressed to wangj@genomics.org.cn or ninglcau@cau.edu.cn.

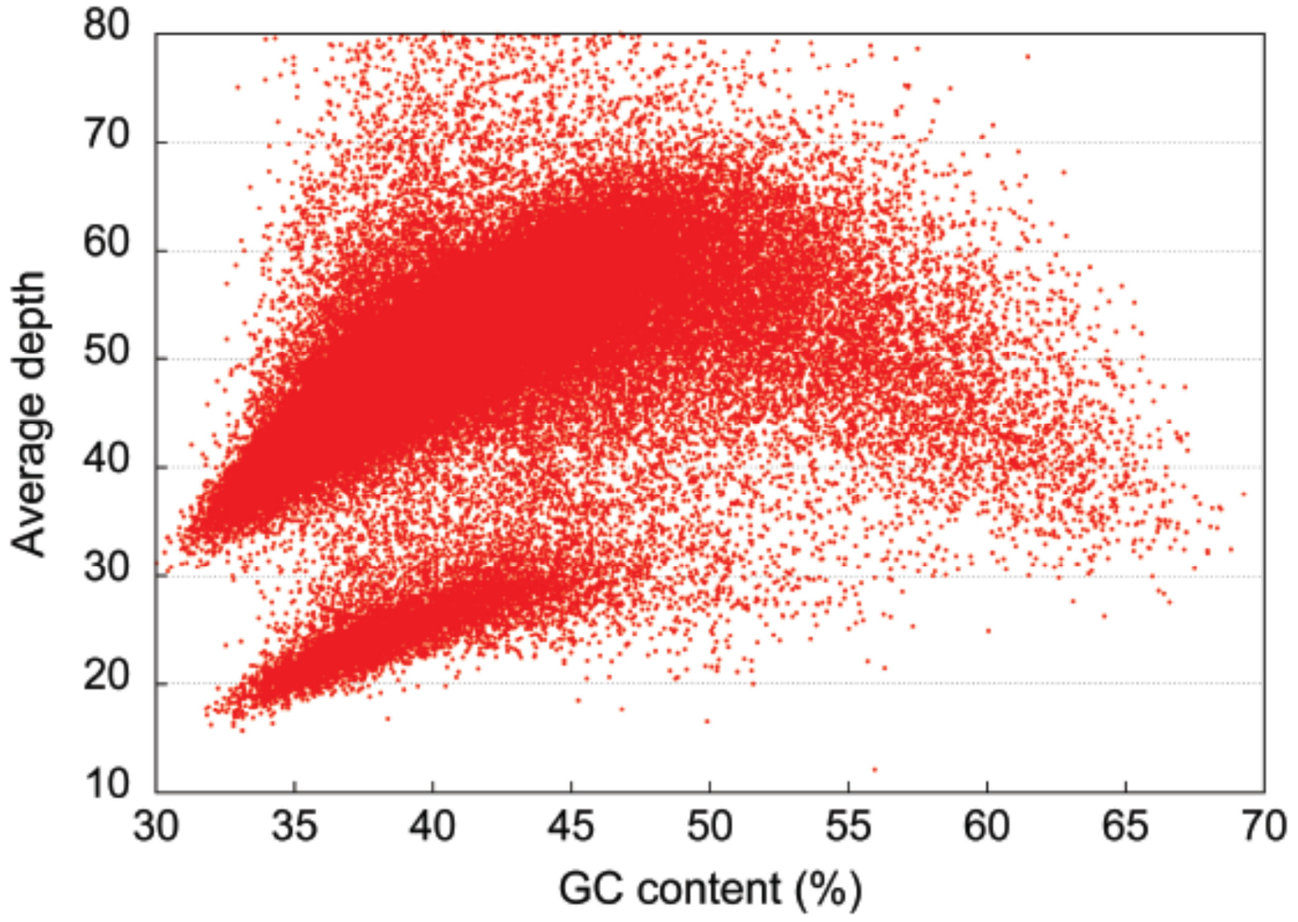
Supplementary Figures



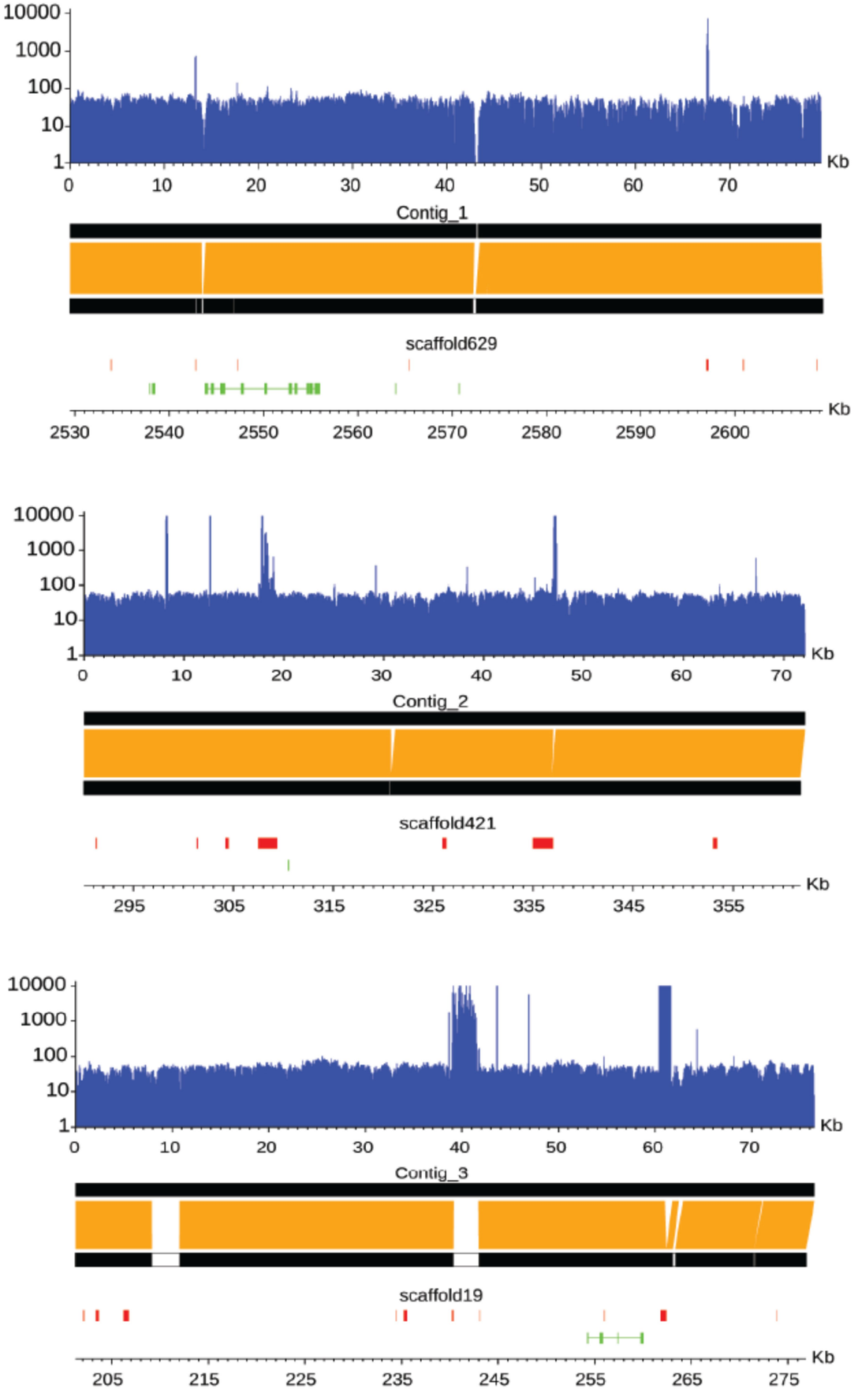
Supplementary Figure 1 | Distribution of the sequencing depth of the duck assembly. All 76.9 Gb duck whole genome sequence reads were mapped to the assembly by SOAPaligner, with a threshold of two mismatches. The sequencing depth at each locus was counted according to the corresponding number of reads in the duck assembly.



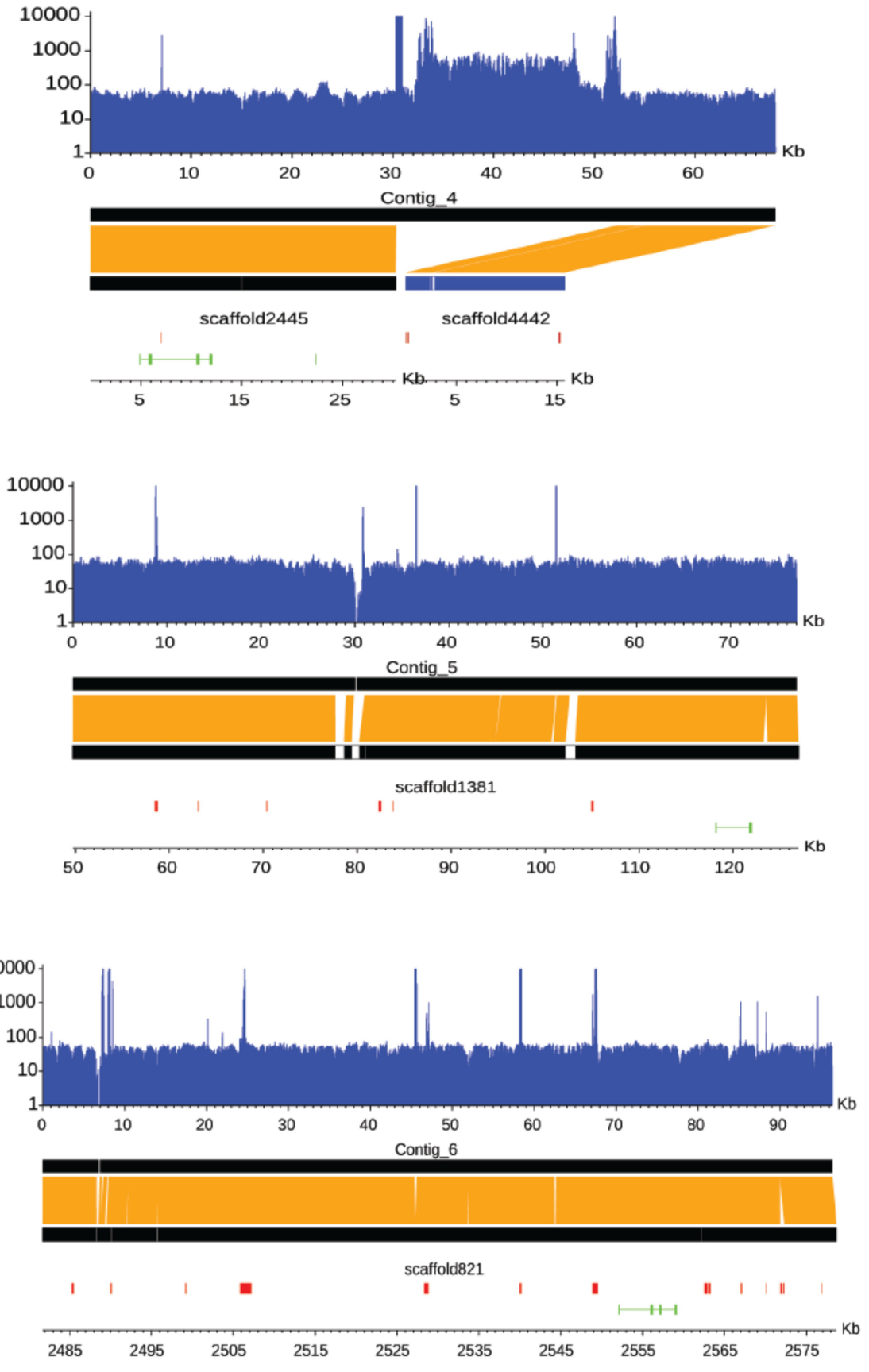
Supplementary Figure 2 | Distribution of 17-mer frequency in the corrected PE reads. Only the reads from the short insert-size libraries (< 500 bp) were included in this analysis. The peak depth is 28 fold. The peak of the 17-mer frequency (M) in the reads is correlated to the real sequencing depth (N), read length (L), and Kmer length (K), and their relationship can be expressed by the experiential formula, $M = N * (L - K + 1) / L$. Then, we divided the total sequence length by the real sequencing depth to estimate the duck genome size to be 1.26 Gb.

