


CORRESPONDENCE

Open Access



Correspondence on Lovell et al.: identification of chicken genes previously assumed to be evolutionarily lost

Susanne Bornelöv^{1,2}, Eyal Seroussi³, Sara Yosefi³, Ken Pendavis⁴, Shane C. Burgess⁴, Manfred Grabherr^{1,5},
Miriam Friedman-Einat^{3*}  and Leif Andersson^{1,6,7*}

Please see related Research article: <http://dx.doi.org/10.1186/s13059-014-0565-1> and Please see response from Lovell et al: <https://www.doi.org/10.1186/s13059-017-1234-y>

Abstract

Through RNA-Seq analyses, we identified 137 genes that are missing in chicken, including the long-sought-after nephrin and tumor necrosis factor genes. These genes tended to cluster in GC-rich regions that have poor coverage in genome sequence databases. Hence, the occurrence of syntenic groups of vertebrate genes that have not been observed in Aves does not prove the evolutionary loss of such genes.

A recent paper reported that 274 protein-encoding genes were missing from sequencing data from 60 bird species [1]. Most of them were organized in conserved syntenic clusters in non-avian vertebrates, suggesting that their loss in the avian lineage had occurred through genomic deletions of gene blocks. This hypothesis was supported by another study reporting that 640 protein-encoding genes were missing from 48 bird genomes [2]; the authors of this second study made a similar suggestion that large segmentally deleted regions had been lost during microchromosome evolution in birds. However, our recent discovery of leptin genes with ~70% GC content in chicken and duck [3], and the new identification of 89 GC-rich genes [4], suggested an alternative hypothesis of a technical barrier to explain the ‘missing genes’. To further explore this, RNA-Seq data from visceral fat, hypothalamus, and pituitary tissues from two types of chickens,

broilers and layers (Additional file 1: Table S1), were used for de novo transcriptome assembly and identification of novel genes.

The initial set of 588,683 transcripts obtained using Trinity [5] was reduced to 257,700 after removing transcripts that were expressed at low levels. We mapped the transcripts to the chicken reference genome build consistent with the previous studies [1, 2] using Blat and Blast, and retained 8395 sequences without alignments. These transcripts were then characterized on the basis of sequence similarity to known genes in other vertebrates using the Trinotate pipeline (<https://trinotate.github.io>), which searches for sequences encoding known protein domains, transmembrane domains, and signal peptides (Additional file 1: Tables S2 and S3a). Genes that were already known in chicken were removed by comparing their gene symbols with those in Ensembl (release 80), RefSeq, and Entrez Gene, resulting in 1878 novel gene-candidate transcripts representing 1063 genes (Additional file 1: Tables S3b and S4).

To increase specificity and to remove multiple transcript isoforms, we tested each transcript by reciprocal Blastn against the full transcriptome assembly (588,683 transcripts), and Blastx against the set of coding sequences predicted by TransDecoder (<https://transdecoder.github.io>), consisting of 111,457 sequences. The remaining set yielded 194 transcripts encompassing 190 distinct high-confidence genes (Additional file 1: Table S5). Through Blastn, we found that 55 loci had already been recovered as annotated genes in an updated genome build (Galgal5) released after the previous studies. In addition, 47 genes mapped to the genome but lacked annotations, while another 51 genes were annotated as uncharacterized or putative proteins (Additional file 1: Table S6). One

* Correspondence: miri.einat@mail.huji.ac.il; leif.andersson@imbim.uu.se

³Agricultural Research Organization, Volcani Center, Rishon LeZion, Israel

¹Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala SE-751 23, Sweden

Full list of author information is available at the end of the article

discrepancy in annotation between our genes and Galgal5 was observed for the *RSADI* transcript, which was annotated as *MYCBPAP* in Galgal5. Closer inspection revealed that these two genes, which are close neighbors in the human genome, have been mistakenly merged into *MYCBPAP* in Galgal5. Therefore, we considered *RSADI* as a novel annotation (Additional file 1: Table S6).

