Toolbox

Physiol Genomics 25: 514–524, 2006.
First published March 22, 2006; doi:10.1152/physiolgenomics.00207.2005.

Chicken genomics resource: sequencing and annotation of 35,407 ESTs from single and multiple tissue cDNA libraries and CAP3 assembly of a chicken gene index

Wilfrid Carre,[1] Xiaofei Wang,[1] Tom E. Porter,[2] Yves Nys,[3] Jianshan Tang,[1,4] Erin Bernberg,[1,4] Robin Morgan,[1,4] Joan Burnside,[1,4] Samuel E. Aggrey,[5] Jean Simon,[3] and Larry A. Cogburn[1]

[1]*Department of Animal and Food Sciences and* [4]*Delaware Biotechnology Institute, University of Delaware, Newark, Delaware;* [2]*Department of Animal and Avian Sciences, University of Maryland, College Park, Maryland;* [3]*Station de Recherches Avicoles, Institut National de la Recherche Agronomique, Nouzilly, France; and* [5]*Department of Poultry Science, University of Georgia, Athens, Georgia*

Submitted 12 August 2005; accepted in final form 8 February 2006

Carre, Wilfrid, Xiaofei Wang, Tom E. Porter, Yves Nys, Jianshan Tang, Erin Bernberg, Robin Morgan, Joan Burnside, Samuel E. Aggrey, Jean Simon, and Larry A. Cogburn. Chicken genomics resource: sequencing and annotation of 35,407 ESTs from single and multiple tissue cDNA libraries and CAP3 assembly of a chicken gene index. *Physiol Genomics* 25: 514–524, 2006. First published March 22, 2006; doi:10.1152/physiolgenomics.00207.2005.—Its accessibility, unique evolutionary position, and recently assembled genome sequence have advanced the chicken to the forefront of comparative genomics and developmental biology research as a model organism. Several chicken expressed sequence tag (EST) projects have placed the chicken in 10th place for accrued ESTs among all organisms in GenBank. We have completed the single-pass 5′-end sequencing of 37,557 chicken cDNA clones from several single and multiple tissue cDNA libraries and have entered 35,407 EST sequences into GenBank. Our chicken EST sequences and those found in public databases (on July 1, 2004) provided a total of 517,727 public chicken ESTs and mRNAs. These sequences were used in the CAP3 assembly of a chicken gene index composed of 40,850 contigs and 79,192 unassembled singlets. The CAP3 contigs show a 96.7% match to the chicken genome sequence. The University of Delaware (UD) EST collection (43,928 clones) was assembled into 19,237 nonredundant sequences (13,495 contigs and 5,742 unassembled singlets). The UD collection contains 6,223 unique sequences that are not found in other public EST collections but show a 76% match to the chicken genome sequence. Our chicken contig and singlet sequences were annotated according to the highest BlastX and/or BlastN hits. The UD CAP3 contig assemblies and singlets are searchable by nucleotide sequence or key word (**http://cogburn.dbi.udel.edu**), and the cDNA clones are readily available for distribution from the chick EST website and clone repository (**http://www.chickest.udel.edu**). The present paper describes the construction and normalization of single and multiple tissue chicken cDNA libraries, high-throughput EST sequencing from these libraries, the CAP3 assembly of a chicken gene index from all public ESTs, and the identification of several nonredundant chicken gene sets for production of custom DNA microarrays.

chicken cDNA libraries; high-throughput DNA sequencing; expressed sequence tags; expressed sequence tag sequence assembly; nonredundant gene sets; *Gallus gallus*

SINCE ITS EARLY DOMESTICATION from the red jungle fowl (*Gallus gallus*) in southeast Asia about 8,000 years ago (35), the domestic chicken (*Gallus domesticus*) has played a key role in the advancement of human culture (13). Today, the domestic chicken continues to serve humans as an important source of high-quality protein from meat and eggs and as a widely used biological model (42). Furthermore, the avian lineage holds a unique position in chordate evolution, and the chicken has recently gained considerable interest from biologists as a model organism for large-scale genomic exploration (5). The recent release and publication (25) of a draft chicken genome sequence, the first for a livestock species, has elevated the chicken to premier status as a model organism for developmental biology and genomics research.

A critical step for assembly and annotation of the chicken genome sequence was the acquisition of an extensive catalog of expressed sequence tags (ESTs) (8). This feat was accomplished by completion of several international chicken EST sequencing projects in a relatively short (<5 yr) period (23). Despite its global agricultural importance, the lack of ESTs and a completed genome sequence once hindered genomics research in the chicken. At the inception of our functional genomics project in 2000, only several thousand chicken ESTs, derived mainly from thymic (46) and bursal (1) lymphocytes, had been determined for the chicken. The first chicken EST database and cDNA clone repository (**http://www.chickest.udel.edu/**) was established (by J. Burnside and R. Morgan) at the University of Delaware (UD) in 2000 with the deposition and annotation of 5,251 chicken ESTs derived from an activated chicken T cell cDNA library (46). The first objective of a second functional genomics project (L. A. Cogburn, T. E. Porter, S. E. Aggrey, and J. Simon) was limited EST sequencing of about 30,000 clones from normalized cDNA libraries for development of tissue-specific chicken microarrays (11). Our chicken cDNA libraries were constructed from metabolic, somatic, neuroendocrine, reproductive, and mixed lymphoid tissues derived mainly from broiler (meat type) chickens. In the midst of our EST sequencing effort, a consortium funded by the British Biotechnology and Biological Sciences Research Council (BBSRC) released a larger and more comprehensive collection of chicken ESTs (3). A total of 332,920 ESTs were derived from 21 normalized libraries representing a wide range of embryonic stages and brain tissues from Leghorn (egg type) chickens and other adult tissues from a mixture of broiler and layer breeds (**http://www.chick.umist.ac.uk/**). The UD chicken

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: L. A. Cogburn, 531 South College Ave., Dept. of Animal and Food Sciences, Univ. of Delaware, Newark, DE 19717 (e-mail: cogburn@udel.edu).

1094-8341/06 $8.00 Copyright © 2006 the American Physiological Society

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.5 on April 8, 2017

cDNA libraries represent several tissues that are either absent or not well represented in other public chicken EST collections. These unique chicken cDNA libraries were derived from the spleen/bursa/thymus/bone marrow/peripheral blood lymphocytes, pituitary/hypothalamus/pineal, abdominal fat, and oviduct mainly from broiler (meat type) chickens.

In this paper, we describe the construction and normalization of single and multiple tissue cDNA libraries, sequencing of 37,577 chicken cDNA clones from these libraries, and the CAP3 assembly of a chicken gene index from all publicly available chicken ESTs. In addition, several nonredundant EST clone sets were clustered from the UD collection for the production of custom chicken cDNA microarrays (10, 11, 28, 31).

MATERIALS AND METHODS

Isolation and purification of total RNA. Total cellular RNA was isolated from each tissue by the guanidine thiocyanate lysis method (9) and CsCl gradient purification (17). Equal amounts of total RNA from individual animals were pooled together and used to prepare poly(A) mRNA (Qiagen; Oligotex mRNA kit, catalog no. 70042), which was oligo(dT) primed for cDNA library synthesis, as described below. All procedures involving the use of animals were approved by the UD Animal Care and Use Committee.

Pilot libraries. The UD chicken EST sequencing project was initiated with the construction of several "pilot" cDNA libraries (*ptr1c*, *pti1c*, *pnf-b*, *pnl-b*, and *pco1c*). These pilot libraries were made by OriGene Technologies (Rockville, MD) with the exception of the pilot oviduct library (*pco1c*), which was made by Life Technologies (Rockville, MD). Two of the pilot libraries (*pnl1s* and *pco1s*) were subtracted with highly redundant clones sequenced from each primary pilot library (*pnl-b* or *pco1c*, respectively) using protocols described by Bonaldo et al. (4). The details of construction and subtraction of pilot chicken cDNA libraries are presented as Supplemental Materials (available at the *Physiological Genomics* web site).[1]

Lymphoid tissue RNA pools. RNA was prepared from bursa, thymus, spleen, and bone marrow of four individual broiler chickens at *day 18* (embryo) and 1, 3, 5, and 7 wk of age and combined with RNA prepared from a pool of peripheral blood lymphocytes of 4-wk-old Leghorn birds. Each tissue comprised ~20% of the final RNA pool for the mixed lymphoid tissue library (*pgn1c*). [Note that the details of the composition of RNA pools used to construct each primary library are presented in the Supplemental Text and Supplemental Fig. S1.]

Liver RNA pools. The liver cDNA library (*pgl1c*) was constructed using liver RNA isolated from broiler (meat type) chickens pooled across different developmental stages and different genetic backgrounds. Equal amounts of RNA from four birds/genetic lines [strains 80 (or 90) and 21 from the Centre for Food and Animal Research (CFAR), Agriculture and Agri-Food Canada, Ottawa, ON, Canada] (29) at four ages (1, 3, 7, and 11 wk) were combined with RNA isolated from the liver of *day 17* embryos from a commercial broiler strain (Ross × Arbor Acres). The final composition of the chicken liver cDNA library is represented by *day 17* embryos (20%) and six ages of CFAR strains 80 (40%) and 21 (40%).

