Available online at www.sciencedirect.com
SCIENCE DIRECT®
Genomics 88 (2006) 252–257
GENOMICS
www.elsevier.com/locate/ygeno

ChickGCE: A novel germ cell EST database for studying the early developmental stage in chickens

Heebal Kim [a,b,1], Dajeong Lim [a,1], Beom Ku Han [c], Samsun Sung [a], Mina Jeon [b], Sunjin Moon [a], Yeonkyung Kang [c], Jungrye Nam [a], Jae Yong Han [a,*]

[a] Department of Food and Animal Biotechnology, Seoul National University, Seoul 151-742, Korea
[b] Graduate Program in Bioinformatics, Seoul National University, Seoul 151-921, Korea
[c] Department of Genomics, Avicore Biotechnology Institute, Hanlim Human Tower, Gyeonggi-Do 435-050, Korea

Received 3 November 2005; accepted 25 March 2006
Available online 22 May 2006

Abstract

We established a database to study germ cells during the early developmental stage in the chicken. The ChickGCE database provides integrated expressed sequence tag (EST) data from chicken testis, ovary, embryonic gonads, and primordial germ cells. We gathered data on 10,294 ESTs from approximately 1000 embryonic gonads, and we experimentally determined 10,851 ESTs from primordial germ cells purified from 7955 embryonic gonads by magnetically activated cell sorting. The EST testis and ovary datasets were retrieved from the public database of The Institute for Genomic Research (TIGR). The EST data were clustered and assembled into unique sequences, contigs, and singletons. The ChickGCE database provides functional annotation, identification, and putative embryonic germ-cell-specific novel transcripts based on the Gene Ontology database, as well as statistical analyses of expression patterns and pair-wise comparisons of two types of tissue- and germ-cell-specific alternative splicing events in the chicken. The new database is accessible online and queries can be answered using several search options, including tissue database searches, keywords, clone IDs, expected values, and BLAST search scores.

Keywords: chicken; germ cell; embryonic gonad; EST database

Transcripts encoding fertility functions exist in meiotic or postmeiotic cells but not in most somatic cells [1,2]. The developmental stage- or germ-cell-specific expression of genes is an essential part of a functional analysis. For example, recent genome-wide analysis of germ-line-enriched and sex-biased expression profiles in *Caenorhabditis elegans* [3] suggested that sex- and germ-line-related genes are nonrandomly distributed in the genome. Functional studies have been applied not only to mammalian cells but also to yeast [4], worm [5], and avian species. The domestic chicken (*Gallus gallus*) is a better model organism than mammals for studying developmental biology and transgenesis [6–8]. An embryogenesis-related gene expression database is available for computer modeling and simulation of development in *C. elegans* [9]; GermOnline (http://www.germonline.org) focuses on mitotic growth and meiotic development for large-scale expression profiling analysis in germ cells [10] and provides microarray expression data for several organisms. However, no database has previously focused on avian-related information.

The number of genome-wide analyses of the chicken has increased. The initial sequencing and analysis of the chicken genome generated a physical map representing about 91% of the genome [11] and a large amount of chicken expressed sequence tag (EST) data has been produced [12–14]. However, information on germ cells that is relative to developmental biology is limited and difficult to obtain experimentally. To date, only four embryonic gonad libraries exist, all for mice; no embryonic gonad or primordial germ cell libraries for the human or chicken are available. Primordial germ cells (PGCs) differentiate into functional gametes, sperm in males or oocytes in females, after sexual

* Corresponding author. Fax: +82 2 874 4811.
E-mail address: jaehan@snu.ac.kr (J.Y. Han).
[1] These authors should be regarded as joint first authors.

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.03.015

maturity. Shin et al. noted that in the majority of vertebrates, sex is determined genetically, but sexual differentiation is initiated only during gonadal development. In the case of chicken PGCs, recent research has compared the chicken to various other organisms [15], and embryonic gonads and PGCs have been determined to be critical in embryonic germ cell development and sex determination.