Among the remaining 38 genes (Additional file 1: Table S6) with no sequence similarity to any genome build are the tumor necrosis factor (*TNF*) and nephrin (*NPHS1*), which have been reported as missing from birds in several studies (Table 1) but which are critically important in vertebrate biology and have extensively been studied in non-avian vertebrates (there are more than 130,000 publications in PubMed on *TNF* and 1300 on *NPHS1*). These genes were subjected to full-cDNA-sequence determination, exon characterization, RT-PCR validation, and expression profiling using RNA-Seq data from red junglefowl (Additional file 2: Figures S1 and S2; Additional file 2: Tables S9 to S12). The similarity in sequences, exon–intron junctions, and characteristic expression profiles confirmed the identification of chicken *NPHS1* and *TNF*, thus resolving the long discussion as to why these genes have been missing from the genome assembly despite their established essential biological function in other species (for examples, see [6–12]).

Mass spectrometry analysis of fat tissue from the same chickens confirmed the identification of *MEPCE*, *NPC1L1*, *PHF1*, *MRPS18*, and *SF3B2* at $P < 0.01$, and the expression of *AMIGO1*, *CYAB*, *FKBP11*, *MGAT1*, *MOGS*, *MRII*, *MTX1*, *POLR3D*, *PEA15*, and *TXNIP* at $P < 0.05$ (Additional file 1: Tables S4, S5, and S8). To further validate the novel genes in the context of species phylogeny, we selected 11 genes with complete coding sequences predicted by TransDecoder (Additional file 3: Table S13) and at least four reported orthologous protein sequences in the NCBI protein database, for analysis of protein identity with the predicted chicken amino acid sequence using pBlast. As expected, the relative degrees of sequence identity were inversely correlated with evolutionary distance for most transcripts ($r = -1$ to -0.7), with three exceptions resulting from high conservation.

Comparing these genes to the genes previously reported as missing [1, 2, 6] recovered 74 overlapping gene symbols (Table 1). A higher proportion of the genes reported missing only in chickens was identified compared to those reported missing in all avian species (15% and 3–4.5%, respectively). The recovered transcripts had very high GC content (68%; Additional file 3: Figure S3b), further supporting the hypothesis that many of the genes that are currently missing from the draft genome eluded previous identification because of their high GC content [3, 4].