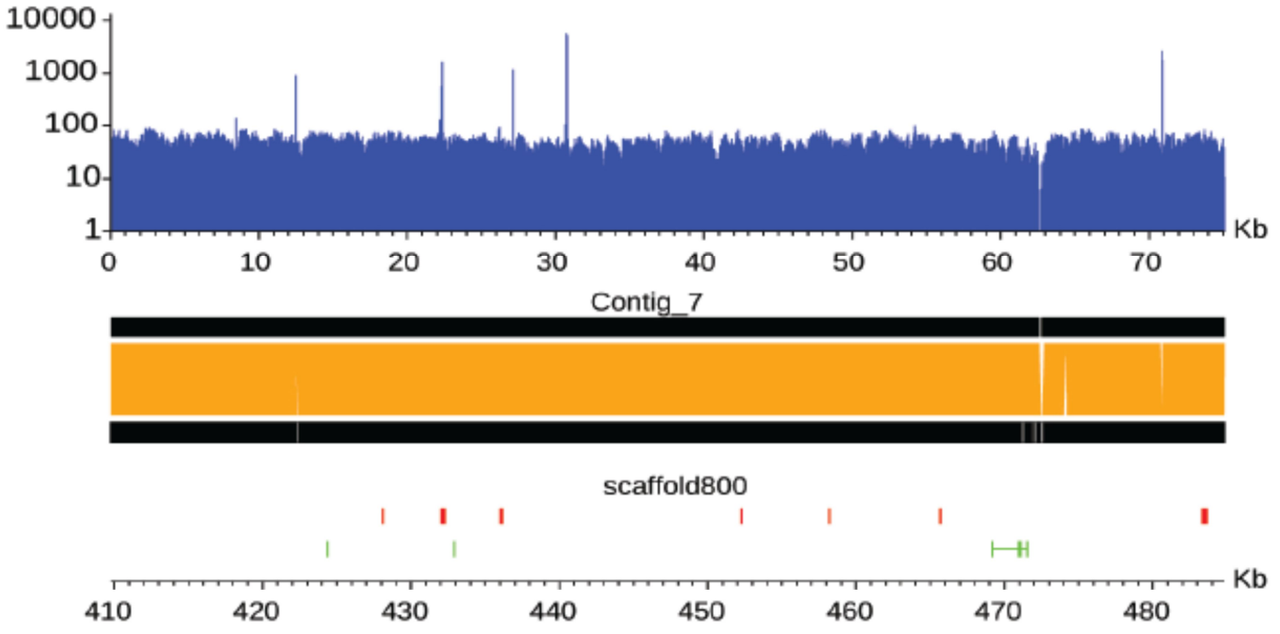


Supplementary Figure 3 | Local GC content versus sequencing depth. We used 10-kb non-overlapping sliding windows along the assembled sequence to calculate the GC content and average sequencing depth. The distribution of the GC content versus sequencing depth of the potential duck Z chromosome was inferred according to ~70 Mb of the chicken syntenic Z chromosome and is shown in the lower left block.

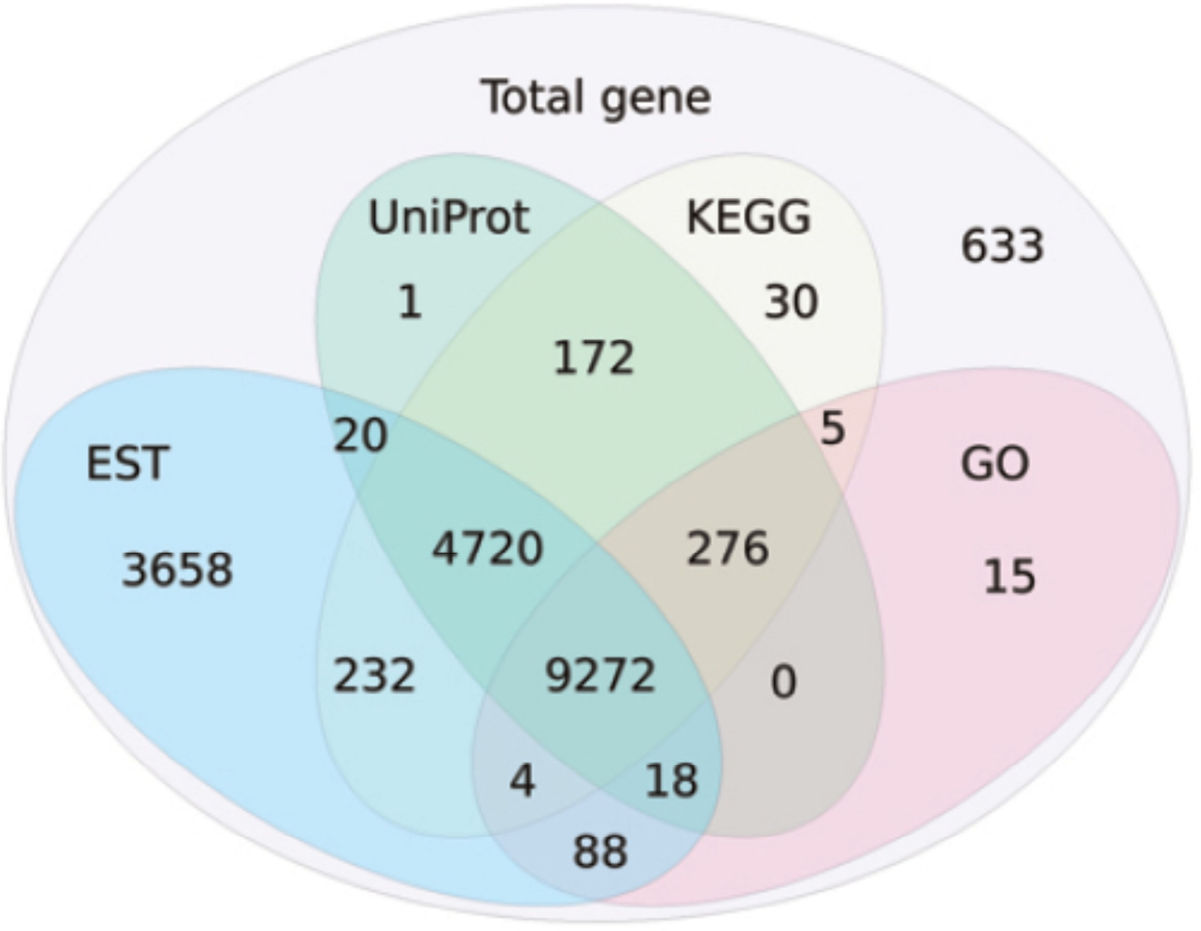


Nature Genetics: doi:10.1038/ng.2657

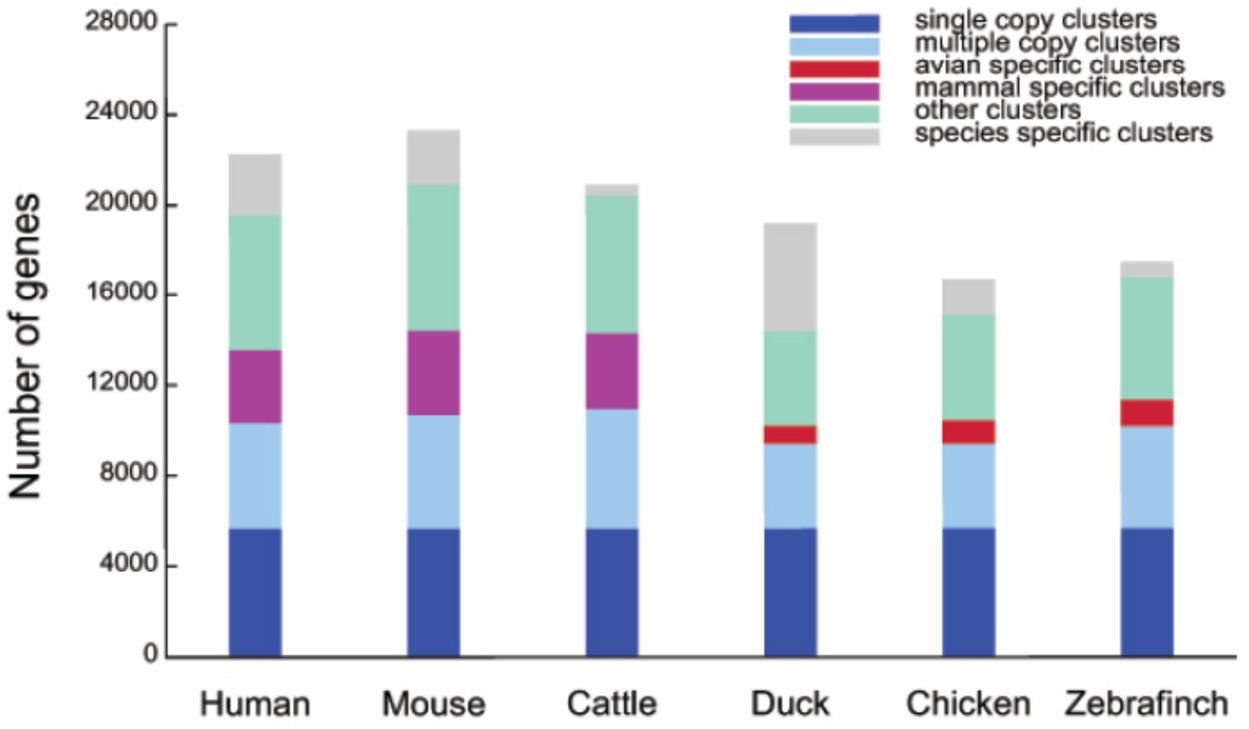

10000
1000
100
10
1
Kb
0
10
20
30
40
50
60
Contig_4
scaffold2445
scaffold4442
Kb
5
15
25
Kb
5
15
10000
1000
100
10
1
Kb
0
10
20
30
40
50
60
70
Contig_5
scaffold1381
Kb
50
60
70
80
90
100
110
120
10000
1000
100
10
1
Kb
0
10
20
30
40
50
60
70
80
90
Contig_6
scaffold821
Kb
2485
2495
2505
2515
2525
2535
2545
2555
2565
2575



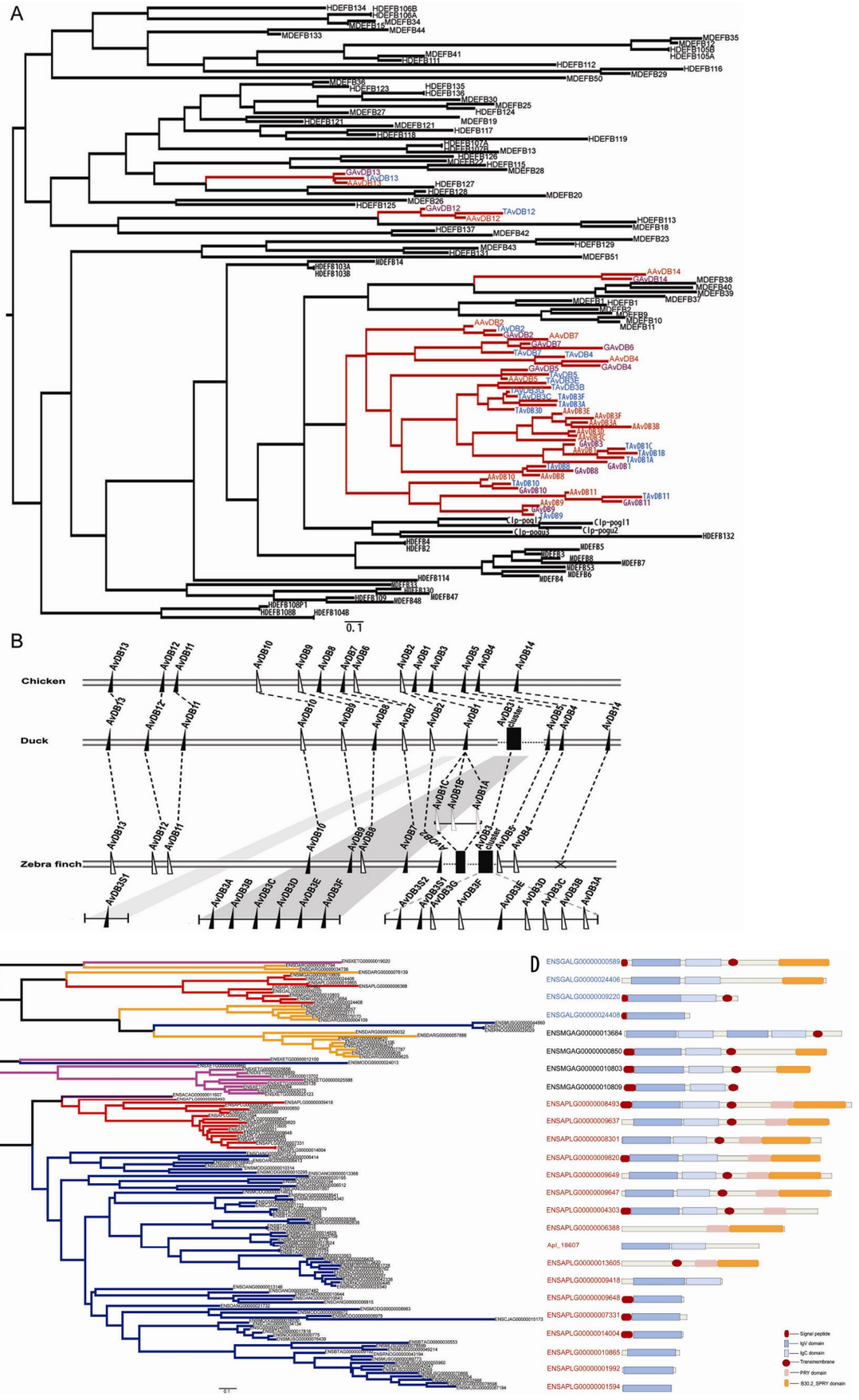
Supplementary Figure 4 | Comparison of the duck assembly and the sequences of 7 BACs. The predicted genes and annotated transposable elements (TEs) on the assembly are shown in green and red, respectively. The remaining unclosed gaps on the scaffolds are marked as white blocks. Contig_1 to 7 represent the sequences of the 7 BACs, which are distributed on chromosomes 1, 3 and 4. The seven BACs, covering 640 kb, aligned over more than 95% of their lengths.