Experimental data on embryonic gonads were previously obtained from approximately 1000 embryos at stage 29, after 6.5 days of incubation [16], and PGCs were generated from embryonic primordial germ cells. To create chicken PGC cDNA libraries, embryonic gonads were collected from 7955 White Leghorn (WL) chicken embryos at 6.5 days and PGCs were extracted from the embryonic gonadal cells using magnetically activated cell sorting (MACS) [17]. For the EST database, we collected 10,294 ESTs from extant chicken data and determined 10,851 ESTs from chicken PGCs from laboratory data. We constructed an online database using the higher quality ESTs following a strict technical methodology. This project was based on gene annotation and expression profiling of other tissues, but using germ-cell-related EST data for the chicken. The functional annotation of gene ontology is presented according to the BIOSIS Controlled Vocabulary file. Gene expression profiling was analyzed by statistical methods using Audic's and Susko's tests. The analysis of germ-cell-specific alternative splicing was developed from our simple algorithm. This paper describes how the user can obtain data from the ChickGCE database.

Results

The database is available at http://chickgce.snu.ac.kr. The Web interface of the ChickGCE database search consists of three search pages. The first is a BLAST page, which is useful for searching sequence similarity against the germ-cell-specific database for gonads and PGCs. BLASTN, TBLASTN, and TBLASTX are available with user-specified *E* values and filtering. To allow searches of nucleotide or amino acid sequences, the input sequence must be in FASTA format or consist of bare sequences. The BLAST results are shown on a new page in an output format similar to that of the National Center for Biotechnology Information (NCBI) site and provide information on the function of the matched sequences.

The second page, "Search by Keywords," provides information using germ-cell-related keywords in the ChickGCE database. Stored information on embryonic gonads and PGCs in the database can be obtained by the user in the form of a local ID, accession number, or description of homology against the nonredundant NCBI protein database using the BLASTX score, BLAST score, and *E* value. The third page, "Search by Clone ID," can locate germ cell ESTs, contigs, or singletons using the tissue dataset. A search with no entry will list all data in ChickGCE on the output page. "Search by Clone ID" retrieves the same information as "Search by Keywords."

ChickGCE also contains simple query interfaces for gene ontology, expression profiling, and novel alternative splicing events (Fig. 1). The gene ontology section is divided into three