Table 1 Characterization of the novel genes reported missing in previous studies

Previously reported list	No. of missing genes	Found in our intermediate set	Found in our high-confidence list	Gene symbols
Predicted absent in birds [1]	274	36 (13%)	8 (3%)	<u>FLT3LG</u> , <u>LPPR2</u> , <u>NPHS1</u> , <u>PLCB3</u> ^a , <u>PRSS8</u> , <u>RCN3</u> , <u>TRMT1</u> , <u>TSPAN31</u>
Predicted missing in chickens but not in all birds [1] ^b	336	152 (45%)	50 (15%)	<u>ALKBH7</u> , <u>ASB16</u> , <u>ATAT1</u> , <u>ATG4D</u> , <u>B9D2</u> , <u>CACNG7</u> , <u>CACNG8</u> , <u>CAMSAP3</u> , <u>CARM1</u> , <u>CCDC106</u> , <u>CCDC120</u> , <u>CCDC22</u> , <u>CIC</u> , <u>CLASRP</u> , <u>CLPP</u> , <u>COPZ1</u> , <u>CYTH2</u> , <u>ESY1</u> , <u>GEMIN7</u> , <u>GPKOW</u> , <u>GTF2F1</u> , <u>JOSD2</u> , <u>KRI1</u> , <u>LMTK3</u> , <u>MAP2K7</u> , <u>METTL1</u> , <u>METTL3</u> , <u>MRPS18B</u> , <u>NDUFB7</u> , <u>PIH1D1</u> , <u>POU6F1</u> , <u>PPP1R12C</u> , <u>PPP1R18</u> , <u>PPPS5</u> , <u>PRKCSH</u> , <u>PRPF31</u> , <u>PRR12</u> , <u>SAMD1</u> , <u>SCAF1</u> , <u>SEMA4C</u> , <u>SLC39A7</u> , <u>SMG9</u> , <u>SSR4</u> , <u>TFPT</u> , <u>TRAPP1</u> , <u>TSR2</u> , <u>U2AF2</u> , <u>UXT</u> , <u>YIF1B</u> , <u>ZNF653</u>
Predicted absent in birds [2]	640	100 (16%)	29 (4.5%)	<u>ADAT3</u> , <u>ALKBH7</u> , <u>C11ORF95</u> , <u>C2ORF68</u> , <u>CCDC22</u> , <u>CDIPT</u> , <u>CGREF1</u> , <u>CIC</u> , <u>CXXC1</u> , <u>FRMD8</u> , <u>HUWE1</u> , <u>IKBKG</u> , <u>KRI1</u> , <u>LMTK3</u> , <u>MBD1</u> , <u>MUS81</u> , <u>NPHS1</u> , <u>OPA3</u> , <u>PHF1</u> , <u>PIH1D1</u> , <u>PLCB3</u> ^a , <u>PPP1R12C</u> , <u>PRKCSH</u> , <u>RCE1</u> , <u>SSSCA1</u> , <u>TFPT</u> , <u>TNF</u> ^d , <u>UXT</u> , <u>ZNF653</u>
Predicted absent by both studies [1, 2]	99	7 (7%)	2 (2%)	<u>NPHS1</u> ^e , <u>PLCB3</u> ^a
Lost adipokines [6]	4	1 (25%)	1 (25%)	<u>TNF</u> ^d

Eleven genes are shared between row 2 (Lovell et al. [1]) and row 3 (Zhang et al. [2]): *ALKBH7*, *CCDC22*, *CIC*, *KRI1*, *LMTK3*, *PIH1D1*, *PPP1R12C*, *PRKCSH*, *TFPT*, *UXT*, and *ZNF653*

^a*PLCB3* was selected manually from the intermediate list of novel genes as a dropout due to misannotation of its quail (*Coturnix japonica*) ortholog (LOC107307599), demonstrating that the intermediate gene list (Additional file 1: Table S4) may contain additional novel genes

^bBased on the genes listed in Tables S4a, S4b, S6a, and S6b in Lovell et al. [1]

^cAlso reported missing in other publications (e.g. [7, 14])

^dAlso reported missing also in Zhang et al. [2] and in additional publications (e.g. [10, 15])

(i) Bold and underlined, (ii) underlined, (iii) underlined by dashed line, and (iv) non-underlined symbols represent (i) novel sequences with no sequence similarity in any genome build, (ii) sequences present in Galgal5 but lacking annotation, (iii) sequences present in Galgal5 as uncharacterized or putative, or (iv) sequences present and annotated in Galgal5, respectively

When exploring the location of novel genes recovered by the updated genome build, we observed that most genes (76%) were located on unplaced scaffolds, probably representing uncharacterized microchromosomes. Among those that mapped to known chromosomes, the majority (80%) were localized to microchromosomes, which are estimated to contain 50% of protein-coding genes in chickens [13]. Surprisingly, many of the mapped genes appeared in clusters. Mapping positions of the human orthologs demonstrated that the organization of 80% of the mapped novel genes was in syntenic clusters (Table 2). The strong tendency of these novel genes to cluster indicated their location in recalcitrant chromosomal regions with high GC content, primarily on microchromosomes. The methods used in this study are detailed in Additional file 4: Detailed materials and methods.