Supplementary Figure 5 | Venn diagram showing the duck reference genes annotated using different databases or supported by EST evidence.



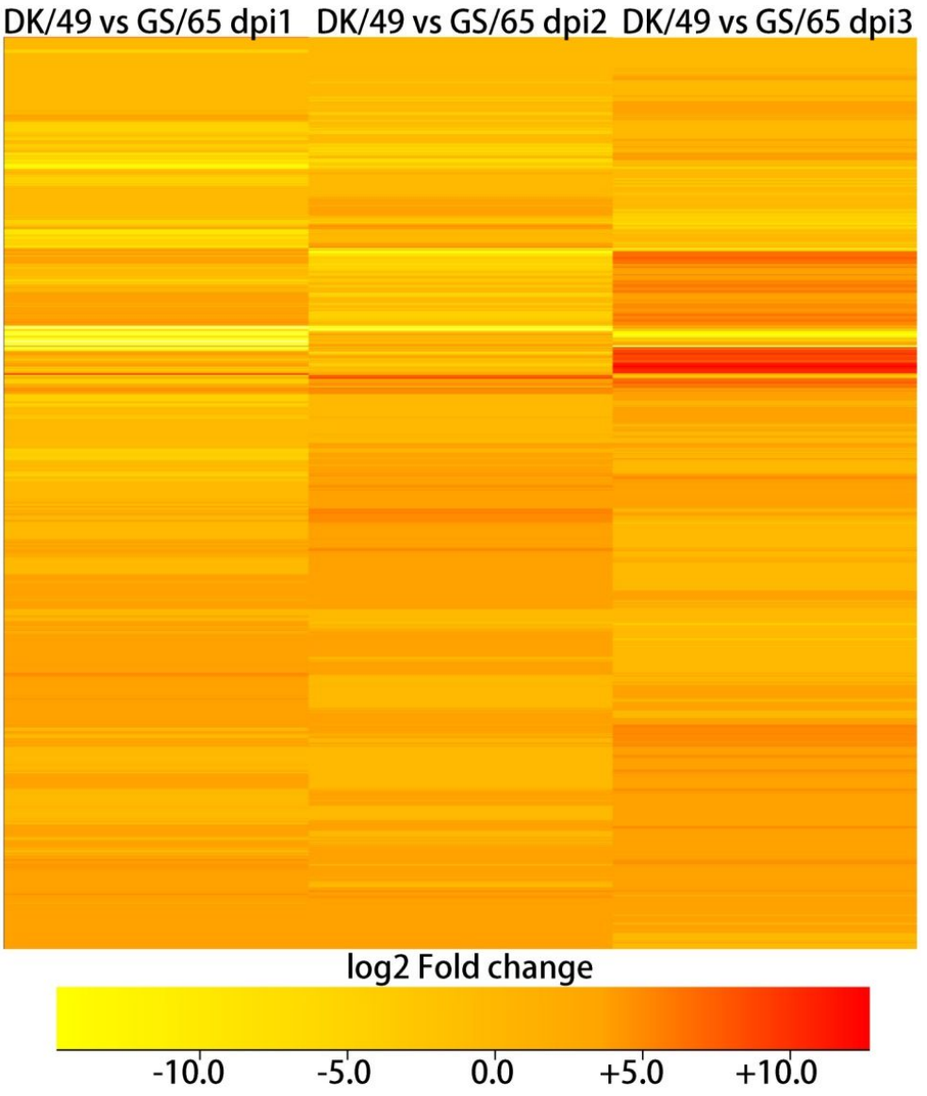
Supplementary Figure 6 | Protein orthology comparison among the reference gene sets of the duck, chicken, zebra finch, human, mouse and cattle assemblies.



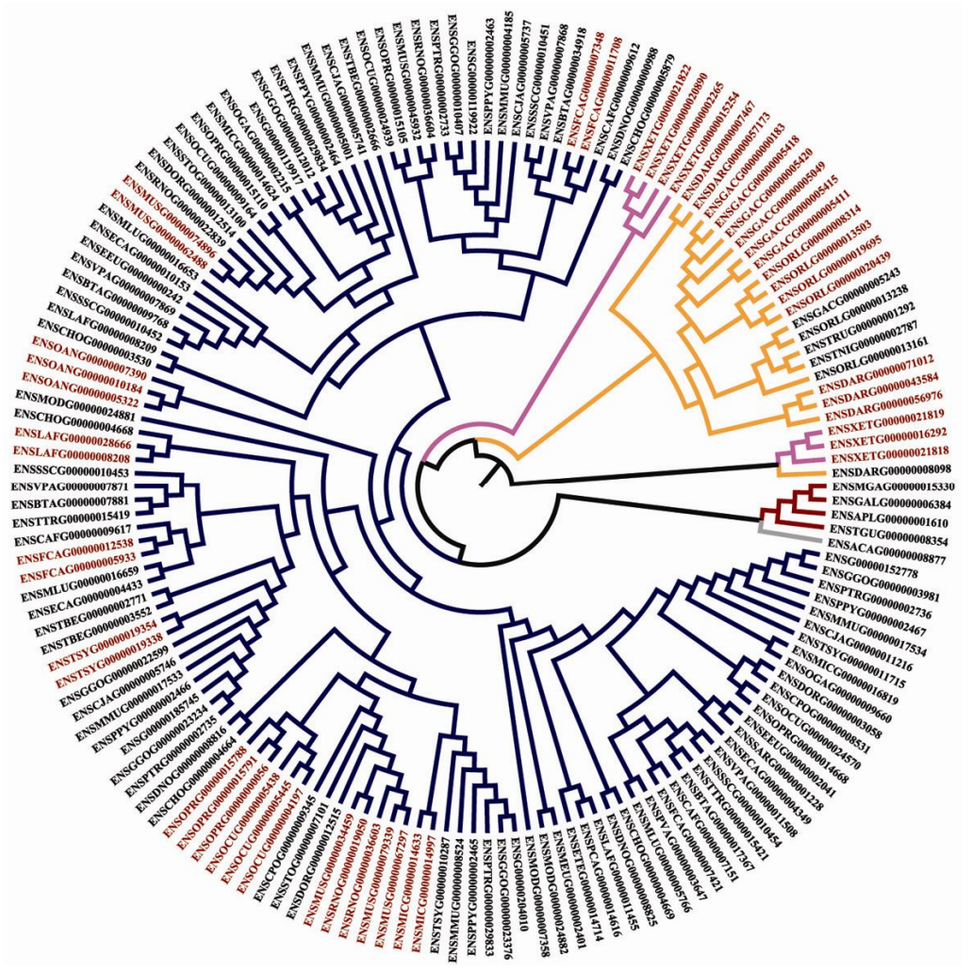
Nature Genetics: doi:10.1038/ng.2657

Supplementary Figure 7 | Two significantly expanded gene families of the duck. Maximum-likelihood trees were constructed with the protein sequences using PHYML version 2.4.4 under the JTT model with 4 substitution rate classes[1]. Domains of butyrophilin-like (BTNL) genes were predicted using SMART software[2]. A. Maximum-likelihood tree of β-defensins in vertebrates. The β-defensin taxa of the duck, chicken and zebra finch are shown in red, purple and blue, respectively. The nodes of the avian β-defensins (AvDBs) are shown in red. AAvDB: *Anas platyrhynchos* β-defensins; GAvDB: *Gallus gallus* β-defensins; TAvDB: *Taeniopygia guttata* β-defensins; HDEFB: *Homo sapiens* β-defensins; MDEFB: *Mus musculus* β-defensins; Clp-pogu: *Pogona barbata* crotamine-like peptides. B. Genomic organization of the β-defensin gene cluster in the chicken, duck and zebra finch. Blank and black triangles represent forward and reverse gene transcripts, respectively. AvDB3S1 and AvDB3S2 are pseudogenes. C. Maximum-likelihood tree of the BTNL family in vertebrates. The mammalian, avian, reptilian, amphibian and teleost clades are shown in blue, red, purple, pink and orange, respectively. D. Domain organization of the duck, chicken and turkey members within the BTNL family. The chicken, turkey and duck genes are presented in blue, black and red, respectively.

Nature Genetics: doi:10.1038/ng.2657



Supplementary Figure 8 | Heatmap of genes that showed significantly different expression in the DK/49 virus-infected ducks as compared with the GS/65 virus-infected ducks on day 1, 2 and 3. This heatmap generated from hierarchical cluster analyses of genes (using Spearman's rank correlation). Genes included in this figure showed significantly different gene expression (FDR<= 0.001, fold change> 2) in at least one experiment. Genes showed in red were upregulated and those showed in yellow were downregulated in the DK/49 virus-infected ducks relative to the GS/65 virus-infected ducks. This heatmap shows significant changes in expressions of 3,232 genes in the DK/49 virus-infected ducks when compared with the GS/65 virus-infected ducks.



Supplementary Figure 9 | Maximum likelihood tree of the interferon-induced protein with tetratricopeptide repeats (IFIT) superfamily based on vertebrate amino acid sequences. Maximum-likelihood trees were constructed with the protein sequences using PHYML version 2.4.4 under the JTT model with 4 substitution rate classes[1]. This tree is rooted with *Danio rerio*. The mammalian, avian, reptilian, amphibian and teleost clades are in blue, red, grey, pink and orange respectively. Genes produced by recent duplication (less than 60 Mya) or lineage-specific duplication are in red. This phylogenetic tree shows that the IFIT repertoire is divergent in mammals and this repertoire in birds is simplified with single gene in the four available genomes (the duck, chicken, turkey and zebra finch), we then name the single gene as avian interferon-induced protein with tetratricopeptide repeats (*AvIFIT*).

Supplementary Tables

Supplementary Table 1 | Summary of the sequenced data of the duck genome

Paired-end libraries (bp)	Paired-end insert size (bp)	Number of libraries	Average read length (bp)	Number of reads ($\times 10^6$)	Total data (Gb)	Sequence coverage (fold)	Physical coverage (fold)
200	185-215	3	71	518	36.7	30.6	43.1
500	450-530	5	47	407	19.1	16	84.9
2K	2-2.4K	3	42	264	11.1	9.23	219.8
5K	5K	1	44	112	4.93	4.11	233.3
10K	10K	1	44	114	5.02	4.18	475.4
Total	All	13	50	1415	76.9	64.1	1,056

The inserted sizes of paired-end libraries were estimated by mapping the reads on the duck assembly. We calculated the sequence coverage on the assumption that the duck genome size is 1.2 Gb.