Abdominal fat RNA pools. A chicken abdominal fat cDNA library (*pft1c*) was constructed from a pool of RNA isolated from the same birds used in preparation of the primary liver library (*pgl1c*). Equal amounts of total RNA from four birds/genetic lines were pooled across four ages. Sequencing from the first normalized fat library (*pgf1n*) revealed a slight contamination (1.94%) with *Escherichia coli* phage protein. A second abdominal fat library (*pgf2c*) was constructed from a mixture of RNA isolated from adult single-comb White Leghorn (SCWL; egg type) chickens, CFAR broiler strains 80 and 21 (7 and 9 wk), and commercial broiler (Ross × Arbor Acres) chickens (*day 19* embryos and 1-day- and 3-wk-old chicks). The final composition of the second fat library (*pgf2c*) represents abdominal fat RNA from three developmental stages (late embryo, juvenile, and adult) of broiler (79%) and egg-type (21%) chickens.

Skeletal muscle and epiphyseal growth plate RNA pools. The skeletal muscle RNA pool was made from equal amounts of breast (white fiber) and leg (red fiber) muscle RNA from CFAR strains 90 and 21 at six ages (1, 3, 5, 7, 9, and 11 wk) and commercial broiler chickens (*day 17* and *day 18* embryos and 1-day-old chicks). The skeletal muscle RNA pool was combined with RNA isolated from the epiphyseal growth plate (36) of commercial (Cobb) broiler chickens (1-, 7-, and 14-day-old chicks) (kindly provided by E. Monsonego-Ornan, Volcani Center, Agricultural Research Organization, Bet-Dagan, Israel). Thus the skeletal muscle/epiphyseal growth plate cDNA library (*pgm1c*) was constructed from an RNA pool made from one-third portions of embryonic and posthatching breast and leg muscle RNA and epiphyseal growth plate RNA from juvenile broiler chickens.

Neuroendocrine tissue RNA pools. The hypothalamus and pituitary gland were collected from commercial broiler chickens (Avian × Avian strain) during late embryonic (*days 12*, *14*, and *19*) and early juvenile development (1, 3, 5, 7, and 9 wk). For embryos, the pituitary glands and hypothalami were pooled together at each age because of the small size of these tissues. The pituitary glands and hypothalami from posthatching chickens were collected and processed separately. Pineal glands were collected from posthatching chickens and processed as a single pool for each age. The final total RNA pool for the neuroendocrine system cDNA library (*pgp1c*) was composed of 40% pituitary, 40% hypothalamic, and 20% pineal total RNA.

Reproductive tract RNA pools. The chicken reproductive tract cDNA library (*pgr1c*) was constructed from RNA isolated from oviduct, ovary, and testes of both broiler (Ross × Arbor Acres) and Leghorn chickens at various stages of sexual development. Testes RNA from 5-, 7-, 13-, 21-, and 35-wk-old broiler males and a year-old Leghorn rooster were pooled together. RNA isolated from immature ovaries of 5-, 7-, and 8-wk-old broiler females and ovaries of 1-yr-old Leghorn laying hens were pooled (the yellow and large white follicles were removed). RNA was also isolated from the magnum, white isthmus, and uterus of commercial laying hens (ISA-Brown) at 3 and 16 h after oviposition to obtain oviduct tissue during different stages of transit of the developing egg. Thus the chicken reproductive tract cDNA library was constructed from an RNA pool composed of 50% oviduct, 25% ovary, and 25% testes RNA.

Construction and normalization of cDNA libraries. The construction and normalization of our chicken cDNA libraries were performed as a custom service (catalog no. 11315-017) by a commercial company [Life Technologies (LTI), Rockville, MD; now Invitrogen, Carlsbad, CA]. The primary libraries were constructed and directionally cloned using SuperScript II H^- RNase RT, ElectroMax DH10B cells, and pCMV Sport 6.0 vector, with the exception of *pft1c*, which was cloned into pSPORT1. The primary libraries contained at least 3×10^6 primary clones. The average insert size (Table 2) was initially estimated by PCR amplification of 23 randomly picked clones/library by the vendor, LTI (Invitrogen).

The libraries were amplified by a semisolid agar procedure to minimize clone size bias and then normalized by LTI's proprietary Subtraction Technology, which is largely based on the procedures described by Soares et al. (40) and Bonaldo et al. (4). The protocols used by LTI (and subsequently Invitrogen) in the construction and normalization of single and multiple tissue cDNA libraries from livestock species have been described in detail elsewhere (15, 39, 41). Our multitissue chicken cDNA libraries were designed to yield the

[1] The Supplemental Material for this article (Supplemental Text, Supplemental Table S1, and Supplemental Figs. S1–S3) is available online at **http://physiolgenomics.physiology.org/cgi/content/full/00207.2005/DC1**.

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.5 on April 8, 2017

maximum number of nonredundant ESTs for development of custom microarrays.

High-throughput DNA sequencing. DNA sequencing was performed at Dupont's high-throughput sequencing facility (Agricultural Products Division, EI du Pont de Nemours, Delaware Technology Park, Newark, DE). Big Dye terminator cycle sequencing reactions (20 μl) were performed using vector primers and a one-fourth dilution of Big Dye (v3.0) on ABI 3700 sequencers (Applied Biosystems). Sequence was obtained from the 5′-end to improve the likelihood of obtaining coding sequence and, therefore, the identity of the cDNA. A quality score of 20 (q20), generated by the Phred basecaller (14), was used as the cutoff parameter for sequence data. After trimming of vector sequence and ambiguous bases at the beginnings and ends of reads, the sequences were stored in a Sybase database at Dupont and continuously clustered and compared within and among all chicken cDNA libraries. For clustering of ESTs within a library, distinct cDNA sequences were identified by basic local alignment search tool (Blast)N analysis with a minimum score of 750 (where the matrix was +5/−4, gap open and extended by −10), a minimum sequence overlap of 75 bp, and a minimum sequence identity of 80%. The number of distinct cDNA sequences represents the sum of unique singlets (only 1 EST) plus the number of unique clusters containing multiple overlapping ESTs (Table 2).

One to three 384-well plates of randomly picked clones were sequenced from each primary cDNA library to evaluate library normalization. EST sequences were annotated with the highest BlastX and BlastN scores and electronically transferred to the UD investigators for batch submission to GenBank and integration into the chicken EST database. The sources of contaminating sequence, expressed as a percentage of all ESTs sequenced, were as follows: bacterial phage head protein (0.18%) introduced during normalization of one library (*pgf1n*), bacteria (*E. coli*) (0.08%), cloning vector (pCMV Sport 6.0) (0.11%), mitochondrial RNA (2.75%), and ribosomal RNA (0.11%).

Bioinformatics. The UD chicken EST database (**http://www.chickest.udel.edu**) was developed by INCOGEN (Williamsburg, VA) under a United States Department of Agriculture (USDA)-National Research Initiative (NRI) grant (to J. Burnside and R. Morgan). The top five Blast results for each EST are stored in the database, which is searchable by key word or clone identification (ID) with a web-based browser that also provides BlastN queries of sequences.

All chicken ESTs found in public databases on July 1, 2004, were assembled into contigs to improve clone annotation and to identify nonredundant clone sets for development of custom chicken microarrays. A total of 517,727 chicken sequences (492,786 ESTs and 24,941 mRNAs) were trimmed of the poly(A) tail, vector, phage, and/or bacterial contaminants using Phil Green's Cross_match program (Washington University, St. Louis, MO, and **http://www.phrap.com**). First, a BlastN analysis was used to group ESTs with overlapping sequence into seven cluster bins. These seven cluster bins were then used to build contigs with the CAP3 sequence assembly program (22), using parameter settings of 90% sequence identity and 40 bp minimum overlap with a maximum overhang length of 50 bp (recommended by X. Huang, Michigan Technological University, Houghton, MI). Furthermore, these contig and singlet sequences were then used in a local BlastN search against the first draft of the chicken genome sequence (**ftp://ftp.ncbi.nih.gov/genomes/Gallus_gallus/**). The parameters used for the BlastN search against the chicken genome were as follows: *E*-value $<10^{-20}$, >95% identity, and >75% coverage of the contig sequence.