Conclusions

Our RNA-Seq study, combined with extensive bioinformatics analysis, recovered 191 novel genes that were missing from previous chicken assemblies, 38 of which are still not present in the most recent genome build (Galgal5), as well as an additional 47 that are at least partially present in Galgal5 but lacking proper annotation. The high GC content (68% on average), the microchromosomal location of the majority of the novel genes (80%) covered by Galgal5, and their high tendency to cluster into syntenic blocks (80%) suggest that the novel genes were not found in earlier analyses because of their position in GC-rich gene clusters, rather than due to chromosomal fragmentation and loss. In addition, the identification and characterization of *NPHS1* and *TNF*, which are expected to be essential for avian physiology, and which are still missing from the latest genome build,

Table 2 Overview of novel genes missing from the Galgal4 assembly but present in Galgal5

Trinity ID	Predicted gene	Galgal5 mapping			Human ortholog (hg38)	Cluster ^a
		Genes	Chromosome	Coordinates		
c192514_g2_i1	<i>RRS1</i>	<i>RRS1</i>	chr2	115,487,692–115,488,635	chr8:66,429,028–66,430,733	–
c144374_g1_i1	<i>KHK</i>	<i>KHK</i>	chr3	104,952,675–104,954,000	chr2:27,086,747–27,100,751	1
c150768_g1_i3	<i>CGREF1</i>	<i>CGREF1</i>	chr3	104,955,106–104,955,990	chr2:27,100,594–27,119,103	1
c191309_g1_i2	<i>ANKRD66</i>	<i>LOC101750448</i>	chr3	110,320,024–110,320,850	chr6:46,746,917–46,759,506	–
c190219_g1_i1	<i>ADO</i>	<i>ADO</i>	chr6	8,089,943–8,090,591	chr10:62,804,857–62,808,483	–
c165457_g1_i6	<i>ABHD14B</i>	<i>LOC107056876</i>	chr12random_Scaffold5645	10,835–12,580	chr3:51,968,510–51,983,409	–
c181867_g2_i3	<i>RSAD1</i>	<i>MYCBPAP</i>	chr18	10,429,164–10,430,334	chr17:50,508,384–50,531,497	–
c160691_g1_i2	<i>BOLA3</i>	<i>BOLA3</i>	chr22	2,880,009–2,880,858	chr2:74,135,398–74,147,994	2
c178063_g1_i8	<i>SEMA4C</i>	<i>SEMA4C</i>	chr22random_Scaffold1011	444–4,447	chr2:96,859,716–96,869,971	2
c156624_g2_i1	<i>CIART</i>	<i>CIART</i>	chr25	2,384,775–2,385,633	chr1:150,282,543–150,287,093	3
c165802_g2_i1	<i>CRTC2</i>	<i>CRTC2</i>	chr25	2,075,046–2,076,072	chr1:153,947,675–153,958,625	3
c189493_g2_i1	<i>C17orf96</i>	<i>LOC107055293</i>	chr27	4,355,476–4,355,902	chr17:38,671,703–38,675,421	4
c151660_g2_i1	<i>KRI1</i>	<i>LOC107055293</i>	chr27	4,357,140–4,357,428	chr19:10,553,078–10,566,037	4
c167546_g1_i3	<i>FBXW9</i>	<i>FBXW9</i>	chr30random_Scaffold7361	448–2,027	chr19:12,688,917–12,696,643	5
c160528_g1_i2	<i>DHPS</i>	<i>DHPS,WDR83</i>	chr30random_Scaffold7361	2,298–5,407	chr19:12,675,721–12,681,902	5
c150426_g1_i4	<i>YIF1B</i>	<i>YIF1B</i>	chr32random_Scaffold22667	160–217	chr19:38,305,118–38,315,963	6
c167964_g1_i2	<i>B9D2</i>	–	chr32random_Scaffold15198	71–292	chr19:41,354,421–41,364,173	6
c164748_g1_i1	<i>OPA3</i>	<i>OPA3</i>	chr32random_Scaffold826	46,400–48,070	chr19:45,546,281–45,584,819	6
c148689_g1_i2	<i>SNRPD2</i>	<i>SNRPD2</i>	chr32random_Scaffold19601	235–1,401	chr19:45,687,454–45,692,333	6
c163802_g1_i1	<i>GRASP</i>	<i>GRASP</i>	chr33	1,916–6,474	chr12:52,006,940–52,015,864	7
c178972_g2_i2	<i>ESYT1</i>	<i>ESYT1</i>	chr33	679,134–685,279	chr12:56,128,056–56,144,671	7
c171696_g1_i1	<i>APOF</i>	<i>APOF</i>	chr33	776,046–776,629	chr12:56,360,569–56,362,823	7
c100851_g1_i1	<i>HOXC4</i>	<i>HOXC4</i>	chr33	1,095,140–1,096,547	chr12:54,016,931–54,055,327	7
c186414_g2_i1	<i>COPZ1</i>	<i>COPZ1</i>	chr33	1,170,192–1,174,833	chr12:54,325,127–54,351,849	7
c146677_g1_i1	<i>DAZAP2</i>	–	chr33	1,573,156–1,573,299	chr12:51,238,292–51,243,933	7