Supplementary Table 2 | Whole-genome assembly statistics

Category	Contig	Scaffold
Total length (kb)	1,069,961	1,105,049
Total number	227,597	78,487
Average size (bp)	4,701	14,079
Longest size (bp)	263,737	5,998,093
N50 size (bp)	26,114	1,233,631
N50 number	11,206	268
N90 size (bp)	3,062	195,458
N90 number	54,048	1,097

Supplementary Table 3 | Distribution of the supersaffolds in the duck assembly

Chromosome	Super-Scaffold[1]	Scaffold	Size (bp)	Strand	Chromosome	Super-scaffold[1]	Scaffold	Size (bp)	Strand
1	1	116	2434702	+	7	1	851	3430928	+
		1139	265011	-			165	886567	-
		1025	343334	-			525	197081	+
		1233	1670434	+			1870	561892	-
		782	818152	-			1645	1698420	-
		210	1392888	+			289	2058593	-
		318	1706257	-			591	705	+
		821	2635718	-			169	2459432	-
		2425	126468	+		2	519	5090982	+
		443	907825	+	8	1	1371	233100	-
		403	1233395	+			67	959237	-
		989	723475	-			93	1882263	+
		940	2865161	-			961	1436605	-
		96	1027715	-			1780	1514742	+
		52	2596476	+			1762	836180	+
		1567	477015	+			1009	2126665	-
		180	616817	+		2	1742	1289923	+
		492	544532	+	9	1	1708	164069	-
		2455	285071	+			1879	523590	-
		534	1178748	-			928	1292746	-
		570	216394	+	10	1	400	1553406	+
		779	928174	-			168	2032535	+
		291	1315036	-			**919**	**358021**	**-**
		912	356010	-			637	1727661	+
		2981	563311	+			811	504667	-
		23	1673207	-		2	**919**	**358021**	**-**
		728	190363	+			5835	363111	+
		884	2318530	+			1160	1401398	-
		307	1828611	+	11	1	**1668**	**1014780**	**+**

			2282	486803	+			1706	1254829	+
			415	2139235	-			126	3538160	+
			1232	1680643	-	12	1	769	935111	+
			738	759020	-	13	1	1517	661392	-
			207	880087	-			1045	685138	+
			1300	93112	-			170	548390	+
			110	1501635	-			144	2518879	-
			213	1916416	-			194	1476656	+
			4940	50796	-		2	1674	1198263	-
			674	1311412	+	15	1	1503	358267	-
			66	1461590	+			731	4130403	+
			611	954322	+	17	1	618	146673	-
			40	944271	+	18	1	1793	1112202	-
			1129	88709	+			1026	840053	-
			1198	1191883	-	20	1	559	1839884	+
			272	1020552	-			1308	844469	+
			11	1170681	-			1064	1128224	-
			969	1774812	+			193	934016	+
			439	760499	+			835	1343068	+
			10	2713871	+	22	1	871	1305515	-
		2	46	1697232	-			2665	111922	+
2		1	562	2401099	+	23	1	2430	441546	+
			121	2650348	+	28	1	C19616247	933	+
			266	1523081	-			C19730698	1075	+
			1438	1763677	-			**1451**	**489999**	**+**
			1035	818659	+			1984	74767	-
			208	595860	+		2	**1451**	**489999**	**+**
			2044	173969	-			2310	588663	-
			2213	1413698	+	Z	1	535	776299	-
			581	2309493	-			117	849200	+
			783	1805837	+			472	795694	+
			1709	425784	-			53	1995231	+

		395	1124893	+			1603	25765	+
		868	995657	-	Chromosome	Super-Scaffold[2]	Scaffold	Size	Strand
		316	3381015	-	9	1	1127	1301368	-
		1153	447100	+			2550	885470	-
		356	1511556	+			2367	1492560	-
		1152	971799	-	11	1	**1668**	**1014780**	**+**
		1034	634130	-			701	1270052	+
		129	2254433	+			428	946750	+
		72	2180811	+			522	1755914	-
		621	1257167	-			502	1230240	+
		643	869075	+	14	1	81	1264033	+
		1521	371169	-			1963	153544	-
3	1	171	2984079	-			3120	520880	-
		1091	2489873	-			2052	149303	-
		365	3588570	-			153	1249431	+
		432	1736154	-	16	1	1229	738052	+
		1303	536836	-			543	4012685	-
		453	1767481	-			455	1799516	-
		192	4356576	+			2914	9792	-
		34	4177285	+	19	1	1743	88479	+
	2	**629**	**2933642**	**-**			179	703014	-
		676	364651	+			202	1757398	+
		3231	131917	-			95	268734	-
	3	596	1385793	-	21	1	667	643900	-
		629	**2933642**	**-**			154	2047242	+
4	1	1208	2268538	-			594	1305795	+
		706	1406006	+	24	1	555	204263	-
		456	3365840	+			390	1819895	+
		42	3476961	-			1363	651972	+
		229	**4198679**	**+**			1332	765404	+
		215	1785268	-	25	1	1259	743608	+
		2530	**458617**	**+**			3812	170879	+

		734	1805680	-			1071	971454	+
	2	**2530**	**458617**	+	27	1	355	635212	-
		1335	602561	-			191	969684	+
		1205	773494	-			2667	236642	+
		347	3900964	+			1119	457412	-
		229	**4198679**	+	29	1	515	178952	+
		405	1473639	-			927	401098	+
		376	5998093	-			2097	448530	-
		1075	778635	-			430	480893	+
5	1	716	4002947	-	Chromosome	Super-Scaffold[3]	Scaffold	Size	Strand
		2336	544428	-	Micro	1	1663	527457	+
		870	3004054	-			394	138193	+
		773	1642078	+	Unknown	2	4811	22132	+
		286	2081588	-			1790	133486	-
	2	1828	220793	-			913	179527	-
		4611	166354	-			446	1086077	-
		997	1343329	-	Unknown	3	743	1768384	-
		2901	711746	+			498	1818954	+
		1358	911330	-			597	3554747	+
	3	1598	1395284	-	Unknown	4	276	2125295	+
6	1	1316	1257041	+			465	140324	+
		427	1626832	-	Unknown	5	906	306270	+
		293	1753535	+			188	56732	+
		319	2433623	-					

"[1]" represents super-saffolds assigned to chromosomes in order. "[2]" represents super-scaffolds assigned to chromosome, but orders of their scaffolds are not clear. "[3]" represents super-scaffolds that are ordered but are unanchored to chromosomes.

Supplementary Table 4 | Summary of the eight duck transcriptomes using Illumina GA and Roche 454 sequencing

Type	Solexa		454
	Liver	Spleen	
Number of reads	56×10^6	52×10^6	1.87×10^6
Total length of all reads (Gb)	3.7	3.5	0.5
Number of uniquely mapped reads	36×10^6	32×10^6	1.82×10^6
Number of multi-mapped reads	1.2×10^6	1.2×10^6	
Number of expressed genes	14,883	15,784	8,700
Number of assembled contigs	3.2×10^5		
Average length of the assembled contig	307 bp		

The liver and spleen transcriptomes were from a healthy 10-week-old female cherry valley duck. The detailed information for the six duck transcriptomes generated with the 454 Roche technology is in Supplementary Table 5.

Supplementary Table 5 | Summary of the sequence data and mapping results of the duck transcriptome using Roche 454 sequencing

Tissue	Number of reads	Total length (Mb)	Number of reads mapped to the genome	Number of reads mapped to genes	Number of genes supported by reads
Brain	113,365	44	83,740	32,876	4,050
Muscle	237,639	93	173,161	84,331	5,322
Intestine*	469,763	114	371,736	132,598	5,550
Spleen*	330,123	78	249,775	82,124	5,232
Lung*	390,856	96	305,992	89,141	5,017
Lung§	335,115	80	246,532	82,018	5,922
Combined	1,876,861	505	1,430,936	503,088	8,700

The brain and muscle transcriptomes were from each five 12-week-old individuals of I444 and I37 INRA duck lines. "*" and "§" represent the tissues from a 6-week-old Beijing duck infected by the BC500 H5N2 and VN1203 H5N1 viruses, respectively.

Supplementary Table 6 | General statistics of the gene sets and integrated predictions in human, chicken and duck

Gene sets		Total genes	Length of gene (bp)	Length of CDS (bp)	GC Ratio of CDS	Number of exons per gene	Length per exon (bp)	Length per intron (bp)
Human		14,128	20,047	1,524	0.49	8.9	170	2,332
Chicken		17,040	16,702	1,322	0.49	8	166	2,203
Duck	Genscan	32,383	23,625	1,345	0.51	8.3	162	3,049
	Augustus	22,739	18,200	1,122	0.53	6.6	169	3,025
	Integrated	19,144	20,574	1,345	0.49	8.2	164	2,664

The length of the gene represents the length of the CDS and the corresponding intron.

Supplementary Table 7 | Summary of homology-based RNA annotations in the duck, turkey, chicken and zebra finch genomes

RNA class	Functional Category	Duck	Turkey	Chicken	Zebra finch	Related reference
5S rRNA	Polypeptide synthesis	2 (+2 5'part)	4	5	42	This study
7SK	Transcription regulation	1	1	1	1	[3,4], this study
Antizyme_FSE	Frame shifting promotion	2	2	3	1	This study
CAESAR	Gene expression regulation	1	1	1-4	0-4	This study
HAR1F	Unknown	1	1	1-2	0-1	This study
Histone3	mRNA transport	0-33	0-26	25-40	1	This study
IRE	Iron metabolism	0-5	0-5	6-9	1	This study
IRES_Cx43	Cap independent translation	0-1	0-1	1-2	0-1	This study
IRES_APC	Apoptotic cascade	0-1	0-1	1-2	0-1	This study
miRNA	Translation control	323	416	461	270	[5], this study
NRON	Immune response	1	1	1-2	0-1	This study
RNase MRP	Mitochondrial replication, rRNA processing	0	1(3'part)	1	1	[6], this study
RNase P	tRNA processing	1(3'part)	0	1	1	[6], this study
SECIS	Selenocystein insertion	1-4	0-2	2-15	0-1	This study
SnoRNA U3	Nucleolar rRNA processing	1	1	1	1	[7], this study
other snoRNAs	processing	217	213	229	213	This study
SRP	Protein transportation	4	3	7	0	This study
Telomerase		0	0	1	0	This study
tRNA	Polypeptide synthesis	241	170	254	219	This study
U1	Splicing	3	3	1	2	[8], this study
U2	Splicing	2	2	1	5	[8], this study
U4	Splicing	2	2	1	2	[8], this study

U5	Splicing	2	1	2	3	[8], this study
U6	Splicing	3	2	4	2	[8], this study
U11	Splicing	1	1	1	1	This study
U12	Splicing	1	1	1	1	[8], this study
U4atac	Splicing	1	1	1	0	This study
U6atac	Splicing	1	1	1	1	This study
U7	Histone maturation	0	2	1	1	This study
vault RNA	Drug resistance	1	1	1	0	This study
Vimentin3	mRNA localization	1	1	1-4	1-4	This study
Y-RNA	DNA replication	3	3	3	3	This study

Supplementary Table 8 | The number of snoRNAs predicted by a homology-based approach

Species	Only *Homo sapiens*	Only *Gallus gallus*	Both	Total
Anas platyrhynchos	30	88	99	217
Gallus gallus	23	103	103	229
Meleagris gallopavo	32	74	107	213
Taeniopygia guttata	29	103	81	213
Anolis carolinensis	39	40	70	149
Danio rerio	37	15	34	86
Ornithorhynchus anatinus	105	43	88	236
Monodelphis domestica	91	33	74	198
Bos taurus	188	33	118	339
Mus musculus	185	29	118	332
Pan troglodytes	178	46	100	324
Homo sapiens	292	51	102	445

"Only *Homo sapiens*", "Only *Gallus gallus*", "Both" and "Total" represents the numbers of snoRNA homologs found by querying using the human snoRNA, chicken snoRNA, both human and chicken snoRNA, and total number of snoRNAs in the 12 listed species, respectively.