RESULTS

Sequencing and analysis of chicken ESTs. We have increased the size of the original UD chicken EST database to 43,928 by sequencing an additional 37,557 clones from several pilot, primary, or normalized chicken cDNA libraries. These chicken EST clones were sequenced from 22 single tissue or multitissue cDNA libraries (Table 1); 246 sequences were removed from this list because they represented either short reads (<100 bp) or contaminating sequences (vector, bacterial, yeast, or viral sequences). Only limited sequencing was completed for the primary and subtracted pilot libraries because of high redundancy of clones sequenced from these unnormalized cDNA libraries. A total of 35,407 chicken EST sequences were batch submitted to the dbEST division of GenBank under the following accession numbers: AW198238–AW198611, AW239568–AW240205, AW355197–AW355737, BG624962–BG625773, BG641470–BG642131, BG709680–BG713744, BI064338–BI067979, BI389587–BI395077, BM425493–BM425861, BM425865–BM427665, BM439357–BM440804, BM485239–BM492051, BQ037060–BQ039025, CA786126–CA786443, CB016320–CB018492, CB270838–CB271074, CD214214–CD216709, and CD217545–CD219105. Forty-eight percent of EST sequences obtained from the pilot and subtracted libraries were not submitted to GenBank because of high redundancy (i.e., same length and 100% sequence identity). However, all UD chicken EST sequences were entered into our EST database and used in the CAP3 assembly of a chicken gene index (**http://cogburn.dbi.udel.edu/**). A searchable database and repository of chicken EST clones in the UD collection are available online at **http://www.chickest.udel.edu**.

The relative abundance of highly expressed genes sequenced from one to three 384-well plates of randomly picked clones in the primary libraries was used to generate an "electronic Northern blot" of representative tissues (Supplemental Fig. S1 and Supplemental Text). The success of subsequent normalization was evaluated by comparing the number of clones that represent the most abundant genes sequenced from the primary library with the number of redundant clones for that gene sequenced from the normalized library (Supplemental Table S1). Two libraries (*pgm1n* and *pgp1n*) completely failed normalization, since the six most highly expressed genes showed a redundancy similar to that of the primary libraries (*pgm1c* and *pgp1c*). Therefore, the two original primary libraries (*pgm1c* and *pgp1c*) were resubmitted to Invitrogen for normalization and subsequently renamed (*pgp2n* and *pgm2n*). The abundance of α_1-actin was reduced 319-fold in the normalized library *pgm2n*. Similarly, the abundance of proopiomelanocortin (POMC) was 97-fold lower in the normalized neuroendocrine library (*pgp2n*) compared with the primary library (*pgp1c*). The redundancy of the most abundant clones found in each of the primary libraries was dramatically reduced in each respective normalized library (Supplemental Table S1).

A total of 14,346 ESTs were sequenced from the primary (unnormalized) chicken cDNA libraries. The average insert size of clones in the primary libraries (Table 2) ranged from 1.6 kb (*pgm1n)* to 2.2 kb *(pgl1c* and *pgr1c*). The average length of EST reads from the primary libraries was 584 bp, and the percent distinct clones ranged from 53.1% (*pgm1c*) to 85.4% (*pgn1c*). A total of 20,091 ESTs were sequenced from six normalized libraries. The average sequencing read from the normalized libraries was 586 bp, and the number of distinct clones sequenced from each library ranged from 74.2% (*pgl1n*) to 91.2% *(pgr1n*). The average insert size of clones from the normalized libraries, based on insert size from PCR amplification of four 96-well plates of nonredundant clones/library, ranged between 1.59 kb (*pgm2n*) and 1.89 kb (*pgl1n*).

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.5 on April 8, 2017

Table 1. *No. of cDNA clones sequenced from single and multiple tissue chicken cDNA libraries*

Name	Type	Tissue Source	No. of ESTs
pgn1c	Primary	Thymus, bursa, spleen, PBL and bone marrow (1–7 wk)	5,642
ptr1c	Pilot	Uninfected (reference) splenic T-cells	655
pti1c	Pilot	MDV-infected splenic T cells	651
		Total lymphoid	6,948*
pnf-b	Pilot	Abdominal fat (3–5 wk)	390
pft1c	Primary	Abdominal fat (1–11 wk CFAR broilers)	358
pgf1n	**Normalized**	**Abdominal fat (1–11 wk CFAR broilers)**	**3,811**
pgf2c	Primary	Abdominal fat (embryo, juvenile, and adult stages)	365
pgf2n	**Normalized**	**Abdominal fat (embryo, juvenile, and adult stages)**	**1,815**
		Total adipose	6,739*
pnl-b	Pilot	Liver (5 wk old)	359
pnl1s	Subtracted	Liver (5 wk old)	260
pgl1c	Primary	Liver (1–11 wk CFAR broilers)	678
pgl1n	**Normalized**	**Liver (1–11 wk CFAR broilers)**	**4,244**
		Total liver	5,541*
pgm1c	Primary	Breast/leg muscle (1–11 wk CFAR broilers) and bone growth plate (1, 7, and 14 Da commercial broilers)	1,013
pgm1n	Unnormalized	Breast/leg muscle (1–11 wk CFAR broilers) and bone growth plate (1, 7, and 14 Da broilers)	330
pgm2n	**Normalized**	**Breast/leg muscle (1–11 wk CFAR broilers) and bone growth plate (1, 7, and 14 Da broilers)**	**4,421**
		Total breast/leg muscle/growth plate	5,764*
pgp1c	Primary	Pituitary gland/hypothalamus/pineal gland (embryo and juvenile commercial broilers)	909
pgp1n	Unnormalized	Pituitary gland/hypothalamus/pineal gland (embryo and juvenile commercial broilers)	4,677
pgp2n	**Normalized**	**Pituitary gland/hypothalamus/pineal gland (embryo and juvenile commercial broilers)**	**3,151**
		Total neuroendocrine	8,737*
pco1c	Pilot	Oviduct (adult SCWL)	375
pco1s	Subtracted	Oviduct (adult SCWL)	84
pgr1c	Primary	Oviduct/ovary/testis (embryo, juvenile, and adult stages)	366
pgr1n	**Normalized**	**Oviduct/ovary/testis (embryo, juvenile, and adult stages)**	**2,649**
		Total reproductive tract	3,828*
		Total ESTs	37,557*

Normalized libraries are shown in boldface. Two "normalized" libraries (*pgp1n* and *pgm1n*) were submitted for sequencing and found to be poorly normalized. The 2 original primary libraries (*pgp1c* and *pgm1c*) were then normalized by Invitrogen and renamed (*pgp2n* and *pgm2n*). Construction and subtraction of pilot cDNA libraries (*ptr1c, pti1c, pnf-b, pnl-b, pnl1s, pco1c,* and *pco1s*) are described in the Supplemental Materials. PBL, peripheral blood lymphocytes; MDV, Marek's disease virus; CFAR, Centre for Food and Animal Research, Agriculture and Agri-Food Canada; EST, expressed sequence tag; SCWL, single-comb White Leghorn. *Totals.

Table 2. *Sequencing statistics for primary and normalized chicken cDNA libraries*

Library	Total Lanes	Good Lanes	Average Read, bp	Distinct cDNA Sequences, %	Insert Size, kb
pgn1c	5,843	5,642	554	82.2	1.97
pft1c	362	358	435	83.5	1.70
pgf1n	**3,859**	**3,811**	**567**	**87.0**	**1.61**
pgf2c	367	365	626	82.5	1.90
pgf2n	**1,834**	**1,815**	**594**	**88.9**	**1.58**
pgl1c	678	678	593	62.1	2.20
pgl1n	**4,336**	**4,244**	**545**	**74.2**	**1.89**
pgm1c	1,024	1,013	622	53.1	1.69
*pgm1n**	336	335	618	70.9	1.60
pgm2n	**4,500**	**4,421**	**591**	**77.8**	**1.59**
pgp1c	920	909	607	70.0	1.87
*pgp1n**	4,738	4,680	609	55.8	2.00
pgp2n	**3,194**	**3,151**	**604**	**85.8**	**1.77**
pgr1c	367	366	604	79.2	2.20
pgr1n	**2,686**	**2,649**	**615**	**91.2**	**1.60**

Insert size was estimated from an analysis of PCR products amplified in four 96-well plates of nonredundant ESTs used in printing of DNA microarrays. No. of distinct cDNA sequences was determined within each library by basic local alignment search tool (Blast) N analysis of all EST sequences within that library. Percent distinct cDNA sequences (within a library) represent the total no. of nonredundant sequences (contigs + singlets) divided by the no. of good sequence lanes. Boldface indicates normalized libraries. *Two libraries (*pgm1n* and *pgp1n*) completely failed normalization and are considered as unnormalized libraries.

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.5 on April 8, 2017

The distribution of EST clusters within each normalized library is presented as the number of clusters containing one EST (a singlet) or two (or more) ESTs per cluster (Supplemental Fig. S2). Most of the ESTs sequenced from the six normalized libraries are distributed as singlets (an average of 2,165 singlets/library) or clusters containing two ESTs (an average of 164 clusters/library), three ESTs (53 clusters/library), four ESTs (18 clusters/library), or more than five ESTs (10 clusters/library). The majority (83%) of the contigs contain three or fewer ESTs, which indicates that the redundancy within the normalized libraries is relatively low.