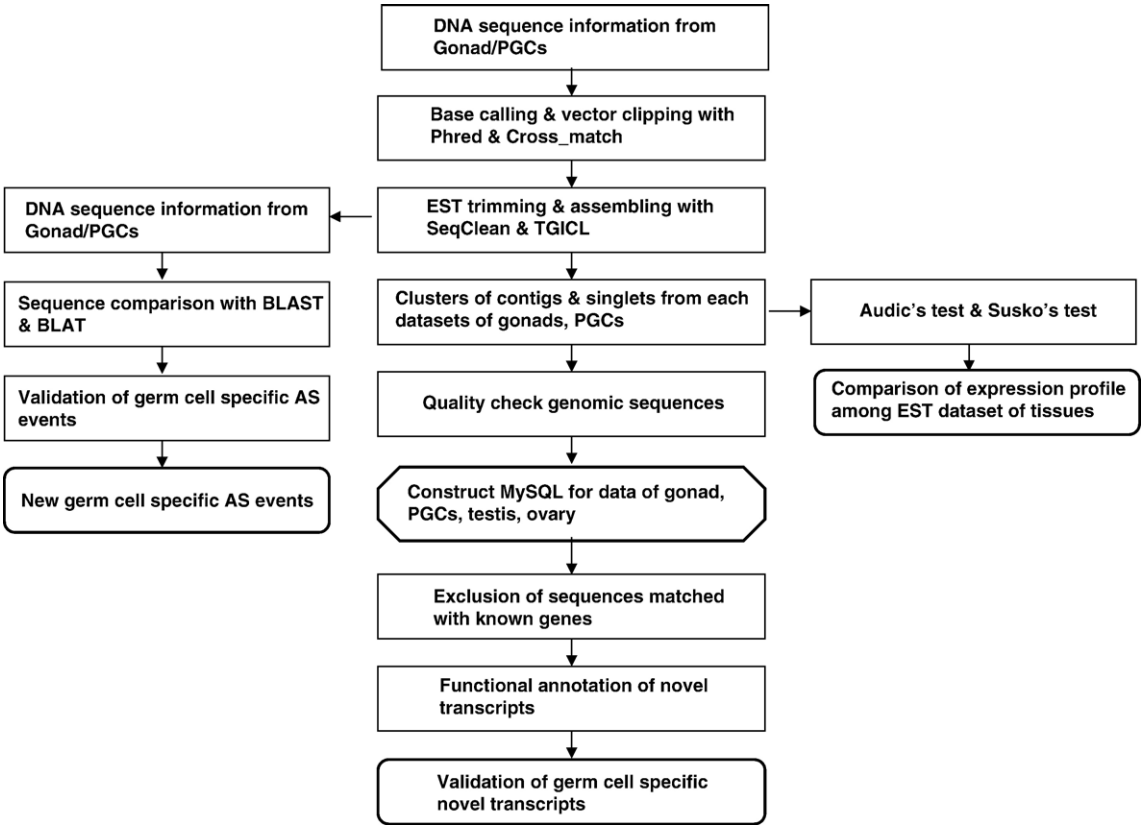


Fig. 1. Data processing pipeline of the ChickGCE.

Table 1
EST datasets in the ChickGCE database

Type	Number of samples	Library ID of TIGR GgGI (supplier)	ESTs	Total ESTs
Embryonic gonads	About 1000 embryos	No ID (Seoul National Univ.)	10,294	10,294
Primordial germ cells	7955 embryos	No ID (Seoul National Univ.)	10,851	10,851
Testis	No information	F6A (Uppsala Univ.)	4,884	10,397
	No information	F6B (Uppsala Univ.)	5,513	
Ovary	No information	AU5 (Univ. of Manchester)	5,566	29,708
	No information	AU6 (Univ. of Manchester)	7,941	
	No information	AU7 (Univ. of Manchester)	8,495	
	No information	AU9 (Univ. of Manchester)	6,212	
	No information	AV0 (Univ. of Manchester)	1,494	

categories: molecular function, biological process, and cellular components. All data are categorized from first-level terms to last-level terms. In the current level, the database displays the description and GO number of the child terms, the percentage of total sequences in ChickGCE, and the percentage of total sequences in the TIGR GgGI dataset. The user can compare two datasets in second-level terms. GO annotation can be used to display a tree view showing the hierarchical structure of the relationships among categories. The user can select tissues and functional annotation, including annotated sequences, nonannotated sequences, and all sequences corresponding to novel transcripts identified by the BLAST results and GO annotation. The putative novel transcripts of gonads and PGCs can be downloaded.

Audic's and Susko's tests were used to determine the significance of gene expression profiles. We needed to correct the significance level relative to the two statistical tests because of the different numbers of multiple tests for each pair of libraries and the different correction methods used for each test, i.e., the Bonferroni and Benjamini corrections for Audic's and Susko's tests, respectively. The corrected significance level is shown on the Web page. A pair of tissue-specific sequences can be described differently based on nine options. This search is a particularly useful tool as little research on the early embryo developmental stage has yet been conducted.

The output indicates whether the case event was an EST query or a TIGR TC ID corresponding to an EST query and reports the case type and case position within the genomic region. If the user then clicks on the case position, the results are displayed graphically, showing the insertion position in a pink box and the deletion position in a light green box. The output format is shown in Fig. 1. The database also provides a sequence view according to the genomic sequence and shows whether a TIGR sequence includes an insertion or deletion, or both. In the sequence displayed, bold letters represent exon–intron boundaries, blue letters indicate exons, and red letters designate case sequences in the genomic sequence view. Each flowchart is described on the Web pages. In the near future, gonad- and PGC-specific alternative splicing variants will be examined using a RT–PCR analysis of various tissues and organs.