Supplementary Table 9 | Type and proportion of transposable elements (TEs) in the duck, chicken, zebra finch and human genomes

TE Type	Duck		Chicken		Zebra finch		Human	
	Length (mb)	% genome	Length (mb)	% genome	Length (mb)	% genome	Length (mb)	% genome
DNA	2.28	0.21	11.87	1.07	3.57	0.29	108.18	3.78
LINE	45.39	4.11	73.77	6.66	40.84	3.30	543.21	19.0
SINE	1.31	0.12	0.99	0.09	1.13	0.09	362.83	12.69
LTR	11.99	1.09	17.58	1.59	43.54	3.52	259.75	9.09
Other	0.05	0.0	0.03	0.0	0.01	0.00	11.69	0.41
Unknown	3.69	0.33	50.59	0.05	0.48	0.04	4.14	0.14
Total	64.67	5.85	104.72	9.45	89.57	7.25	1,289.79	45.12

Supplementary Table 10 | Distribution of the lineage-specific duplications in the duck

Family ID	Tree description	No. of LSDs[1]	No. of LSDs[2]
439534	PHD finger 7 testis development NYD SP6	3	1
64146	E3 sumo ligase inhibitor of activated STAT	3	2
921797	pyrin marenostrin	3	2
146790	polycystic kidney disease 2	3	-
754946	60S ribosomal l7	3	-
1841972	myosin heavy chain muscle	4	1
208527	keratin	4	3
981241	regakine 1	4	3
528978	**novel**	**5**	
423271	beta keratin related	6	3
423353	keratin	6	3
423441	keratin	9	4
1059131	**Immunoglobulin family**	**9**	-
1752102	**olfactory receptor**	**14**	**2**
21274	mucin muc intestinal mucin	-	1
24792	sec24 related	-	1
32137	c jun amino terminal kinase interacting 1 jnk interacting 1 jip 1 jnk map kinase scaffold 1 islet brain ib 1 mitogen activated kinase 8 interacting 1	-	1
43630	cas scaffolding family member 4	-	1
70526	transcription factor tcf transcription factor	-	1
128351	bmi1 proteinpredicted	-	1
184257	bruno 3 transcript variant	-	1
191144	deltex 3	-	1
213189	zinc finger suppressor of hairy wing homolog	-	1
246332	Singnal regulatory beta2, Tryosine phosphatase non-receptor type substrate	-	1
259607	histone h5	-	1
261434	zinc finger swim doamin containing	-	1
264863		-	1
270947	g coupled receptor 84	-	1
279775	histamine receptor	-	1
281515	mas related g coupled receptor member	-	1
332227	dual adapter phosphotyrosine and 3 phosphotyrosine and 3 phosphoinositide b cell adapter molecule of 32 kda b lymphocyte adapter bam32	-	1
363250	spermatogenesis associated 5	-	1
365148	mif4g domain containing	-	1
376687	ring finger 11	-	1

408972	aldehyde dehydrogenase aldehyde dehydrogenase family 3 member aldehyde dehydrogenase	-	1
432904	run and sh3 domain containing 1	-	1
449109	ubiquitin	-	1
482861	coiled coil domain containing 70	-	1
493031	alcohol dehydrogenase ec_1.1.1.1 alcohol dehydrogenase	-	1
493641	proteasome subunit alpha type 3 ec_3.4.25.1 proteasome subunit	-	1
498707	ubiquitin conjugating enzyme e2 r1 ec_6.3.2.19 ubiquitin ligase r1 ubiquitin conjugating enzyme e2.32 kda completmeting ubiquitin conjugating enzyme e2 cdd34	-	1
500040	traf interacting	-	1
500845	endoculcease viii 2 ec_3.2.2-ec_4.2.99.18 dna glycosylas/ap/lyase neil2 dna apurinic or apyrimidinic site lyase neil2 nei 2 nei homolog 2 neh2	-	1
520229	glutamate cysteine ligase regulatory subunit gamma glutamylcysteine synthetase regulatory subunit gamma ecs regulatory subunit gcs light chain glutamate cysteine ligase modifier subunit	-	1
552320	sugar phosphate exchanger 2 solute carrier family 37 member 2	-	1
560323		-	1
564990		-	1
580419	alpha amylase ec_3.2.1.1	-	1
585794	nad dependent deacetylase sirtuin 2 ec_3.5.1.-sir2	-	1
597116	liver fatty acid binding	-	1
646199	udp transporter solute carrier family 35 member	-	1
675769	tryptophan 5 dydroxylase ec_1.14.16.4 tryptophan 5 monooxygenase	-	1
676777		-	1
695115	ubiquitin associated 1 ubap 1	-	1
704594	dynactin subunit 3	-	1
719497	neurolysin mitochondrial ec_3.4.24.16 neurotensin endopeptidase mitochondrial oligopeptidase m microsomal endopeptidase mep	-	1
731395		-	1
736270	60s ribosomal I21	-	1
755451		-	1
761839	aurora borealis	-	1
773015	ribosomal s12	-	1
780261	fad dependent oxidoreductase domain containing 2	-	1

785145	vasculin gc rich promoter binding 1	-	1
797431	n acylneuraminate cytidyletransferase ec_2.7.7.43 cmp n acetylneuraminic acid synthase cmp neunac synthase	-	1
828138	active regulator of sirt1 40s ribosomal s19 binding 1 rps19 binding 1	-	1
830758	avidin	-	1
1022512	type opioid receptor or 1	-	1
1200071	acquaporin 8	-	1
1261143	mhc class ii b family peptide loading	-	1
1295656	btb/poz domain containing	-	1
1483408		-	1
1580773	chemokine binding 2 chemokine binding d6 cc chemokine receptor d6	-	1
1705946	d dopamine receptor dopamine receptor	-	1
1750140	leucine rich repeat containing 10	-	1
1896404	cytidine deaminase	-	1
1906918	nipped b scc2 homolog	-	1
438484	cd1 1 antigen splice variant	-	2

[1]Gene families had the number of lineage-specific duplications (LSDs) being larger than 2. [2]LSDs were identified using thresholds of lineage-specific homologous sequence identity < 97% and lineage specific dS < median dS. Gene families significantly ($p < 0.0005$) expanded in the duck are shown in bold.

Supplementary Table 11 | The enrichment of 440 positively selected duck genes, classified according to their molecular and cellular functions

Name	P value	Number of molecules
Amino acid metabolism	5.53E-05-3.83E-02	9
Small molecule biochemistry	5.53E-05-4.98E-02	36
Cellular assembly and organization	2.15E-04-4.40E-02	32
Cell signaling	3.66E-04-4.72E-02	8
Cellular function and maintenance	3.66E-04-4.72E-02	16

The categories of the positively selected duck genes, classified by their molecular and cellular functions, were analyzed using the Ingenuity Pathways Analysis (IPA) system.

Supplementary Table 12 | The enrichment of significantly differential expressed duck genes in two H5N1-viruses infections with known molecular and cellular functions

Group	Name	P value	Number of molecules
DK/49 infections vs Control (DEG set1)	Cell-To-Cell Signaling and Interaction	9.79E-21-4.27E-05	607
	Cellular Movement	1.97E-20-4.11E-05	687
	Cellular Function and Maintenance	2.03E-14-3.80E-05	785
	Antigen Presentation	9.80E-13-3.92E-05	292
	Cell Signaling	1.80E-12-1.71E-05	386
GS/65 infections vs Control (DEG set2)	Cell-To-Cell Signaling and Interaction	1.79E-17-1.10E-04	364
	Cellular Movement	4.91E-16-1.19E-04	415
	Antigen Presentation	3.56E-14-1.10E-04	195
	Cellular Development	4.18E-12-6.16E-05	566
	Cellular Function and Maintenance	6.10E-11-6.16E-05	299
DK/49 vs GS/65 infections (DEG set3)	Cellular Movement	7.65E-20-3.29E-06	428
	Cellular Development	6.74E-17-6.73E-06	649
	Cellular Growth and Proliferation	8.49E-17-6.73E-06	663
	Molecular Transport	6.02E-16-6.39E-06	502
	Lipid Metabolism	3.12E-15-2.46E-06	291

Significantly differential expressed genes (DEGs) identified in the DK/49-virus or GS/65-virus infections versus control were merged to DEG set1 (5,038) and set2 (2,741), respectively. DEGS detected in the DK/49-virus infections against GS/65-virus infections were combined into DEG set3 (3,232) (Table 1). The categories of DEGs with known molecular and cellular functions were analyzed using the Ingenuity Pathways Analysis (IPA) system.

Supplementary Table 13 | Description of genes responsive to influenza A virus in the lungs of ducks infected with one of two H5N1 viruses on days 1, 2 and 3 post-inoculation.

Gene	Full name	Gene	Full name
ADAR	adenosine deaminase, RNA-specific	*GDF7*	growth differentiation factor 7
ANGPT1	angiopoietin 1	*GDF9*	growth differentiation factor 9
ANGPT2	angiopoietin 2	*GDF10*	growth differentiation factor 10
AvDB1	avian defensin, beta 1	*GDF11*	growth differentiation factor 11
AvDB2	avian defensin, beta 2	*HGF*	hepatocyte growth factor
AvDB3	avian defensin, beta 3	*Anpl-DRA*	major histocompatibility complex, class II, DR alpha
AvDB3A	avian defensin, beta 3 type 1	*HSP90A*	heat shock protein 90kDa alpha (cytosolic), class A
AvDB3B	avian defensin, beta 3 type 2	*IFIH1*	interferon induced with helicase C domain 1
AvDB3C	avian defensin, beta 3 type 3	*IFITM10*	interferon induced transmembrane protein 10
AvDB3D	avian defensin, beta 3 type 4	*IFITM3*	interferon induced transmembrane protein 3
AvDB3E	avian defensin, beta 3 type 5	*IFITM5*	interferon induced transmembrane protein 5
AvDB3F	avian defensin, beta 3 type 6	*IFNA*	interferon, alpha
AvDB4	avian defensin, beta 4	*IFNE*	interferon, epsilon
AvDB5	avian defensin, beta 5	*IFNG*	interferon, gamma
AvDB6	avian defensin, beta 6	*IFNK*	interferon, kappa
AvDB7	avian defensin, beta 7	*IGF1*	insulin-like growth factor 1
AvDB8	avian defensin, beta 8	*IgM*	immunoglobulin heavy chain constant region (mu)
AvDB9	avian defensin, beta 9	*IL6*	interleukin 6

AvDB10	avian defensin, beta 10	*IL8A*	interleukin 8 type 1
AvDB11	avian defensin, beta 11	*IL8B*	interleukin 8 type 2
AvDB12	avian defensin, beta 12	*IL10*	interleukin 10
AvDB13	avian defensin, beta 13	*IL12A*	interleukin 12A
AvDB14	avian defensin, beta 14	*IL12B*	interleukin 12B
AvIFIT	avian interferon-induced protein with tetratricopeptide repeats	*IL13*	interleukin 13
BDNF	brain-derived neurotrophic factor	*IL17A*	interleukin 17A
BMP1	bone morphogenetic protein 1	*IL17D*	interleukin 17D
BMP2	bone morphogenetic protein 2	*IL18*	interleukin 18 interon-gamma-inducing factor
BMP3	bone morphogenetic protein 3	*IL19*	interleukin 19
BMP4	bone morphogenetic protein 4	*IL22*	interleukin 22
BMP5	bone morphogenetic protein 5	*IL28A*	interferon, lambda 2
BMP8	bone morphogenetic protein 8	*INHBA*	inhibin, beta A
BTNL	butyrophilin-like	*INHBB*	inhibin, beta B
CAMP	cathelicidin antimicrobial peptide	*INHBC*	inhibin, beta C
CCL4L2	chemokine (C-C motif) ligand 4-like 2	*KITLG*	KIT ligand
CCL5	chemokine (C-C motif) ligand 5	*LEFTY*	left-right determination
CCL6	chemokine (C-C motif) ligand 6	*LEP*	leptin
CCL17	chemokine (C-C motif) ligand 17	*LIF*	leukemia inhibitory factor
CCL19	chemokine (C-C motif) ligand 19	*MC1R*	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)

CCL20	chemokine (C-C motif) ligand 20	*MSTN*	myostatin
CCL21	chemokine (C-C motif) ligand 21	*NGFB*	nerve growth factor, beta
CCL23	chemokine (C-C motif) ligand 23	*NLRC3*	NLR family, CARD domain containing 3
CCL24	chemokine (C-C motif) ligand 24	*NLRC5*	NLR family, CARD domain containing 5
CCR7	chemokine (C-C motif) receptor 7	*NODAL*	nodal homolog
CD3E	*CD3e molecule, epsilon (CD3-TCR complex)*	*NRG2*	neuregulin 2
CD4	*CD4 molecule*	*NRG3*	neuregulin 3
CD40LG	CD40 ligand	*PDGFD*	platelet derived growth factor D
CD44	*CD44 molecule*	*PGF*	placental growth factor
CD8A	*CD8a molecule*	*TAP1*	transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)
CSF1R	colony stimulating factor 1 receptor	*TAP2*	transporter 2, ATP-binding cassette, sub-family B (MDR/TAP)
CSF2RA	colony stimulating factor 2 receptor, alpha	*TGFB2*	transforming growth factor, beta 2
CSF2RBA	colony stimulating factor 2 receptor, beta type 1	*TGFB3*	transforming growth factor, beta 3
CSF2RBB	colony stimulating factor 2 receptor, beta type 2	*TLR15*	toll-like receptor 15
CSF3R	colony stimulating factor 3 receptor	*TLR1B*	toll-like receptor 1 type 2
CX3CL1	C-X3-C motif chemokine 1	*TLR21*	toll-like receptor 21
CXCL12	chemokine (C-X-C motif) ligand 12	*TLR2A*	toll-like receptor 2 type 1
CXCL13L1	chemokine (C-X-C motif) ligand 13 like1	*TLR2B*	toll-like recetpor 2 type 2