^aThis column indicates clusters of neighboring genes that are largely supported by the human orthologs

emphasizes the importance of striving towards a repertoire of known and characterized genes that is as complete as possible.

Additional files

Additional file 1: Overview of the RNA-Seq data and filtration of the novel gene candidates. **Table S1.** Information about the RNA-Seq data. **Table S2.** The initial set of 2810 candidate novel transcripts. **Table S3.** Annotation, characterization, and filtering of the novel transcripts. **Table S4.** The intermediate set of 1878 transcripts representing 1063 candidate novel genes. **Table S5.** The high-confidence set of 194 transcripts representing 191 novel genes. **Table S6.** The 191 novel genes not included in Galgal4; 54 of these are correctly annotated while 137 are missing or lack correct annotation in Galgal5. **Table S7.** Characterization of the novel genes according to predicted cellular localization. **Table S8.** Identification of the novel genes in Galgal5 genome assembly and by Mass-Spec analysis in adipose tissue. (XLS 2362 kb)

Additional file 2: Characterization of *NPHS1* and *TNF*. **Figure S1.** Predicted full length cDNA sequence of *NPHS1* and its characterization. **Figure S2.** Predicted full length cDNA sequence of *TNF* and its characterization. **Table S9.** Coding sequence of chicken *NPHS1* and *TNF* predicted transcripts. **Table S10.** List of *NPHS1* and *TNF* exons in human, turtle, and chicken. **Table S11.** List of primers used for RT-PCR. **Table S12.** Probes used for expression profiling in the Sequence Read Archive (SRA) database. (PDF 1672 kb)

Additional file 3: Characterization of the high confidence novel genes. **Table S13.** Phylogenetic analysis of representative novel genes. **Figure S3.** Characterization of the novel transcripts. (PDF 232 kb)

Additional file 4: Detailed materials and methods. Animals and tissue sampling. RNA-seq. Bioinformatic analysis. RT-PCR. Mass spectrometry analysis (MS). (PDF 283 kb)

Abbreviations

NPHS1: Nephhrin; *TNF*: Tumor necrosis factor

Acknowledgment

We thank Mr Mark Ruzal for growing the chickens. The study was supported by the ERC project BATESON (awarded to LA), Israel Academy of Sciences 876/14, and by the Chief Scientist of the Israeli Ministry of Agriculture 0469/14 (awarded to MFE and ES).

Availability of data and materials

The raw sequences that were used to build the trinity transcripts of the novel genes, as well as the cDNA sequences of the chicken *NPHS1* and chicken and turkey *TNF*, are available in the ENA BioProject repository [PRJEB13623, www.ebi.ac.uk/ena/data/view/PRJEB13623].