Discussion

We describe a novel functional database related to the germ line for determining gene annotation, novel transcripts, expression profiling, and alternative splicing events in the chicken. The ChickGCE database was established using extant germ cell EST datasets for the chicken and by analyzing gonads and PGCs in the early embryonic developmental stages experimentally. ChickGCE is a useful tool for those studying developmental biology and biological mechanisms in the chicken. For example, the number of putative novel transcripts obtained using BLAST with the GgGI dataset was 1215 sequences containing 131 contigs and 1084 singletons, and with a BLAST best hit, 10% of all putative novel transcripts were derived from chicken data. There were 14 and 22 germ-cell-specific alternative splicing events in the embryonic gonadal dataset and PGC dataset, respectively. The database model was designed to analyze two case events simultaneously. Of the 14 events in the gonad dataset, 7 alternative splicing (AS) events were case 1 (insertion) events, and 7 were case 2 (deletion) events. Of the 22 events in the PGC dataset, 14 AS events were case 1 events, and 8 were case 2 events. ChickGCE allows users to obtain and apply useful data related to developmental biology in the chicken.

Materials and methods

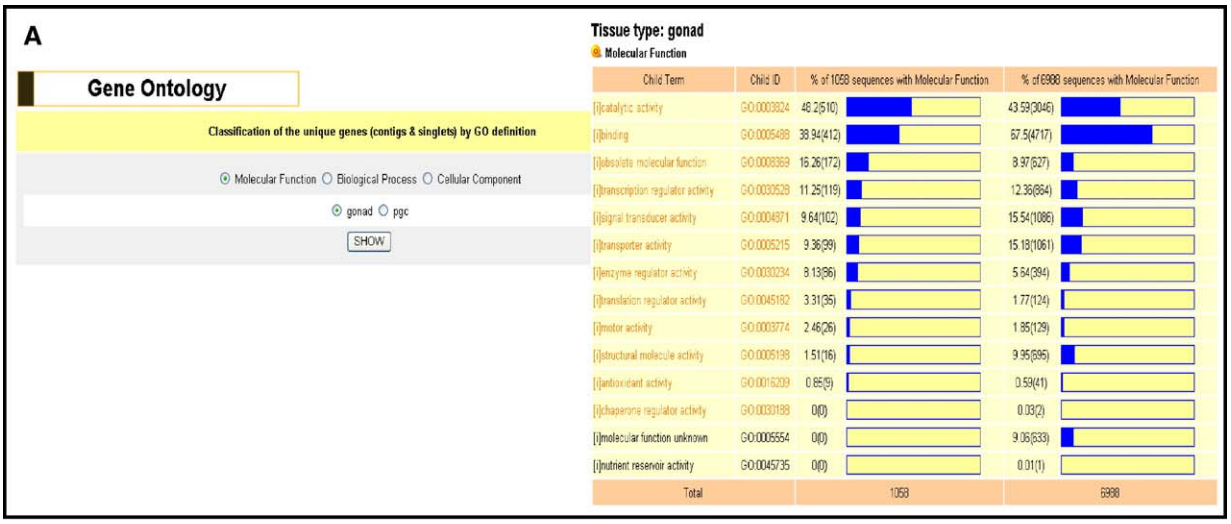
Construction of EST data

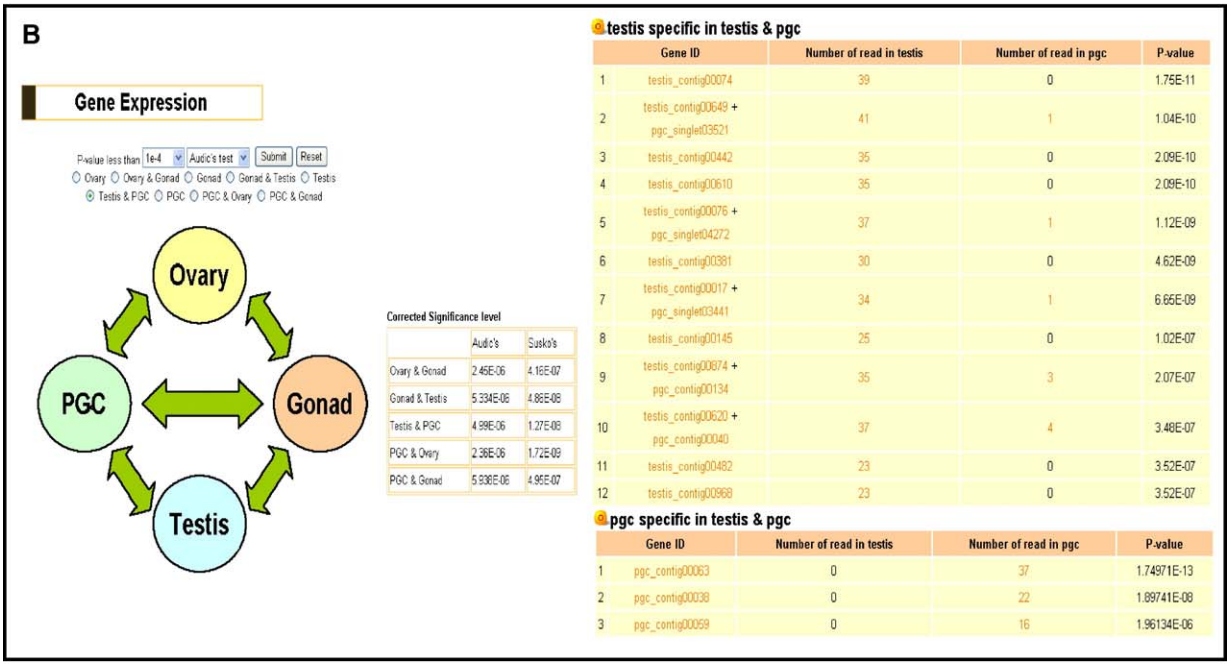
To obtain a large number of freshly prepared embryonic gonads from eggs and to retrieve primordial germ cells from the gonads, embryos at stage 29 (6.5 days of incubation) were produced from an inbred strain of WL chickens maintained at the University Animal Farm, Seoul National University, Korea. We collected 10,294 ESTs from an extant chicken cDNA library that was constructed from approximately 1000 highly pure embryonic gonads; the EST sequences were determined by priming the 5′ end of cDNA followed by analysis with an ABI 3700 DNA sequencer [16]. For PGCs, we used MACS to determine experimentally 10,851 ESTs from primordial germ cells, purified from 7955 embryonic gonads. The PGC population ratio increased approximately 47.4-fold after MACS separation in comparison with before MACS (35.1% vs 0.74%) [17]. The registered accession numbers from the NCBI dbEST database are *CV852525–CV862818* for gonads and *DR410159–DR421006* for PGCs [16,17]. The chicken EST trace data were base-called using phred [18] and vector-clipped by a cross-match program [19] with vector sequences. Vector-screened EST sequences were filtered and trimmed by removing contaminating