CXCL13L2	chemokine (C-X-C motif) ligand 13 like 2	*TLR3*	toll-like receptor 3
CXCL14	chemokine (C-X-C motif) ligand 14	*TLR4*	toll-like recetpor 4
DDX58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58	*TLR5*	toll-like recetpor 5
DHX58	DEXH (Asp-Glu-X-His) box polypeptide 58	*TLR7*	toll-like receptor 7
EFNA1	ephrin-A1	*TNFSF4*	tumor necrosis factor (ligand) superfamily,member 4
EGF	epidermal growth factor	*TNFSF6*	tumor necrosis factor superfamily, member 6
EPO	erythropoietin	*TNFSF10*	tumor necrosis factor superfamily member 10
FGF8	fibroblast growth factor 8	*TNFSF11*	tumor necrosis factor (ligand)superfamily member 11
FGF9	fibroblast growth factor 9	*TRA*	T cell receptor alpha
FGF10	fibroblast growth factor 10	*TRD*	T cell receptor deta
FGF12	fibroblast growth factor 12	*TRG*	T cell receptor gama
FGF13	fibroblast growth factor 13	*UAA*	MHC class I antigen alpha chain, UAA gene
FGF18	fibroblast growth factor 18	*VEGFC*	vascular endothelial growth factor C
FGF23	fibroblast growth factor 23	*XCL1*	chemokine (C motif 1/2) ligand 1

Supplementary Note

Genome Sequencing and Assembly

The genomic DNA of a female Beijing duck (*Anas platyrhynchos*) (Gold Star Duck Production, Beijing, China) was extracted from 10 ml blood collected from a wing vein with the Puregene Tissue Core Kit A (Qiagen, Quesseldorf, Germany) according to the manufacture's protocol. This sample was used for whole genome shotgun sequencing with the Illumina GA Solexa technology. Similar to the methods used in the giant panda genome project[9], eight standard DNA libraries with a short insert size (185-530 bp) were constructed using the paired-end DNA sample prep kit (Illumina, California, USA), and five mate-paired libraries with a long insert size (2-10 kb) were constructed with the paired-end cluster generation kit V2 according to their corresponding manuals (Illumina, California, USA) (Supplementary Table 1).

The paired-end (PE) sequencing was performed on the Genome Analyzer platform as described in the manual (Illumina, California, USA). The clusters were generated using the Illumina cluster station. The workflow was as follows: template hybridization, isothermal amplification, linearization, blocking, sequencing primer hybridization, and sequencing of Read 1. After sequencing the first read, we prepared the second read as follows: denaturation, de-protection, re-synthesis, linearization, blocking, primer hybridization, and sequencing of the dsDNA fragments in the opposite.

Nature Genetics: doi:10.1038/ng.2657

We removed the duplicate reads introduced by the PCR and base-calling and the adapter sequences contained in the raw reads. Next, we assembled the duck genome with the high-quality reads using the pipeline developed in the giant panda genome project with SOAPdenovo[9]. Based on the high-quality reads of the duck assembly (Supplementary Table 2), we estimated the size of the duck genome to be 1.26 Gb according to the 17-mer frequency distribution, which was close to the C-value estimated by biochemical analysis in red blood cells[10] (Supplementary Figure 1-3).

To assess the quality of the duck assembly, we compared it with the duck sequences of seven BACs[11], 240 microsatellite markers[12], and 319,996 ESTs assembled in this project using BLASTN (E value $< 1\times10^{-5}$). These analyses suggested that this assembly covered more than 95% of the duck genome (Supplementary Figure 4). In addition**,** we aligned the duck and chicken assemblies to the human genome using Narcisse. This effort showed that the coverage of these two avian assemblies on the human genome was similar, indicating that the quality of the duck and chicken assemblies was comparable.

We then constructed super-scaffolds and created chromosomal sequences according to the duck genetic map[12] and the comparative physical map[13] using the following pipeline:

(1) Sequences of duck microsatellite markers and genes were assigned to the duck scaffolds by BLASTN

(2) Chicken BACs mapped to the duck chromosomes by FISH were assigned to the chicken genome sequence by the alignment of sequences of the BAC ends or microsatellite markers with BLASTN.

(3) The duck assembly was aligned to the chicken genome sequences by BLASTZ with its default settings[14].

(4) The super-scaffolds were positioned along the chromosomes based on the genetic and/or physical positions of their markers and initially oriented based on the relative marker order along the super-scaffolds.

(5) The super-scaffolds based on the genetic map were integrated with those based on the physical map using the cytogenetic map and the comparative genomic map between the chicken and duck.

(6) The comparative genomic map between the chicken and duck were used to aid in the orientation and confirm the order.

This effort resulted in the construction of a total of 47 super-saffolds, which contained 225 scaffolds and spanned 289 Mb (Supplementary Table 3).

Repetitive Element Annotation

The annotation of repetitive elements (TE) was performed using the following pipeline:

(1) The assembly was searched against the nucleotide repetitive database of Repbase (Release 14.07)[15] using RepeatMasker (version 3.2.6).

(2) The assembly was searched against the protein repetitive database provided in the RepeatMasker software using RepeatProteinMask.

(3) Tandem repeats were identified by Tandem repeat finder[16] using the defaults of "Match=2, Mismatch=7, Delta=7, PM=80, PI=10, Minscore=50, and MaxPeriod=12".

(4) RepeatModeler was used to construct a *de novo* duck repeat library, which was then used as the database to identify repetitive elements using RepeatMasker.

Genomic Variation

We mapped the raw reads on the duck assembly using SOAPaligner, and only the un-gapped mapped reads were selected to call heterozygous SNPs with SOAPsnp[17]. At each position, the frequencies of all potential alleles were calculated and the allele with the highest frequency was selected as the final allele. A rank sum test was applied to adjust the frequency of heterozygosity. The final frequencies were transformed to quality scores in the Phred scale. In addition, five thresholds were used to filter out the unreliable SNPs: (1) a quality cutoff of Q20; (2) an overall sequencing depth of less than 130; (3) a copy number of the flanking sequences < 2; (4) the number of the unique mapped reads for each allele > 5; and (5) an interval between the SNPs > 5 bp.

The SNPs were discovered from two sources:

(1) We identified 2,789,606 SNPs by mapping all the genomic reads on the duck assembly. We estimated the heterozygosity rate to be 2.61×10^{-3} for the autosomes and 2.08×10^{-3} for the coding regions. These heterozygosity rates were higher than those of mammals, such as the human assembly, with 0.69×10^{-3} for the autosomes

and 0.34×10^{-3} for the coding regions[18], and the panda assembly, with 1.35×10^{-3} for the autosomes and 0.66×10^{-3} for the coding regions[9].

(2) We mapped the transcriptome reads from the liver and spleen of one cherry valley duck to the duck assembly, which increased the number of SNPs to 2,957,169. Thus, the average was approximately 2.76 SNPs per kb along the 1.07 Gb genome. The fraction of SNPs in the intergenic, intronic, and exonic regions were 63%, 34.3% and 2.7%, respectively.

Reference Gene Sets

Two reference gene sets were predicted using the developed pipelines from BGI and ENSEMBL.

BGI Gene Set

We used five different approaches to the predict gene sets, and those gene sets were subsequently used to derive a final reference gene set for the duck assembly.

(1) EST-based gene prediction

All 319,996 ESTs assembled from the eight duck transcriptomes (see Assembly of Transcriptomes, Supplementary Table 4-5) were aligned to the duck assembly using BLAT[19]. We searched the best match for each EST, and removed the hits with identity < 95% or with coverage < 95%. Then, we used PASA[20] to assemble the hits and constructed a transcript set without ORFs.

(2) Transcriptome-based gene prediction

We mapped the transcriptome reads from the liver and spleen of the cherry valley duck (see Read Mapping to the Genome and Genes, Supplementary Table 4) to the duck genome using TopHat[21] with its default settings and constructed another transcript set using the mapped transcriptome reads with Cufflinks[22].

(3) Homology-based gene prediction (HSP)

We used the pipeline that built the human and chicken Ensembl genes databases (version 57) to predict the genes in the duck assembly using four steps:

(a) Rough alignment: We aligned the longest protein sequence of each human and chicken gene to the duck assembly with TBLASTN. We then built a gene-like structural library with a threshold of E value < 1×10E-5 and further extended the aligned regions at both ends by 500 bp.

(b) Precise alignment: We aligned the referenced protein sequences to the above gene-like structural library using GeneWise[23].

(c) Transcript building: We combined the transcripts that overlapped by more than 1 bp in the duck assembly. For each gene, the transcript having the highest coverage on its referenced protein was selected.

(d) Pseudogene filtering: We filtered the single-exon genes that were derived from retro-transposition and contained a frame error. For multi-exon genes, we removed those genes with more than 2 frame shift errors/in-frame stop codons.

(4) *De novo* gene prediction

We predicted genes using Genscan[24] and Augustus[25] with the defaults trained from *Homo sapiens*. We then filtered those genes with a threshold of coding length less

than 150 bp. Then, we aligned the predictions to a TE protein database using BLASTP (E value < 1×10^{-5}) and filtered the TE-derived genes with a coverage >50%.

(5) GLEAN gene set

We built a GLEAN gene set using GLEAN[26] based on the above gene sets.

(6) Final gene set

We clustered all the predicted transcripts to create a final gene set that was based on the homology set, *de novo* set, GLEAN set, and the syntenicy relationship between the chicken and duck assemblies (Blastz/chain/net).

The clustering process included the following steps:

a) For the chicken-based set, we kept genes that were located within the syntenic block and with an alignment length >30% of the target chicken gene length. The genes outside the syntenic region but with an alignment rate >70% were also retained. The human-based genes with an alignment rate ≤ 70% were removed. For the GLEAN and *de novo* sets, only the genes that overlapped with EST contigs longer than 100 bp were kept.

b) For each gene locus, we clustered all the remaining genes with a cutoff of genomic overlap greater than 1 bp. The selection of the gene at each locus conformed to the following order: GLEAN gene > chicken-based gene > human-based gene > *de novo* gene (Supplementary Table 6).

The BGI pipeline finally annotated 19,144 protein coding genes, 1529 psedogenes and 891 ncRNA genes. Comparing the final gene set to the merged gene set based on (3) and (4) (which contained 18,392 protein coding genes), we found that 1,058 protein

genes were only annotated based on either ESTs based prediction (1) and/or transcriptome based prediction (2). In addition, untranslated regions of 7,075 protein coding genes were defined using the duck transcriptomes (Supplementary Table 5). These observations supported that transcriptomes were important resources to the improvement of gene annotation.

ENSEMBL Gene Set

A reference gene set was also built using a modified version of the Ensembl genebuild pipeline (PMID: 15123590). The pipeline mainly relies on the alignment of proteins from both the target species and other species.

(1) The available duck protein sequences (1,369) from NCBI and Uniprot were aligned to the genome with Genewise (PMID: 15123596).

(2) Uniprot proteins of birds, mammals and other vertebrates filtered to only contain entries in the protein existence (PE) levels 1-3 were also mapped to the genome with Genewise.

(3) The canonical translations of the protein coding gene models for chicken in Ensembl release 56 were aligned to the genome using Exonerate (PMID: 15713233).

(4) The combined sets of the transcript models from (1), (2) and (3) were filtered according to the following: a. the transcript length; b. internal consistency in the models from (1), (2) and (3); c. comparisons of the splice site boundaries to those of the duck ESTs from NCBI, Uniprot and assembled in this project and the

chicken cDNA alignments.

(5) When no model was generated by this process, supplementary models were included from two sets: models generated by ortholog projection and models based on low-coverage Uniprot alignments. The ortholog projection involved identifying the Ensembl genes with orthologous relationships between human, mouse, chicken, dog and finch in release 56 and then attempting to project such genes from human to duck via a genomic alignment. The low-coverage Uniprot alignments were those with coverage scores between 30 and 70. Typically, a cut-off of 70 was used.

(6) The combined models from (4) and (5) were scanned for pseudogenes and merged with the results of the Ensembl non-coding RNA pipeline.

The Ensembl pipeline finally annotated 15,634 protein coding genes and 249 pseudogenes.