Sequence alignment and cluster analyses. To improve annotation of our EST clones, we assembled a gene index from all chicken ESTs and mRNAs found in public databases on July 1, 2004 (Table 3). The CAP3 sequence cluster program (22) was used at the recommended stringency (40 bp overlap with 90% sequence identity). Considering only the 43,928 ESTs found in the UD collection, 38,186 ESTs were clustered into 13,495 contigs (in silico cDNAs), while an additional 5,742 ESTs were classified as singlets, which represent the nonoverlapping sequences. Thus the UD collection represents 19,237 nonredundant EST sequences (contigs + singlets). There are 6,223 unique sequences that are only found in the UD collection (i.e., UD specific), where 76% of these UD-specific sequences match the draft chicken genome sequence. The UD-specific sequences represent 481 contigs and 5,742 singlets. (Within the 5,742 UD-specific singlets, 85% of the high-scoring ESTs matched the genome sequence, while 67% of the low Blast score ESTs and 75% of unknown ESTs matched the chicken genome sequence.) The CAP3 assembly of 492,786 chicken ESTs and 24,941 mRNAs found in public databases (as of July 1, 2004) shows that 438,535 sequences (414,980 ESTs + 23,555 mRNAs) form 40,850 contigs, while 79,192 sequences (77,806 ESTs + 1,386 mRNAs) represent nonoverlapping singlets. The present CAP3 assembly of a chicken gene index (Table 3) closely corresponds to The Institute for Genomic Research (TIGR) *Gallus gallus* Gene Index (*Gg*GI; release 8.0) (**http://www.tigr.org/tdb/tgi/gggi/**), which shows 493,547 chicken ESTs and 23,057 expressed transcripts (ETs or mRNAs) assembled into 116,777 nonredundant sequences from 42,988 contigs (tentative consensus sequences; TCs), 72,941 singlets, and 848 mature transcripts (ETs).

As expected, CAP3 clustering of the 517,727 chicken ESTs found in public databases and Blast analysis of the resulting contigs improved gene identification in the UD EST collection. CAP3 clustering of UD chicken ESTs alone showed that 58% of the UD ESTs were classified with a high BlastX score (>200), while 24% had a low BlastX score (50–200) and 18% remained unknown (<50). The number of UD ESTs with a high BlastX score (>200) was 78% after CAP3 assembly of all public chicken ESTs, while the number of UD ESTs either with a low BlastX score (13%) or classified as unknown (9%) was reduced by 50%. A subset of 20,680 UD CAP3 contigs was used in a BlastX analysis against the nonredundant human protein sequences in GenBank. Contigs with no Blast hit (E-value $<10^{-5}$) or partial alignment were removed from the analysis. The average amino acid identity of putative chicken proteins, derived from high-fidelity UD CAP3 contigs, with a Blast hit to the human protein database was 71% (median = 73%). This number likely reflects the average amino acid identity between human proteins and their chicken homologs.

The UD CAP3 database of contigs and unassembled singlets is searchable by Blast or key word queries under the Gene Index button (**http://cogburn.dbi.udel.edu/**). For example, a key word query for CCAAT/enhancer-binding protein-β (*C/EBPβ*) or a BlastN search with its cDNA sequence against our CAP3 database generates a web page (Fig. 1) that displays the ESTs used to assemble UD CAP3 Contig_23098.4. A BlastN search of the UD CAP3 Contig_23098.4 sequence against the TIGR *Gg*GI shows 99% nucleotide identity to *Gg*GI TC158932 (*C/EBPβ*).

A BlastN search of the 40,850 UD CAP3 contigs against the draft chicken genome sequence showed a 96.7% match and wide chromosomal distribution (Fig. 2*A*). Ten percent of the contigs (4,350) showed a significant BlastN hit to a genomic sequence that is yet unassigned to a specific chicken chromosome (unknown *GGA*), while 1,345 contigs (3.3%) showed no match to the draft chicken genome sequence. The chromosomal distribution (contigs/chromosome) of the high-fidelity CAP3 contigs shows that the largest gene assignment was to *GGA*1, which is the largest chicken chromosome (Fig. 2*A*). Very few of the CAP3 contigs mapped to *GGA*16 and *GGA*W (female chromosome), which are among the most poorly assembled chicken chromosomes (25). In contrast, gene density (contig/Mb), based on the size (Mb) of each chromosome sequenced (25), shows that chicken microchromosomes exhibit a higher number of genes (contigs) per megabase of chromosome that was actually sequenced (Fig. 2*B*). Three-dimensional plots of UD contigs (Supplemental Fig. S3*A*) or UD singlets (Supplemental Fig. S3*B*) across chromosomes and tissue source of ESTs show a higher distribution of UD EST se-

Table 3. *Cluster analysis of chicken EST sequences and assembly of a chicken gene index*

	*UD ESTs	Unique to UD Collection	*Public ESTs (Version 4)	†TIGR *Gg*GI (release 8.0)
Total ESTs (mRNA or ET)	**43,928**		492,786 (24,941)	493,547 (23,057)
ESTs in contigs	**38,186**	**1,503**	438,535	420,606
Contigs (TCs) (>1 UD EST) (only 1 UD EST)	**13,495 (6,582) (6,913)**	**481**	40,850	42,988 TCs
Singlets	**5,742**	**5,742**	79,192	72,941
Total nonredundant sequences	**19,237**	**6,223**	120,042	116,777

*The CAP3 sequence assembly program was used to assemble the Univ. of Delaware (UD) ESTs with all chicken ESTs found in public databases on July 1, 2004 (Version 4; **http://cogburn.dbi.udel.edu/**). This UD CAP3 assembly was made with 492,786 chicken ESTs found in public databases, including 24,941 chicken mRNAs in GenBank and 43,928 ESTs in the UD collection. †The Institute for Genomic Research (TIGR) *Gallus gallus* Gene Index (*Gg*GI), Release 8.0, was also assembled with the CAP3 program (**http://www.tigr.org/tdb/tgi/gggi/**). ET, expressed transcript; TC, tentative consensus sequence (TIGR definitions).

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.5 on April 8, 2017

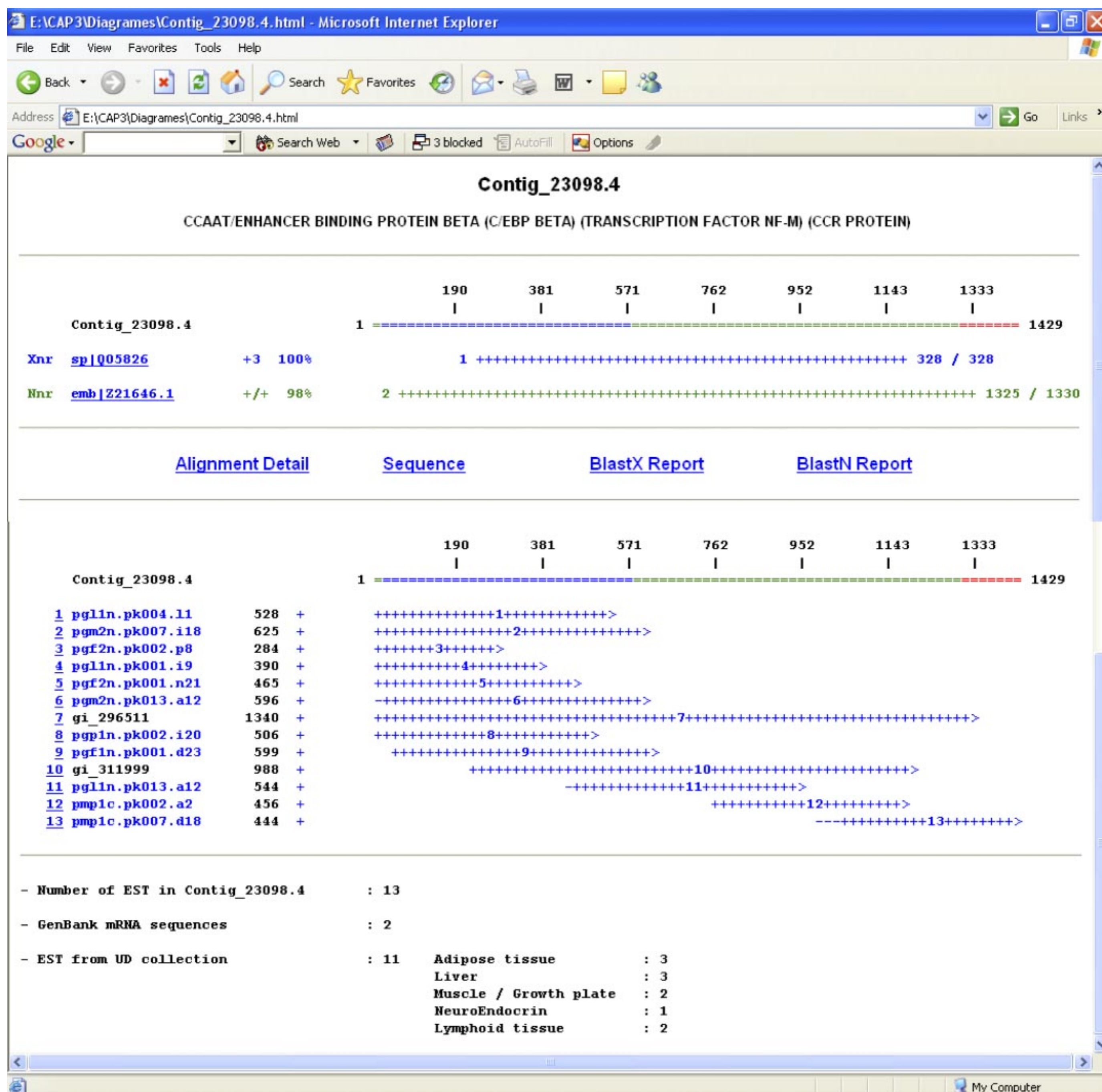

Fig. 1. Web page generated from a key word query (under the Gene Index button) for "CCAAT/enhancer-binding protein-β (or *C/EBPβ*)" or a Blast search of the *C/EBPβ* cDNA sequence against our CAP3 database (**http://cogburn.dbi.udel.edu/**).

quences within the macrochromosomes (*GGA1–GGA5*). The ESTs derived from lymphoid tissue cDNA libraries yielded a larger proportion of unassembled singlets. A BlastN search of the 5,742 UD-specific singlets against the chicken genome shows that 74.4% of these unique ESTs match the genome sequence.