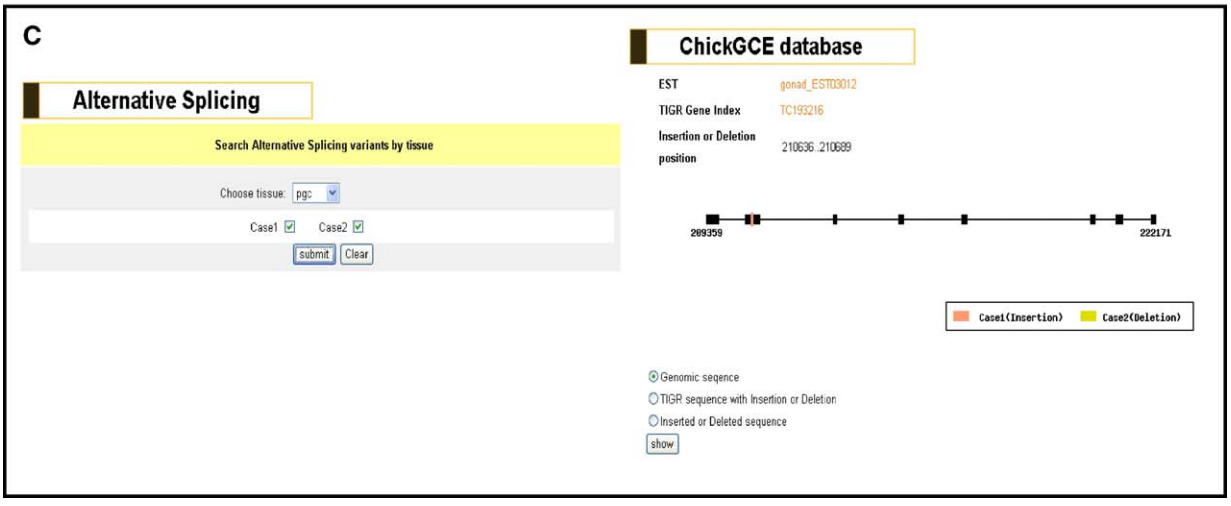
Fig. 2. Simple query and output format of the ChickGCE. Users can search for individual genes and sequence information by search options. (A) Of search interfaces, Gene Ontology enables the user to obtain the distribution of gonad and PGCs by three GO terms. (B) Gene expression can be selected options, *P* value, statistical method, and tissues. The output format shows a list of expression-specific genes in each tissue. (C) Alternative splicing events are divided into two parts, case 1 and case 2 events. Users can see the output by graphical view with search options, case events, and tissues.

sequences using SeqClean (http://www.tigr.org/tdb/tgi/software). We filtered out ESTs of fewer than 300 bp from our dataset. The ESTs were clustered and assembled using TIGR Gene Indices Clustering Tools [20].

For the final step, unique sequences were aligned against chicken genome sequences by the University of California at Santa Cruz genome browser (http://genome/ucsc.edu/) using BLAT [21]. We retrieved 934 contigs and 3450







singletons from gonads and 156 contigs and 4937 singletons from PGCs. We downloaded testis and ovary ESTs from a TIGR GgGI cDNA library search (http://www.tigr.org/tdb/tgi/gggi/searching/xpress_search.html). Only pure tissue or organ data were collected. These consisted of 10,397 ESTs from two testis libraries (Cat. Nos. F6A and F6B) and 29,708 ESTs from five ovary libraries (Cat. Nos. AU5, AU6, AU7, AU9, and AV0). There was no need to perform a base-calling procedure with these sequences because trace data did not exist. The subsequent procedures, however, were the same as described above. Using the same assembling and clustering procedure applied to our dataset, we obtained 1087 contigs and 4519 singletons from the testis ESTs and 4090 contigs and 13,490 singletons from the ovary ESTs. The information on the EST datasets is listed in Table 1. Our EST datasets came from nonnormalized libraries and the others have no clear information about the normalization.

Gene ontology annotation and identification of putative novel transcripts

Two EST datasets, gonad and PGC, were subjected to the Gene Ontology (GO) database for annotation and categorization. GO annotation and identification were executed with a sequence similarity search against the tentative consensus sequences of GgGI release 10.0 (January 28, 2005) using BLASTN [22]. GO categorization used the GO flat files of the Gene Ontology Consortium (http://www.geneontology.org). Cutoff values were 95% identity, 60% coverage, and an E value <0.00001 for GO identification. If a query sequence did not satisfy the above conditions, it was placed in the putative novel transcript group with the BLAST no-hit sequences. For GO annotation of the novel transcripts, we performed a BLASTX search against the nonredundant NCBI protein database downloaded from ftp://ftp.ncbi.nih.gov/blast/db/(June 29, 2005).