Annotation of Non-coding RNAs

We used RNAfold, RNAcofold, RNAlifold and RNAplulex from the ViennaRNA package[27] to determine the putative structure of non-coding RNAs. The non-coding RNA reference sequences were collected from NCBI, Rfam, and miRBase[28], as well as from the resource reported by Xie et al.[29] and Shao et al.[30].

tRNA

Using tRNAscan-SE[31] with its defaults, we predicted a total of 241 tRNAs in the duck. This repertoire is similar to those of the chicken (254) and zebra finch (219) but

slightly larger than that of the turkey (170) (Supplementary Table 7).

snoRNA

Small nucleolar RNAs (snoRNAs) are one of the most abundant groups of ncRNAs in the genome. Their main function is to guide the modification of other ncRNAs, mainly ribosomal RNAs (rRNAs), transfer RNAs (tRNAs) and small nuclear RNAs. There are two main classes of snoRNA: the C/D box snoRNAs, which are associated with methylation, and the H/ACA box snoRNAs, which are associated with pseudouridylation. Using the human and chicken snoRNAs[29-30], we performed homologous searches in the chimpanzee, mouse, cattle, opossum, platypus, chicken, turkey, zebra finch, duck, lizard and zebra fish. The thresholds for the sequence identity, minimal relative length and E value of the Blast-hits were 50%, 50% and 1E-03, respectively. We then predicted the structures of the Blast-hits with mlocARNA[32] and filtered those Blast-hits not containing the typical H/ACA- or C/D-boxes. For the duck snoRNA set, we combined the query sets based on the human and chicken snoRNAs, removed the redundancies and merged the overlapping sequences. This effort identified a total of 217, 229, 213 and 213 snoRNAs in the duck, chicken, turkey and zebra finch, respectively (Supplementary Table 7-8). Among the 217 duck snoRNAs, 131 are C/D-box snoRNAs distributed among 86 families, and 86 are H/ACA-box snoRNAs among 73 families. Interestingly, that 9 H/ACA and 16 C/D-box snoRNA families are only observed in these four bird species. The detailed analysis suggested that the family size of 128 of 176 snoRNAs families

in the duck was same as the corresponding features in the chicken. This observation, together with the observation that 90% (139/155) of the chicken annotated snoRNAs families[30] were found in the duck, suggested that the snoRNAs were conserved between these two species. However, 25 and 23 families presented a slight expansion in the duck and chicken, respectively, resulting in a change in the family size of less than two between them. Using the functional annotation of the chicken snoRNA[28], we inferred that approximately half of the duck snoRNAs might bind to an antisense element on the 28S rRNA.

microRNA

Similar to the snoRNA approach, we identified homologs in the chicken, turkey, duck and zebra finch through Blast[33] using 465 chicken pre-microRNA reference sequences from the mirBase database[28]. All identified microRNA homologs in one organism were added to the query set for the next organism according to the phylogenetic tree. We then derived the homologous microRNA from the Blast-hits using three thresholds: (1) sequence identity > 85%; (2) E value < 1×10^{-3}; (3) coverage of the queried sequence > 90%. This procedure detected 461, 416, 323 and 270 microRNAs in the chicken, turkey, duck and zebra finch, respectively (Supplementary Table 7). As expected, some of the annotated chicken microRNAs are not confirmed with our pipeline, possibly due to an annotation error.

Other RNA Families

We also identified other ncRNA families using the following methods:

(a) To identify ribosomal RNAs, splicedosomal RNAs and SRP RNA, we performed a homologous search with a threshold *E* value < 1E-3 using ncRNA genes from the NCBI, Rfam[34] and Noncode databases[35]. For the duck ncRNA sets, we performed an additional homologous search to identify paralogs. We then predicted the snRNAs by aligning all Blat-hits to the annotated snRNAs[8]. For the 7SK RNA, the 5S and 5.8S rRNAs we used against the complete set of Rfam entries; for the SSU and LSU rRNAs, we also used chicken homologs from NCBI. For more diverged genes, such as minor snRNAs, RNase MRP, masc/men RNA, U7 snRNA and telomerase, we used GotohScan[6] in addition. In cases where no good candidates were found, we also employed descriptor-based search tools, such as RNABOB[36].

(b) In the second step, the known and predicted sequences were aligned using ClustalW[37]. To identify functional secondary structures, RNAfold, RNAduplex, RNAalifold, and RNAcofold were used. The combined primary and secondary structures were visualized in the emacs editor using the ralee-mode[38] and manually checked.

(c) Putatively functional sequences were distinguished from likely pseudogenes by analyzing the flanking genomic sequence. For the polymerase III transcripts (U6 snRNA, U6atac RNA, 7SK RNA, 5S RNA, RNase MRP, RNase P, vault RNA, and Y-RNA), a promoter analysis and the pattern detecting tool MEME[39] were used to identify the TATA box and PSE element specific for duck in the 100nt upstream region.

(d) Additional consistency checks were employed for individual RNA families,

including phylogenetic analysis by neighbor-joining[40].

(e) Accepted candidate sequences were compared by Blast with the chicken, turkey and zebra finch genomes to determine their copy numbers in the genome assembly.

This procedure identified 77, 71, 36 and 36 other ncRNAs in the chicken, zebra finch, duck and turkey, respectively (Supplementary Table 7).

Y-RNA Cluster

Y-RNAs are RNA components of the Ro RNP particle[41] and form a small family of short polymerase III transcripts that are grouped into a gene cluster in tetrapods[42]. A BLAST search using the known vertebrate Y-RNAs uncovered four loci in duck, with one being a Y3 pseudogene. The other three loci were identified unambiguously as homologs of the human Y1, Y3, and Y4 genes (using a ClustalW alignment and neighbor-joining to infer the gene phylogeny with 1000 bootstrap replicates).

The Y-RNA cluster is located anti-sense between the *EZH2* and *PDIA4* protein-coding genes, an arrangement that is conserved among sauropsids. Of note, the distances between *PDIA4* and Y1 and between Y1 and Y3 were more constrained than the distances between the other members of the cluster. Each Y-RNA had its own polymerase III promoter sequence consisting of a TATA box and a PSE element.

Other RNA Motifs

We also looked for motifs in other RNA families annotated by RFAM in the chicken and zebra finch (Supplementary Table 7). No drastic changes in the numbers within

birds have been observed. Compared with other tetrapods, the spliceosomal RNAs showed a significantly reduced number of pseudogenes in birds.

Gene Function Annotation

The duck reference genes were searched against the databases of BlastProDom, Coil, FPrintScan, Gene3D, HMMPanther, HMMPfam, HMMPIR, HMMSmart, HMMTigr, ProfileScan, ScanRegExp and Superfamily by InterPro (version 18.0)[43] to annotate all the motifs and domains. The gene descriptions were classified according to the Gene Ontology annotation[44], which was extracted using the InterPro output. The genes were also compared with the SwissProt/TrEMBL (Release 14.1) and KEGG (Release 51) databases[45] using BLASTP (E value < 1×10^{-5}) (Supplementary Figure 5).

Genome Evolution Analysis

Homolog Identification

We constructed gene families using TreeFam[46]. One gene family is defined as a group of genes that descended from a single gene in the last common ancestor and with an outgroup gene standing on the edge of the family tree. Nine predicted gene sets from the human, mouse, cattle, platypus, chicken, duck, zebra finch, lizard and frog (outgroup) were used to identify the orthologs and paralogs. All gene sets, except the duck set, were downloaded from Ensembl Release 55. We examined the conserved genes in the genomes of the duck, chicken, zebra finch, human, mouse and cattle. The orthology was resolved for >75% of the duck genes (Supplementary Figure 6). There

are 25,229 orthologous groups with representatives in these six species, which represent 14,402 duck, 15,122 chicken, and 16,777 zebra finch genes. Among the 25,229 groups, 5,646 are strictly 1:1 orthologous. We also identified 711 avian-specific orthologous groups, many of which were related to the cytoskeleton, transmembrane receptor activity, intermediate filament, integral to membrane and cell surface receptor linked signal transduction. Moreover, we predicted 4,742 duck "orphan" genes, with 4,092 of these being supported by EST sequences (Supplementary Figure 5-6).

GeneTree analysis

We built gene families[47] using the genomes in Ensembl 59 plus the duck and turkey genomes. In brief, we grouped proteins based on their Smith-Waterman pairwise alignment score and aligned them using M-Coffee[48]. We then used TreeBeST to construct five phylogenetic trees with different combinations of evolutionary and substitution models. TreeBeST merged the trees guided by the species tree. It also runs a reconciliation step in which the resulting gene tree is compared with the species tree so that duplication and speciation events can be inferred. We obtained 19,549 gene families with the predicted genes in Ensembl 59 and the duck and turkey predicted gene sets. Among these, 7,944 families contained 15,279 duck genes, and the remaining 7,335 gene families lacked any representative in the duck assembly.

We extracted the orthologs and paralogs from the above 19,549 gene families. Any two genes related to a speciation event are defined as orthologs, whereas those related

to a duplication event are defined as paralogs. The quality of the assemblies were affected by issues such as gaps and incorrectly ordered contigs, and some genes were predicted as several split genes with the available genome sequences. We detected and removed the split genes from the final list of paralogs by checking the homologous gene sets using the following thresholds: (1) the number of homologs in one species was slightly larger than that in other species; (2) the length of multiple homologs in one species was slightly shorter than the corresponding length in other species; and (3) there was no overlap between the multiple homologs in one species.

We next focused on paralogs in the four avian genomes. The time of duplication event was inferred from the gene family trees. A large number of lineage-specific duplications were observed in both the chicken and the zebra finch lineages. Further investigation showed that the majority of these recent duplications involved genes in the unassembled genomes. After filtering these genes, the results were much more consistent across the species.

Evolutionary Analysis of Gene Families

To estimate the changes in the gene repertoire in the duck and the other three avian genomes, we inferred the most likely gene family size at all internal nodes, calculated the global birth and death rates of gene families and characterized the lineage-specific gene duplications (LSDs) using 15,751 gene families from 17 species and CAFÉ (computational analysis of gene family evolution) tool[49].

Likelihood analysis of gene repertoire

First we applied an updated version of the likelihood model developed by Hahn et al[49-50] and 15,751 gene families constructed through the above genetree analysis to estimate the rates of gene gain and loss (Figure 1). The method models gene family evolution as a stochastic birth-to-death process, taking into account the phylogenetic tree topology and branch lengths. Assuming that all genes have an equal probability of changing from an initial number of genes, $X_0 = s$, to size c over time t, $X_t = c$ is given by

$$P(X_t = c \mid X_0 = s) = \sum_{j=0}^{\min(s,c)} \binom{s}{j} \binom{s+c-j-1}{s-1} \alpha^{s+c-2j} (1-2\alpha)^j$$

where $\alpha = \lambda t \div (1+\lambda t)$. Because $X_0 = 0$ will result in a probability of zero for birth and death, we restricted this analysis to families in which $X_0 > 0$. Thus, lineage-specific families were excluded from this likelihood analysis. We also tried various models assuming a single or multiple rates for each of the major groups (mammals, birds, fish and amphibians). This analysis suggested that the A4 parameter model maximized the likelihood (p value << 0.01), with λ values of 0.0019, 0.0017, 0.0012 and 0.0011 for amphibians, mammals, fish and birds, respectively.

We then calculated the number of gene gains and losses on each branch by comparing the sizes of all parent-daughter node pairs using the maximum likelihood sizes of the ancestral gene family. The difference in size between these two values was inferred to be the number of genes gained or lost: a larger daughter size implies a gene gain, whereas a smaller daughter size implies a gene loss.

Identifying lineage-specific duplications (LSDs)

Two methods were used to count the LSDs.

First we identified the LSD gene families and counted the number of LSDs with a threshold of 2. In this case, only gene families for which the number of recent duplicated genes for one species was > 2 were counted. This analysis found 5, 76, 577 and 1752 LSDs in the turkey, duck, chicken and zebra finch, respectively.

Second, we filtered the gene families with thresholds of lineage-specific homologous sequence identity <97% and lineage-specific dS < median dS. This filter identified 11, 88, 88 and 999 LSDs in the turkey, duck, chicken and zebra finch respectively.

Comparison in LSDs showed that large numbers of duplication events counted using the first method in four families was disappeared when they counted using the second method (Supplementary Table 10). Among them, one, three, three and twelve LSDs in the 146790, 754946, 528978 and 1059131 families respectively were transcribed in either the lung, brain and/or spleen tissues of the duck. Molecular phylogenetic analysis and sequence alignment of the duck genes of the 1059131 family on the reference gene set and assembly (both the galGal3.0 and galGal4.0) of the chicken further supported the significant expansion of the BTNLs in the duck (Supplementary Figure 7C). Similarly, large number of duplication events (14) in the olfactory family counted using the first method was lessened to a small number (2) when it counted using the second method. These observations suggested that thresholds (lineage-specific homologous sequence identity >97% and lineage-specific dS < median dS) of the second method might be too strict to count LSDs.