For microarray applications, 19,237 nonredundant UD sequences (cDNA clones) were clustered into three major physiological systems (Fig. 3). These unique gene sets represent the metabolic/somatic system (5,603 unique ESTs), the neuroendocrine/reproductive system (3,786 unique ESTs), and the immune system (5,270 unique ESTs). The metabolic/somatic gene set (3,788 unique contigs and 1,815 unique singlets) represents ESTs sequenced from liver, abdominal fat, and skeletal muscle/bone growth plate cDNA libraries (Table 1). The neuroendocrine/reproductive gene set (3,786 unique contigs and 1,190 unique singlets) represents ESTs derived from the pituitary, hypothalamus, pineal, oviduct, ovary, and testis. The immune system gene set contains an equal number of contigs and singlets from two unnormalized libraries con-

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.5 on April 8, 2017

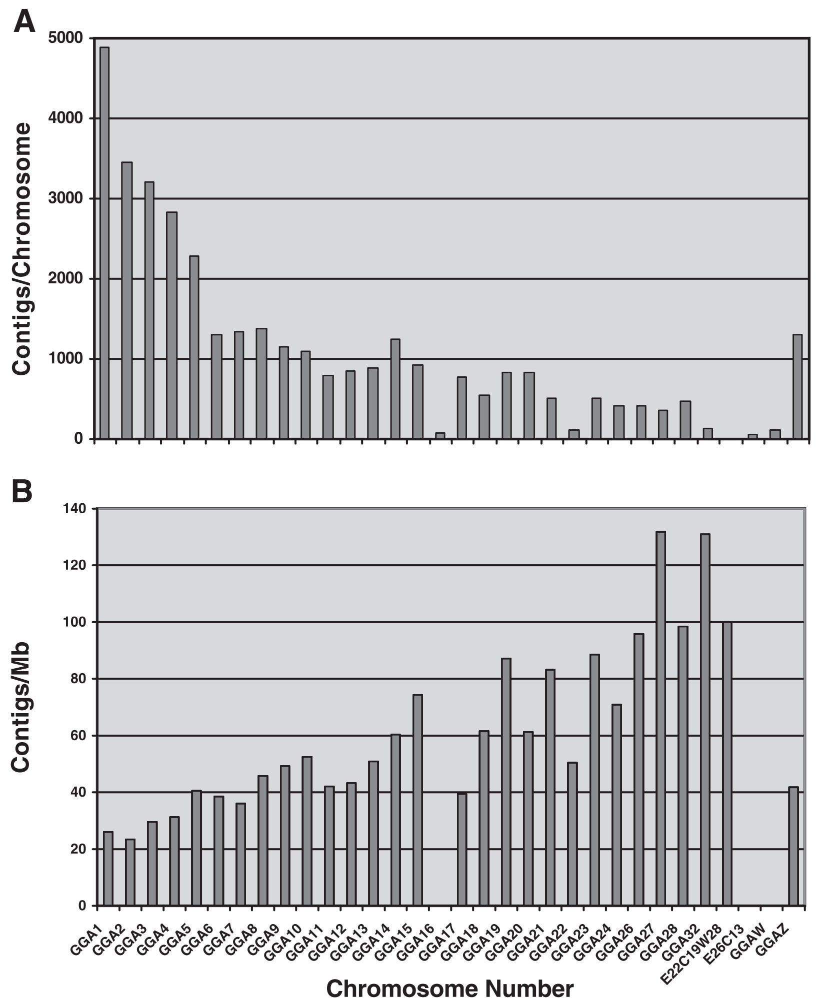

Fig. 2. Basic local alignment search tool (Blast)N analysis of the 40,850 CAP3 contig sequences (in silico cDNAs) against the chicken genome sequence shows a 96.7% match and wide chromosomal distribution (*A*). Chromosomal gene density (contigs/Mb) is based on sequenced chromosome size (*B*) from a physical map of the chicken genome (48). Gene density is not shown for 3 chromosomes that were poorly sequenced (*GGA*16, E26C13, and *GGA*W).

structed from either activated T cells (*pat*) (46) or mixed lymphoid tissues (*pgn1c*). The number of common genes shared between or among tissue groups is shown in the overlapping arcs of the Venn diagram (Fig. 3).

These unique gene sets were used to produce several custom chicken cDNA microarrays. Microarray platforms were entered into the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (**http://www.ncbi.nih.gov/geo**) for the 3.2K UD liver (GPL1742), the 7.4K UD metabolic system (GEO platform GPL1737), the 5K chicken neuroendocrine system (GPL1744), and the 14K DEL-MAR chicken integrated system (GPL1731) microarrays. The cDNA probes on each microarray are annotated with the highest BlastN and BlastX hits and hyperlinked to the UD EST or contig sequences in our CAP3 database (**http://cogburn.dbi.udel.edu/**) and to the chromosomal location on the chicken genome sequence (**http://genome.ucsc.edu/cgi-bin/hgGateway**). Some of the unique lymphoid clones from the UD collection (1,983 cDNAs) were incorporated into the 13K mixed-tissue microarray (GPL1836) produced at the Fred Hutchinson Cancer Research Center (Seattle, WA) by another chicken genomics consortium (7).

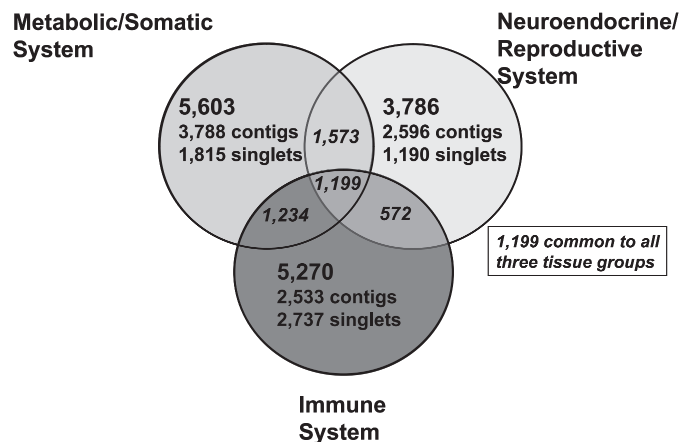

Fig. 3. Venn diagram of nonredundant chicken expressed sequence tag (EST) clusters that represent 3 major physiological (metabolic/somatic, neuroendocrine/reproductive, and immune) systems. The 19,237 unique sequences in the University of Delaware collection were clustered according to 3 major tissue groups. The no. of contigs and singlets within each cluster is indicated in smaller font. The no. of common genes among the 3 systems (1,199) and between each system pair is also indicated.

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.5 on April 8, 2017

DISCUSSION

A major goal of our functional chicken genomics project (L. A. Cogburn, T. E. Porter, S. E. Aggrey, and J. Simon) was limited EST sequencing of ~30,000 clones from tissues with the greatest agricultural importance for development of custom high-density cDNA microarrays. Our cDNA libraries were constructed from developmentally and genetically complex pools of RNA to increase novel gene discovery and reduce overall redundancy. This approach of pooling RNA samples from different animals, developmental stages, and tissues before normalization has yielded several high-quality cDNA libraries that were deeply sequenced for porcine (15) and bovine (39, 41) gene discovery. Presently, we have sequenced and functionally annotated an additional 35,407 chicken ESTs from several single and multiple tissue cDNA libraries. These sequences were entered into GenBank and the UD chick EST database (**http://www.chickest.udel.edu**) as accrued. The total UD collection has made a significant contribution to the present number of chicken ESTs in GenBank and to the assembly of the TIGR *Gg*GI (see Attribution at homepage: **http://www.tigr.org/tdb/tgi/gggi/**). Furthermore, the UD EST collection was important for the recent functional annotation of the chicken genome sequence (25).