Authors' contributions

SB performed the transcript assembly and produced the novel gene lists. ES extended and characterized the *NPHS1* and *TNF* predicted cDNAs and proteins. MFE and SY performed the biological experiments, prepared the RNA for sequencing, confirmed the deduced cDNA sequences of *NPHS1* and *TNF* by RT-PCR, and performed the *NPHS1* and *TNF* expression profiling. KP and SCB performed the MS analysis. MG helped to design the bioinformatic approaches. MFE, SB, and LA designed the experiments and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

All animal procedures were carried out in accordance with the National Institutes of Health Guidelines on the Care and Use of Animals and Protocol IL536/14, which was approved by the Animal Experimentation Ethics Committee of the Agricultural Research Organization, Volcani Center, Rishon, Israel.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala SE-751 23, Sweden. ²Present Address: Wellcome Trust Medical Research Council Stem Cell Institute, University of Cambridge, Cambridge CB2 1QR, UK. ³Agricultural Research Organization, Volcani Center, Rishon LeZion, Israel. ⁴College of Agriculture and Life Sciences, University of Arizona, Tucson, AZ 85721-0036, USA. ⁵Bioinformatics Infrastructure for Life Sciences, Uppsala University, Uppsala, Sweden. ⁶Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala SE-750 07, Sweden. ⁷Department of Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77843-4458, USA.

Published online: 14 June 2017

References

- Lovell PV, Wirthlin M, Wilhelm L, Minx P, Lazar NH, Carbone L, et al. Conserved syntenic clusters of protein coding genes are missing in birds. *Genome Biol.* 2014;15:565.
- Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science.* 2014;346:1311–20.
- Seroussi E, Cinnamon Y, Yosefi S, Genin O, Smith JG, Rafati N, et al. Identification of the long-sought leptin in chicken and duck: expression pattern of the highly GC-rich avian leptin fits an autocrine/paracrine rather than endocrine function. *Endocrinology.* 2016;157:737–51.
- Hron T, Pajer P, Paces J, Bartunek P, Elleder D. Hidden genes in birds. *Genome Biol.* 2015;16:164.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 2011;29:644–52.
- Dakovic N, Terezol M, Pitel F, Maillard V, Elis S, Leroux S, et al. The loss of adipokine genes in the chicken genome and implications for insulin metabolism. *Mol Biol Evol.* 2014;31:2637–46.
- Miner JH. Life without nephrin: it's for the birds. *J Am Soc Nephrol.* 2012;23:369–71.
- Wajant H, Pfizenmaier K, Scheurich P. Tumor necrosis factor signaling. *Cell Death Differ.* 2003;10:45–65.
- Wagner N, Morrison H, Pagnotta S, Michiels JF, Schwab Y, Tryggvason K, et al. The podocyte protein nephrin is required for cardiac vessel formation. *Hum Mol Genet.* 2011;20:2182–94.
- Uysal B, Donmez O, Uysal F, Akaci O, Vuruskan BA, Berdeli A. Congenital nephrotic syndrome of *NPHS1* associated with cardiac malformation. *Pediatr Int.* 2015;57:177–9.
- Li X, Chuang PY, D'Agati VD, Dai Y, Yacoub R, Fu J, et al. Nephrin preserves podocyte viability and glomerular structure and function in adult kidneys. *J Am Soc Nephrol.* 2015;26:2361–77.
- Kestila M, Lenkkeri U, Mannikko M, Lamerdin J, McCready P, Putaala H, et al. Positionally cloned gene for a novel glomerular protein—nephrin—is mutated in congenital nephrotic syndrome. *Mol Cell.* 1998;1:575–82.
- Smith J, Bruley CK, Paton IR, Dunn I, Jones CT, Windsor D, et al. Differences in gene density on chicken macrochromosomes and microchromosomes. *Anim Genet.* 2000;31:96–103.
- Yaoita E, Nishimura H, Nameta M, Yoshida Y, Takimoto H, Fujinaka H, et al. Avian podocytes, which lack nephrin, use adherens junction proteins at intercellular junctions. *J Histochem Cytochem.* 2016;64:67–76.
- Magor KE, Miranzo Navarro D, Barber MR, Petkau K, Fleming-Canepa X, Blyth GA, Blaine AH. Defense genes missing from the flight division. *Dev Comp Immunol.* 2013;41:377–88.