Evolutionary analysis of orthologous genes

We downloaded the 1:1 orthologous gene sets from ENSEMBL for the chicken, turkey and zebra finch. The 1:1 orthologous gene sets for the duck and the other three birds was created by reciprocal best hit analysis between the duck and chicken. As a result, a total of 8,409 1:1 orthologs for the four birds were collected. The phylogenetic trees were obtained from Timetree.

Orthologous gene sets were aligned using prank with its default settings[51], and poorly aligned sites were eliminated using Gblocks[52]. Then, we used the maximum likelihood method (Codeml of PAML 4[53]) to estimate the dN (rate of non-synonymous substitutions), dS (rate of synonymous substitution) and ω (ratio of non-synonymous substitutions to the rate of synonymous substitutions) with the F3X4 codon frequencies under the branch-site model (model =2, NSsites =2). Orthologs with dS >3 or ω >5 were filtered[54].

Cytokine Analysis

Cytokines were identified using the following three steps:

(1) Collect the cytokines in the gene sets of the human, mouse, duck, chicken and zebra finch using TreeFam according to the cytokine genes list in KEGG;

(2) Identify cytokines by aligning the reference cytokine protein sequences to the genome sequence and NCBI nr database using BLAST; and

(3) Search the homologs of the cytokine genes in the above five genomes using the known motifs from the Pfam database[55].

Nature Genetics: doi:10.1038/ng.2657

The identification and comparison of the cytokine genes suggested the cytokine repertoire in birds was more succinct than that of mammals (Table 2).

Transcriptome Analysis using Roche 454 Sequencing

Data Preparation

Six duck transcriptomes were separately sequenced using Roche 454 technology.

(1) Five animals each from the I444 and I37 INRA duck lines, which were involved in a QTL cross[56], were fed *ad libitum*. After slaughtering, muscle and brain samples were harvested, immediately frozen in liquid nitrogen and stored at -80°C. Total RNA was extracted using the NucleoSpin RNA L kit (Macherey-Nagel EURL). Double-strand cDNA was prepared, and polyT tails were removed as described[57] from 100 μg of total RNA. The fragments were then sequenced using the Roche 454 Life Sciences Genome FLX Sequencer following the manufacturer's instructions for the Titanium series (454 Life Science, Roche), slightly modified[57].

(2) Infections were performed as previously described[58]. Briefly, the H5N1 A/Vietnam/1203/04 HPAI was generated by reverse genetics, and H5N2 A/mallard/BC/500/05 LPAI was isolated by screening environmental samples. Outbred White Beijing ducks (*A. platyrhynchos*) were purchased from Ideal Poultry or Metzer Farms and inoculated at six weeks of age. A total of 10^6 of 50% egg infectious doses of BC500 and VN1203 were administered via the natural route, in the nares, eyes, and trachea. The ducks were killed, and the tissues were

collected at day 3 post infection. Tracheal and cloacal swabs were collected to monitor the viral shedding. Intestine, lung and spleen samples from the ducks infected by H5N2 and lung sample from the duck infected by H5N1 were collected. RNA was extracted from tissues using TRIzol according to the manufacturer's instructions. Double-strand cDNA was prepared, and the polyT tails were removed. The fragments were then sequenced using the Roche 454 Life Sciences Genome FLX Sequencer following the manufacturer's instructions for the Titanium series (454 Life Science, Roche).

Data Analysis

Six transcriptomes produced using 454 sequencing technology were mapped to the duck assembly by GMAP[59], using its default parameters, and to the predicted BGI reference gene set by BLASTN (E value < 1×10^{-5}).

Transcriptome Analysis using Illumina Sequencing

Data Preparation

The cDNA libraries were prepared according to the manufacturer's instructions (Illumina). The mRNA samples of the liver and spleen from a 10-week-old-female cherry valley duck were purified from total RNA using Dynal Oligo(dT) bead and fragmented into small pieces of approximately 200 nucleotides using RNA Fragmentation Reagents (Ambion). The cleaved mRNA fragments were converted into single cDNAs using SuperScript II (Invitrogen) and primed with random primers, and then double-strand cDNA was synthesized using RNaseH (Invitrogen) and DNA

Pol I (Invitrogen). Subsequently, the cDNA was subjected to end-repair and phosphorylation using Klenow polymerase (Enzymatics), T4 DNA polymerase (Enzymatics) and T4 polynucleotide kinase (to blunt-end the DNA fragments) (Enzymatics). These end-repaired cDNA fragments were 3'-adenylated using Klenow (exo-) DNA polymerase (Enzymatics). Then, Illumina PE adapters were ligated to the ends of these 3'-adenylated cDNA fragments. Gel-electrophoresis was used to separate the cDNA fragments from any unligated adapters. Those cDNA fragments with a size between 180–220 bp were selected. The cDNA libraries were amplified by 12 cycles of PCR with Phusion polymerase (NEB), and the 75 cycle paired-end sequencing was performed on the Illumina Genome Analyzer.

Read Mapping to Genome and Genes

After removing duplicate reads and the adapter sequence contained in the raw reads, we aligned the high-quality reads to the genome using SOAPaligner with a threshold of three mismatches. For the multi-position hits, one of the best matching loci was chosen randomly. Only the unique mapped reads were used for the gene expression level analysis. The insert size used to map the high quality reads on the duck assembly and the predicted genes are set as 0~10,000 bp and 0~1,000 bp, respectively. These results are summarized in Supplementary Table 4.

Assembly of Transcriptomes

All high-quality short reads of the two transcriptomes performed with Illumina Genome Analyzer were assembled using the SOAPdenovo software, producing 339,803 contigs. Subsequently, these assembled contigs and six transcriptomic reads

sequenced with 454 Roche technology were re-assembled using the Phrap software. Finally, we obtained 319,996 contigs with an average length of 307 bp and used these contigs to evaluate the genome assembly and predict the duck genes.

URLs

Narcisse, http://narcisse.toulouse.inra.fr; RepeatMasker and RepeatModeler, http://www.repeatmasker.org; Ensembl 59, http://e59.ensembl.org; TreeBeST, http://treesoft.sourceforge.net/treebest.shtml; Timetree, www.timetree.org; KEGG, http://www.genome.jp/kegg/; ENSEMBL, http://www.ensembl.org/; Phrap, http://www.phrap.org/; Uniprot, http://www.uniprot.org/docs/pe_criteria; Ingenuity Systems Pathway Analysis (IPA), Ingenuity® Systems, www.ingeuity.com.

References

1. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704 (2003).
2. Letunic, I., Doerks, T. & Bork, P. SMART 6: recent updates and new developments. *Nucleic Acids Res* **37**, D229-32 (2009).
3. Gruber, A.R. et al. Invertebrate 7SK snRNAs. *J Mol Evol* **66**, 107-15 (2008).
4. Marz, M. et al. Evolution of 7SK RNA and its protein partners in metazoa. *Mol Biol Evol* **26**, 2821-30 (2009).
5. Hertel, J. et al. The expansion of the metazoan microRNA repertoire. *BMC Genomics* **7**, 25 (2006).
6. Hertel, J. et al. Non-coding RNA annotation of the genome of Trichoplax adhaerens. *Nucleic Acids Res* **37**, 1602-15 (2009).
7. Marz, M. & Stadler, P.F. Comparative analysis of eukaryotic U3 snoRNA. *RNA Biol* **6**, 503-7 (2009).
8. Marz, M., Kirsten, T. & Stadler, P.F. Evolution of spliceosomal snRNA genes in metazoan animals. *J Mol Evol* **67**, 594-607 (2008).
9. Li, R. et al. The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311-7 (2010).
10. Vendrey, R. & Vendrey, C. Sur la teneur absolue en acide desoxyribonucleique du noyau cellulaire chez quelques especes doiseaux et de poissons. *Comptes Rendus de l'Academie des Sciences* **230**, 788-790 (1950).
11. Huang, Y. et al. Molecular evolution of the vertebrate TLR1 gene family--a complex history of gene duplication, gene conversion, positive selection and co-evolution. *BMC Evol Biol* **11**, 149 (2011).
12. Huang, Y. et al. A genetic and cytogenetic map for the duck (Anas platyrhynchos). *Genetics* **173**,

287-96 (2006).

13. Skinner, B.M. et al. Comparative genomics in chicken and Pekin duck using FISH mapping and microarray analysis. *BMC Genomics* **10**, 357 (2009).
14. Schwartz, S. et al. Human-mouse alignments with BLASTZ. *Genome Res* **13**, 103-7 (2003).
15. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-7 (2005).
16. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-80 (1999).
17. Li, R. et al. SNP detection for massively parallel whole-genome resequencing. *Genome Res* **19**, 1124-32 (2009).
18. Wang, J. et al. The diploid genome sequence of an Asian individual. *Nature* **456**, 60-5 (2008).
19. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).
20. Haas, B.J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654-66 (2003).
21. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-11 (2009).
22. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-5 (2010).
23. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988-95 (2004).
24. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in Drosophila genomic DNA. *Genome Res* **10**, 516-22 (2000).
25. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel.

Bioinformatics **19 Suppl 2**, ii215-25 (2003).

26. Elsik, C.G. et al. Creating a honey bee consensus gene set. *Genome Biol* **8**, R13 (2007).
27. Hofacker, I.L. et al. Fast folding and comparison of RNA secondary structures. *Monatsch. Chem* **125**, 169-177 (1994).
28. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**, D154-8 (2008).
29. Xie, J. et al. Sno/scaRNAbase: a curated database for small nucleolar RNAs and cajal body-specific RNAs. *Nucleic Acids Res* **35**, D183-7 (2007).
30. Shao, P., Yang, J.H., Zhou, H., Guan, D.G. & Qu, L.H. Genome-wide analysis of chicken snoRNAs provides unique implications for the evolution of vertebrate snoRNAs. *BMC Genomics* **10**, 86 (2009).
31. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-64 (1997).
32. Will, S., Reiche, K., Hofacker, I.L., Stadler, P.F. & Backofen, R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol* **3**, e65 (2007).
33. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).
34. Griffiths-Jones, S. et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, D121-4 (2005).
35. Liu, C. et al. NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res* **33**, D112-5 (2005).
36. Riccitelli, N.J. & Luptak, A. Computational discovery of folded RNA domains in genomes and in

vitro selected libraries. *Methods* **52**, 133-40 (2010).

37. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
38. Griffiths-Jones, S. RALEE--RNA ALignment editor in Emacs. *Bioinformatics* **21**, 257-9 (2005).
39. Bailey, T.L., Williams, N., Misleh, C. & Li, W.W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* **34**, W369-73 (2006).
40. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-25 (1987).
41. Lerner, M.R., Boyle, J.A., Hardin, J.A. & Steitz, J.A. Two novel classes of small ribonucleoproteins detected by antibodies associated with lupus erythematosus. *Science* **211**, 400-2 (1981).
42. Mosig, A., Guofeng, M., Stadler, B.M. & Stadler, P.F. Evolution of the vertebrate Y RNA cluster. *Theory Biosci* **126**, 9-14 (2007).
43. Mulder, N.J. et al. New developments in the InterPro database. *Nucleic Acids Res* **35**, D224-8 (2007).
44. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).
45. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).
46. Li, H. et al. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* **34**, D572-80 (2006).
47. Vilella, A.J. et al. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**, 327-35 (2009).

48. Wallace, I.M., O'Sullivan, O., Higgins, D.G. & Notredame, C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* **34**, 1692-9 (2006).

49. Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res* **15**, 1153-60 (2005).

50. De Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269-71 (2006).

51. Loytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A* **102**, 10557-62 (2005).

52. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564-77 (2007).

53. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-91 (2007).

54. Castillo-Davis, C.I., Kondrashov, F.A., Hartl, D.L. & Kulathinal, R.J. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. *Genome Res* **14**, 802-11 (2004).

55. Finn, R.D. et al. The Pfam protein families database. *Nucleic Acids Res* **38**, D211-22 (2007).

56. Vitezica, Z.G., Marie-Etancelin, C., Bernadet, M.D., Fernandez, X. & Robert-Granie, C. Comparison of nonlinear and spline regression models for describing mule duck growth curves. *Poult Sci* **89**, 1778-84 (2010).

57. Leroux, S. et al. Non PCR-amplified Transcripts and AFLP fragments as reduced representations of the quail genome for 454 Titanium sequencing. *BMC Res Notes* **3**, 214 (2010).

58. Barber, M.R., Aldridge, J.R., Jr., Webster, R.G. & Magor, K.E. Association of RIG-I with innate immunity of ducks to influenza. *Proc Natl Acad Sci U S A* **107**, 5913-8 (2010).

59. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-75 (2005).