Several international EST projects, including ours, have contributed to the total of 517,727 chicken EST sequences in GenBank (as of July 1, 2004). Our CAP3 assembly of these sequences into a chicken gene index has revealed a similar number of contigs (40,850) and singlets (79,192) as those found in the TIGR Chicken Gene Index (release 8.0) (**http://www.tigr.org/tdb/tgi/gggi/**). The number of contigs represented in these two chicken EST assemblies exceeds the original estimate of 35,000 genes expressed in the chicken genome (3). This large number of chicken contigs could be due in part to the presence of nonoverlapping fragments of identical transcripts. A more recent analysis of the chicken transcriptome, based on an analysis of 19,626 finished cDNAs and 485,337 public ESTs, suggests that there are at least 19,000 chicken genes (23). And, analysis of the first draft of the chicken genome sequence provides an estimate of 20,000–23,000 chicken genes (25). Furthermore, alternative splicing of exons can generate an even greater number of putative transcripts (25). Extensive alternative promoter usage, splicing, and polyadenylation contribute to a diverse transcriptome in the mouse of 181,047 transcripts (45). The total number of genes predicted for the chicken is similar to the human genome, which harbors from 20,000 to 25,000 protein-coding genes (24).

The UD EST clone collection represents the second largest catalog and repository of chicken EST clones, which were derived from tissues of major agricultural and biomedical importance. The UD EST collection has a minimum overlap with and is complementary to chicken sequences found in the larger BBSRC database (**http://www.chick.umist.ac.uk/**) (3) and the bursal (B) lymphocyte transcript database (**http://pheasant.gsf.de/DEPARTMENT/DT40/dt40Transcript.html**) (1, 6). Our chicken cDNA libraries were constructed from mixed lymphoid tissue and metabolic (liver and abdominal fat), somatic (breast and leg muscle/bone growth plate), neuroendocrine (pituitary/hypothalamus/pineal), and reproductive tissues (oviduct/ovary/testes). Several UD libraries represent novel tissues (i.e., lymphoid, abdominal fat, pituitary, hypothalamus, pineal, and oviduct) that are either not found in or underrepresented in other public chicken EST databases (3). Furthermore, 6,223 unique sequences (481 contigs and 5,742 singlets) are found only in the UD collection. Many of the unique singlets are from the immune system cluster, which is composed of a nearly equal number of contigs and singlets. The UD collection contains 5,742 unique singlet sequences (not found in other public EST collections), of which 74% match to the chicken genome sequence. One explanation of the higher rate of singlets not matching the chicken genome assembly could be the presence of high GC content sequences, which would reduce the frequency of G+C-rich ESTs sequenced from our libraries. However, the initial draft sequence is incomplete, with as many as 10% of the protein-coding genes still missing from the Ensembl gene set and with very poor coverage of microchromosomes and two chromosomes in particular, *GGA*16 and *GGA*W (25). Therefore, it seems reasonable that an even higher number of our unique ESTs would match the completely finished chicken genome sequence when it becomes available. Furthermore, a large number of contig and singlet sequences match a chicken genome sequence that is not yet assigned to a specific chromosome (i.e., *chr*Un). The finished chicken genome sequence could reveal an even greater density of genes on the microchromosomes, which have a higher recombination rate and a higher G+C content than the macrochromosomes (25).

The UD chicken EST collection contains a large number of lymphoid ESTs (12,261 clones) sequenced mainly from two unnormalized cDNA libraries: an activated T cell library (46) and a mixed lymphoid tissue library. Although unnormalized, the mixed lymphoid tissue (*pgn1c*) had a very low redundancy rate (18%) even after sequencing of 5,642 randomly picked clones. Numerous clusters of differentiation (CD) antigens, cytokines, cytokine receptors, and coagulation/complement factors were identified from ESTs sequenced from the UD lymphoid tissue cDNA libraries. The large number of unknown singlets found in the lymphoid tissue libraries (Supplemental Fig. S3*B*) may reflect unusual/rare clones that are less likely to have matches in the database.

The UD collection contains ESTs sequenced from other novel chicken cDNA libraries. For example, 6,739 chicken ESTs were sequenced from our adipose tissue cDNA libraries compared with only 2,672 ESTs sequenced from the BBSRC adipose tissue library, which completely failed normalization (3). *Osteonectin*, or *SPARC* (secreted protein acidic and rich in cysteine), was very abundant in the primary abdominal fat cDNA library (*pft1c*). This adipose-specific autocrine/paracrine factor (an "adipokine") is implicated in development of obesity in mice (44). Other adipokines identified in the UD chicken EST collection are *adiponectin* and *visfatin*. *Visfatin* is a newly discovered adipokine secreted from visceral fat that is thought to be a missing link between obesity and diabetes; two contigs in our CAP3 database (UD_Contig_2318.1 and UD_Contig_2318.2) represent chicken homologs of *visfatin*, previously identified as pre-B cell colony-enhancing factor (*PBEF1*) (37). Surprisingly, three very important genes involved in lipid metabolism in mammals, hormone-sensitive lipase (*HSL*), resistin (*RETN*), and leptin (*LEP*), have not yet been identified in our collection or among the 578,445 ESTs now sequenced from the chicken. However, the existence of chicken *LEP* (2, 43) remains very controversial (16, 34). The

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.5 on April 8, 2017

single EST clone in the BBSRC collection (clone ID no. ChEST698d23), originally identified as chicken *LEP*, appears to be a contaminating sequence that corresponds to bovine *LEP* [i.e., 98% identical to bovine *LEP* (TIGR *Bt*GI TC292189)]. Furthermore, this BBSRC EST sequence for "chicken" *LEP* fails to show a BlastN hit to the chicken genome sequence. Searches over the genomic region of the chicken comparable with human *LEP* synteny also revealed no evidence of a chicken *LEP* gene. In an exhaustive PCR analysis of chicken *LEP* with multiple primer sets designed from two published cDNA sequences (2, 43), we have consistently failed to produce an amplified PCR product using liver, fat, and muscle total RNA as template (W. Carre, X. Wang, and L. A. Cogburn, unpublished observations). However, two EST clones corresponding to the chicken *LEP* receptor (*LEPR*) cDNA sequence (21, 32) were found in our neuroendocrine library (*pgp1n*).

Our liver cDNA library is also populated by a large number of genes involved in lipogenesis (*adipophilin*, *fatty acid-binding protein*, *Spot 14*, *fatty acid desaturase*, *malic enzyme*, *Δ-9 desaturase*, etc.). The large number of lipogenic genes found in the chicken's liver reflects a specific feature of avian metabolism, where the liver is the major site of lipogenesis (18, 19). The chicken homolog of *apolipoprotein AV* (*ApoAV*) is represented by UD_Contig_12151.1, which was assembled from 13 UD ESTs (11 ESTs from the liver) and 3 public ESTs. This newest member of the apolipoprotein gene cluster (*ApoAV*) was recently revealed by a comparative analysis of the human and mouse genome sequences (33). Three single nucleotide polymorphisms (SNPs) were found across the *ApoAV* locus in humans that are associated with plasma triglyceride levels. One SNP in the promoter region of human *ApoAV* has garnered a great deal of attention as an important determinant of plasma triglyceride levels and a potential molecular marker for diagnosis of cardiovascular disease (38). Detailed sequence alignment and BlastN analysis of chicken *ApoAV* (UD Contig_12151.1) against the chicken genome sequence revealed its chromosomal location (*GGA*24_random at 115,638–116,824) within an apolipoprotein gene cluster and several potentially important polymorphisms: nine SNPs in the coding region and a 7-bp insertion/deletion polymorphism located in the proximal promoter region near the TATA box (L. A. Cogburn, X. Wang, and W. Carre, unpublished observations). Thus the genetic complexity of our cDNA libraries makes the UD EST collection a valuable resource for discovery of important chicken genes and identification of polymorphisms (11).

A large number of ESTs (8,734) were derived from our neuroendocrine cDNA libraries, which were constructed from the pituitary, hypothalamus, and pineal gland. Numerous pituitary-specific hormones [*POMC*, *growth hormone* (*GH*), and *prolactin* (*PRL*)], hormone receptors [*leptin receptor*, *growth hormone-releasing hormone receptor* (*GHRH-R*)], and transcription factors [*Pit-1*, *sterol response element-binding protein 2* (*SREBP2*)] were sequenced from the neuroendocrine libraries. A number of these gene sequences are unique to our neuroendocrine cDNA libraries and to the UD collection. These include the preprohormone *POMC*, which yields multiple peptide products from proteolytic cleavage to generate β-endorphin, β-lipotropin (β-LPH), α-melanocyte-stimulating hormone (α-MSH), and adrenocorticotropic hormone (ACTH). In birds, *POMC* products play critical roles in the regulation of growth, metabolism, and the adaptive stress responses. However, several genes were noticeably absent from our ESTs sequenced from the neuroendocrine libraries [i.e., *pre-pro-thyrotropin-releasing hormone* (*TRH*), *corticotropin-releasing hormone* (*CRH*), *somatotropin release-inhibiting factor* (*SRIF*), *gonadotropin-releasing hormone* (*GnRH*), and *luteinizing hormone beta subunit* (*LH-β*)]. Interestingly, three elements of the somatotropic axis that regulate animal growth and development were unique to these libraries and the UD EST collection: *GH*, *GHRH*, and *GHRH-R* genes. Another interesting example of a gene that is unique to our collection is Contig_13370.2 (assembled from 11 UD ESTs from the neuroendocrine libraries), which represents the chicken homolog of *β3-tubulin*. Furthermore, the sequencing of a large number of redundant ESTs from the "unnormalized" neuroendocrine library (*pgp1n*) has contributed to the discovery of SNPs in a number of important pituitary hormones (i.e., *POMC*, *GH*, *PRL*, etc.) (27).

An initial chicken SNP discovery effort, initiated by another UD group (27), identified 1,210 SNPs from a subset of 23,427 UD ESTs. However, a more comprehensive polymorphism map was recently developed for the chicken that contains 2.8 million SNPs or about five SNPs per kilobase of genome sequence (26). The chicken genetic variation map (**http://chicken.genomics.org.cn/index.jsp**) is based on comparison of the 0.25 × coverage of genome sequence from three distinct domestic breeds (broiler, layer, and Chinese silkie) against the 6.6 × coverage of the red jungle fowl genome sequence (25). Thus the integration of the genetic variability from 549,157 EST sequences (**http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html**) with the chicken genome sequence (25) and the polymorphism map (26) now provides abundant genomic resources required for fine mapping of quantitative trait loci (QTL) and the eventual identification of polymorphic genes that control many important phenotypic traits.

This recent cache of chicken genomic resources has provided us with the first global view of the homology between the chicken and mammalian genomes and proteomes. BlastX analysis of a set of high-fidelity UD CAP3 contigs against the human protein database showed a median amino acid identity of 73% for the putative chicken homologs, which is very similar to the recent estimate of 76% homology derived from the draft chicken genome sequence (25). About 17% of the contigs and 39% of the singlets in the UD EST collection have no BlastX hit against the GenBank nonredundant (nr) database. Some of these sequences could reflect highly divergent or rarely expressed transcripts. One contig (UD Contig_13866.1), which was assembled from 269 ESTs derived from many tissues and is therefore authentic, is a good example of a highly divergent transcript. This unknown transcript is abundantly expressed in the unnormalized mixed lymphoid tissue library (*pgn1c*) (Supplemental Fig. S1*A*). Even after an extensive Blast search, the identity of this putative gene remains unknown. Detailed analysis of unknown sequences could lead to the identification of additional orthologs and paralogs. For example, we have discovered two contigs that represent chicken paralogs (*THRSPα* and *THRSPβ*) of the human thyroid hormone-responsive Spot 14 protein gene (*THRSP*) (20). These high-fidelity contigs encode amino acid sequences that are only 29% identical to the THRSP human protein. Our CAP3 assembly of two unique contigs for *Spot 14* (*THRSP*) allowed us to

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.5 on April 8, 2017

identify this unique gene duplication, which was not revealed by the BBSRC or TIGR chicken EST assemblies. THRSP is an important transcription factor that controls expression of several metabolic genes in the lipogenic pathway (47). We have identified insertion/deletion polymorphisms near the DNA-binding domain of chicken *THRSPα* that are associated with a QTL for abdominal fatness located on *GGA*1 (49). *THRSPβ* has a very high G+C content, which we discovered in a single shotgun sequence that was not included in the draft chicken genome sequence. Thus our chicken genomic resources (ESTs, CAP3 database, and tissue-specific microarrays) have been very useful for gene discovery, expression profiling, and identification of major genes that control economically important production traits in the broiler chicken (10, 11, 49).

In summary, we have sequenced 35,407 chicken ESTs from developmentally and genetically complex cDNA libraries that are either absent from or not well represented in other public EST databases. The UD ESTs have been integrated into a comprehensive catalog of expressed chicken genes that will aid the discovery of sequence polymorphisms. The CAP3 assembly of our ESTs with publicly available sequences was used for annotation and selection of nonredundant sets of cDNA clones. The UD EST collection contains 19,237 nonredundant cDNA sequences derived from major physiological (immune, metabolic/somatic, and neuroendocrine/reproductive) systems. Unique system-specific gene sets (Fig. 3) have been used for production of custom chicken cDNA microarrays for transcription profiling. These initial chicken cDNA microarrays have given us the first glimpse of the chicken's transcriptome (7, 12, 30). The availability of high-density microarrays, the immediate access to the CAP3 chicken EST assemblies, and the large number of physical cDNAs in the UD collection (43,928 EST clones) greatly enhance the value of our genomic resources for the chicken. Several chicken EST sequencing projects, including the present one, have now placed the chicken in 10th place for accrued ESTs among all organisms represented in GenBank (**http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html**). Furthermore, the large international chicken EST collection was essential for the recent assembly and annotation of the first draft of the chicken genome sequence (25). These important new developments, acquisition of large public collections of ESTs, a completed genome sequence, and a dense polymorphism map, emphasize an important new role for the chicken (*G. gallus*) in developmental biology and genomics research and a continuing role in the advancement of biomedical sciences.

ACKNOWLEDGMENTS

We express appreciation to E. Monsonego-Ornan, Volcani Center, Agricultural Research Organization, Bet-Dagan, Israel, for supplying the epiphyseal growth plate RNA used in construction of the chicken skeletal muscle/epiphyseal growth plate cDNA library (*pgm1c*). We greatly appreciate the excellent efforts of Maureen Dolan and Mike Hanefy in the high-throughput DNA sequencing of our chicken cDNA libraries at Dupont Agricultural Products, Delaware Technology Park, Newark, DE. We also express appreciation to G. R. Gao and Y. Niu at the Delaware Biotechnology Institute (DBI) for the use of the SunFire 4800 Workstation and to Dianne York (DBI Microarray Core Facility) for management of EST clones and for rearraying unigene sets used in production of microarrays. We gratefully acknowledge the earlier contributions of Luc Sofer, L. Amarin Cogburn, V. G. Tirunagaru, and J. Cui in the development of these genomic resources.

GRANTS

This work was supported by grants from the USDA-NRI (to J. Burnside and R. Morgan; award no. 00-35205-9407); the Initiative for Future Agricultural and Food Systems (USDA-IFAFS) (to L. A. Cogburn, T. E. Porter, S. E. Aggrey, and J. Simon; award no. 00-52100-9614) Cooperative State Research, Education, and Extension Service (USDA-CSREES) Animal Genome Program; the Maryland Agricultural Experiment Station (to T. E. Porter); and the Institut National de la Recherche Agronomique [INRA, Analyse du Genome des Animaux d'Elevage (AGENAE) Program] (to Y. Nys). The University of Delaware Life Sciences Core Instrumentation (Bioinformatics and Microarray) Centers at Delaware Biotechnology Institute are supported by grants from the National Center for Research Resources (grant no. 2P20RR-16472-05) and the National Science Foundation (grant no. EPS-0447610).

REFERENCES

1. **Abdrakhmanov I, Lodygin D, Geroth P, Arakawa H, Law A, Plachy J, Korn B, and Buerstedde JM.** A large database of chicken bursal ESTs as a resource for the analysis of vertebrate gene function. *Genome Res* 10: 2062–2069, 2000.
2. **Ashwell CM, Czerwinski SM, Brocht SM, and McMurtry JP.** Hormonal regulation of leptin expression in broiler chickens. *Am J Physiol* 45: 226–232, 1999.
3. **Boardman PE, Sanz-Ezquerro J, Overton IM, Burt DW, Bosch E, Fong WT, Tickle C, Brown WRA, Wilson SA, and Hubbard SJ.** A comprehensive collection of chicken cDNAs. *Curr Biol* 12: 1965–1969, 2002.
4. **Bonaldo MF, Lennon G, and Soares MB.** Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* 6: 791–806, 1996.
5. **Brown WRA, Hubbard SJ, and Wilson SA.** The chicken as a model for large-scale analysis of vertebrate gene function. *Nat Rev Genet* 4: 87–98, 2003.
6. **Buerstedde JM, Arakawa H, Watahiki A, Carninci PP, Hayashizaki YY, Korn B, and Plachy J.** The DT40 web site: sampling and connecting the genes of a B-cell line. *Nucleic Acids Res* 30: 230–231, 2002.
7. **Burnside J, Neiman P, Tang J, Basom R, Talbot R, Aronszajn M, Burt D, and Delrow J.** Development of a cDNA array for chicken expression analysis. *BMC Genomics* 6: 13, 2005.
8. **Burt DW.** The chicken genome and the developmental biologist. *Mech Dev* 121: 1129–1135, 2004.
9. **Chirgwin JM, Przybla AE, MacDonald RJ, and Rutter WJ.** Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* 18: 5294–5299, 1979.
10. **Cogburn LA, Wang X, Carre W, Rejto L, Aggrey SE, Duclos MJ, Simon J, and Porter TE.** Functional genomics in chickens: developement of integrated-systems microarrays for transcriptional profiling and discovery of regulatory pathways. *Comp Funct Genom* 5: 253–261, 2004.
11. **Cogburn LA, Wang X, Carre W, Rejto L, Porter TE, Aggrey SE, and Simon J.** Systems-wide chicken DNA microarrays, gene expression profiling and discovery of functional genes. *Poult Sci* 82: 939–951, 2003.
12. **Cui J, Sofer L, Cloud SS, and Burnside J.** Patterns of gene expression in the developing chick thymus. *Dev Dyn* 229: 480–488, 2004.
13. **Dohner JV.** *The Encyclopedia of Historic and Endangered Livestock and Poultry Breeds*. New Haven, CT: Yale Univ. Press, 2001.
14. **Ewing B and Green P.** Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. *Genome Res* 8: 186–194, 1998.
15. **Fahrenkrug SC, Smith TPL, Freking BA, Cho J, White J, Vallet J, Wise T, Rohrer G, Pertea G, Sultana R, Quackenbush J, and Keele JW.** Porcine gene discovery by normalized cDNA-library sequencing and EST cluster assembly. *Mamm Genome* 13: 475–478, 2002.
16. **Friedman-Einat M, Boswell T, Horev G, Girishvarma G, Dunn IC, Talbot RT, and Sharp PJ.** The chicken leptin gene: has it been cloned? *Gen Comp Endocrinol* 115: 354–363, 1999.
17. **Glisin V, Grkvenjakov R, and Byus C.** Ribonucleic acid isolation by cesium chloride centrifugation. *Biochemistry* 13: 2633–2637, 1974.
18. **Goodridge AG and Ball EG.** Lipogenesis in the pigeon: in vivo studies. *Am J Physiol* 213: 245–249, 1967.
19. **Goodridge AG, Garay A, and Silpananta P.** Regulation of lipogenesis and the total activities of lipogenic enzymes in a primary culture of hepatocytes from prenatal and early postnatal chicks. *J Biol Chem* 249: 1469–1475, 1974.

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.5 on April 8, 2017

20. **Grillasca JP, Gastaldi M, Khiri H, Dace A, Peyrol N, Reynier P, Torresani J, and Planells R.** Cloning and initial characterization of human and mouse Spot 14 genes. *FEBS Lett* 401: 38–42, 1996.
21. **Horev G, Einat P, Aharoni T, Eshdat Y, and Friedman-Einat M.** Molecular cloning and properties of the chicken leptin-receptor (CLEPR) gene. *Mol Cell Endocrinol* 162: 95–106, 2000.
22. **Huang X and Madan A.** CAP3: a DNA sequence assembly program. *Genome Res* 9: 868–877, 1999.
23. **Hubbard SJ, Grafham DV, Beattie KJ, Overton IM, McLaren SR, Croning MDR, Boardman PE, Bonfield JK, Burnside J, Davies RM, Farrell ER, Francis MD, Griffiths-Jones S, Humphray SJ, Hyland C, Scott CE, Tang H, Taylor RG, Tickle C, Brown WRA, Birney E, Rogers J, and Wilson SA.** Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags. *Genome Res* 15: 174–183, 2005.
24. **Human Genome Sequencing Consortium.** Finishing the euchromatic sequence of the human genome. *Nature* 431: 931–945, 2004.
25. **International Chicken Genome Sequencing Consortium.** Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432: 695–716, 2004.
26. **International Chicken Polymorphism Map Consortium.** A genetic variation map for chicken with 28 million single-nucleotide polymorphisms. *Nature* 432: 717–722, 2004.
27. **Kim H, Schmidt CJ, Decker KS, and Emara MG.** A double-screening method to identify reliable candidate non-synonymous SNPs from chicken EST data. *Anim Genet* 34: 249–254, 2003.
28. **Liu HC, Cheng HH, Tirunagaru VG, Sofer L, and Burnside J.** A strategy to identify positional candidate genes conferring Marek's disease resistance by integrating DNA microarrays and genetic mapping. *Anim Genet* 32: 351–359, 2001.
29. **Mao JNC, Burnside J, Postel-Vinay MC, Chambers J, Pesek J, and Cogburn LA.** Ontogeny of growth hormone receptor gene expression in tissue of growth-selected strains of broiler chickens. *J Endocrinol* 156: 67–75, 1998.
30. **Mirnics ZK, Caudell E, Gao YH, Kuwahara K, Sakaguchi N, Kurosaki T, Burnside J, Mirnics K, and Corey SJ.** Microarray analysis of Lyn-deficient B cells reveals germinal center-associated nuclear protein and other genes associated with the lymphoid germinal center. *J Immunol* 172: 4133–4141, 2004.
31. **Neiman PE, Ruddell A, Jasoni C, Loring G, Thomas SJ, Brandvold KA, Lee R, Burnside J, and Delrow J.** Analysis of gene expression during myc oncogene-induced lymphomagenesis in the bursa of Fabricius. *Proc Natl Acad Sci USA* 98: 6378–6383, 2001.
32. **Ohkubo T, Tanaka M, and Nakashima K.** Structure and tissue distribution of chicken leptin receptor (cOb-R) mRNA. *Biochim Biophys Acta* 1491: 303–308, 2000.
33. **Pennacchio LA, Olivier M, Hubacek JA, Cohen JC, Cox DR, Fruchart JC, Krauss RM, and Rubin EM.** An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* 294: 169–173, 2001.
34. **Pitel F, Monbrun C, Gellin J, and Vignal A.** The chicken LEP (OB) gene has not been mapped. *Anim Genet* 31: 281, 2000.
35. **Price EO.** *Animal Domestication and Behavior*. Wallingford, UK: CABI, 2002.
36. **Reich A, Jaffe N, Tong A, Lavelin I, Genina O, Pines M, Sklan D, Nussinovitch A, and Monsonego-Ornan E.** Weight loading young chicks inhibits bone elongation and promotes growth plate ossification and vascularization. *J Appl Physiol* 98: 2381–2389, 2005.
37. **Samal B, Sun Y, Stearns G, Xie C, Suggs S, and McNiece I.** Cloning and characterization of the cDNA encoding a novel human pre-B-cell colony-enhancing factor. *Mol Cell Biol* 14: 1431–1437, 1994.
38. **Seydel C.** Genome research: possible new heart disease risk factor. *Science* 294: 33a, 2001.
39. **Smith TPL, Grosse WM, Freking BA, Roberts AJ, Stone RT, Casas E, Wray JE, White J, Cho J, Fahrenkrug SC, Bennet GL, Heaton MP, Laegreid WW, Rohrer GA, Chitko-McKown CG, Pertea G, Holt I, Karamycheva S, Liang F, Quackenbush J, and Keele JW.** Sequence evaluation of four pooled-tissue normalized bovine cDNA libraries and construction of a gene index for cattle. *Genome Res* 11: 626–630, 2001.
40. **Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, and Efstratiadis A.** Construction and characterization of a normalized cDNA library. *Proc Natl Acad Sci USA* 91: 9228–9232, 1994.
41. **Sonstegard TS, Capuco A, White J, Tassell VT, Conner EE, Cho J, Sultana R, Shade L, Wray JE, Wells KD, and Quackenbush J.** Analysis of bovine mammary gland EST and functional annotation of the *Bos taurus* gene index. *Mamm Genome* 13: 373–379, 2002.
42. **Stern CD.** The chick: a great model system becomes even greater. *Devel Cell* 8: 9–17, 2005.
43. **Taouis M, Chen JW, Daviaud C, Dupont J, Derouet M, and Simon J.** Cloning the chicken leptin gene. *Gene* 208: 239–242, 1998.
44. **Tarte-Deckert S, Chavey C, Monthouel MN, Gautier N, and van Obberghen E.** The matricellular protein SPARC/osteonectin as a newly identified factor up-regulated in obesity. *J Biol Chem* 276: 22231–22237, 2001.
45. **The FANTOM Consortium and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group).** The transcriptional landscape of the mammalian genome. *Science* 309: 1559–1563, 2005.
46. **Tirunagaru VG, Sofer L, Cui J, and Burnside J.** An expressed sequence tag database of T-cell-enriched activated splenocytes: sequence analysis of 5251 clones. *Genomics* 66: 144–151, 2000.
47. **Towle HC, Kaytor EN, and Shih HM.** Regulation of the expression of lipogenic enzyme genes by carbohydrate. *Annu Rev Nutr* 17: 405–433, 1997.
48. **Wallis JW, Aerts J, Groenen MAM, Crooijmans RPMA, Layman D, Graves TA, Scheer DE, Kremitzki C, Fedele MJ, Mudd NK, Cardenas M, Higginbotham J, Carter J, McGrane R, Gaige T, Mead K, Walker J, Albracht D, Davito J, Yang SP, Leong S, Chinwalla A, Sekhon M, Wylie K, Dodgson J, Romanov MN, Cheng H, de Jong PJ, Osoegawa K, Nefedov M, Zhang H, McPherson JD, Krzywinski M, Schein J, Hillier LD, Mardis ER, Wilson RK, and Warren WC.** A physical map of the chicken genome. *Nature* 432: 761–764, 2004.
49. **Wang X, Carre W, Zhou H, Lamont SJ, and Cogburn LA.** Duplicated Spot 14 genes in the chicken: characterization and identification of polymorphisms associated with abdominal fat traits. *Gene* 332: 79–88, 2004.

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.5 on April 8, 2017