Digital gene expression profiling

A significance test of the gene expression profiles between the pair of cDNA libraries was performed using Audic's [23] and Susko's statistical tests [24]. The program for Susko's test was downloaded from http://www.mathstat.dal.ca/~tsusko. These tests assume that EST frequencies are effective indicators of differences in gene expression and can therefore validate gene expression profiles from different tissues, organs, and cells. Bonferroni [25] and Benjamini corrections were applied to Audic's and Susko's tests because multiple-test problems occurred during this procedure. The number of ESTs assembled into a contig was considered to be the number of genetic sequences, and a singleton was considered a single genetic sequence. We tested whether different transcripts were expressed equally from the germ-cell-specific dataset. The stand-alone BLAST program, BLASTN, was used to cluster pairs of sequences that occurred in both libraries, with a minimum size of 200 bp and at least 95% identity.

Identification of germ-cell-specific alternative splicing events

We used a simple algorithm to compare EST queries based on known alternative splicing transcripts using known EST transcript–genomic sequence alignment. TIGR gene indices included alternative splicing variants identified from the given 529,525 ESTs and 25,660 expressed transcripts. Only the sequence with the best BLAST alignment of similarity against the known AS sequence was retained. If the best hit alignment was collinear, it was regarded as an alternative splicing event. If the best hit alignment contained an insertion or deletion of an EST query, it was assumed to be an alternative splicing form never seen before, at least in the current AS database; insertion and deletion events were deemed case 1 and case 2, respectively. The novel alternative splicing events were validated using BLAT sequence similarity searches against genomic sequences. This was followed by (1) checking the alignment between the case region and the genomic sequences using BLAT, (2) confirming the exon–intron boundary, (3) noting the intron phases, and (4) validating the translation of alternative splicing. For the first step, we aligned the case region of the putative alternative splicing isoform with the genomic sequence using the program BLAT. In a case 1 event, an EST fragment is inserted in a sequence of the current alternative splicing isoform, and this position should be a perfect match with the genomic region of the current alternative splicing isoform without a gap or mismatch in the alignment step. For a case 2 event, an EST fragment is deleted from an existing isoform sequence that is already aligned to the genome sequence. Therefore, this event should confirm whether the case 2 fragment matches the sequence of the existing alternative splicing isoform. Since genomic contamination occurs in some cDNA libraries, this stage is important to validate the putative alternative splicing isoforms. We allowed GT-AG, AT-AC, or GC-AG introns as splicing signals. The region (±10 bases) of the case position according to the alternative splicing events was checked against the exon–intron boundary within a genomic region, then against the initial AS dataset to remove any isoforms that did not meet prior conditions (i.e., incomplete sequences or unknown introns that may be due to sequencing errors). After checking the boundary with the genomic sequence, the case position was corrected under the assumption that the intron boundary sequences were located in the region of the case position. We confirmed whether the corrected case position was consistent with the EST and isoform. After checking the introns using the BLAT alignment, we determined the intron phases. This step excluded transcript isoforms with nonproductive protein sequences. Alternative splicing events in untranslated regions were also excluded from our analysis. The novel alternative splicing events detected in translated coding regions were thought to reflect surrounding intron phases. If the intron phases surrounding the case events are not symmetric, the translation of spliceosomal sequences might result in a frameshift.

Database construction and implementation

The database and Web interface were developed using MySQL, PHP, HTML, and Javascript. The standard server requires Apache 1.3.19, MySQL 4.0.21, and PHP 4.3.9 with GD library 2.0.33 and uses the FreeBSD 5.2 operating system. A summary of the data processing pipeline is shown in Fig. 2.

Acknowledgment

This work was supported by a grant (Code # 20050301034362) from BioGreen 21 Program, Rural Development Administration, Republic of Korea.

References

[1] E.M. Eddy, Male germ cell gene expression, Recent Prog. Horm. Res. 57 (2002) 103–128.
[2] D.J. McLean, P.J. Friel, D. Pouchnik, M.D. Griswold, Oligonucleotide microarray analysis of gene expression in follicle-stimulating hormone-treated rat Sertoli cells, Mol. Endocrinol. 16 (2002) 780–792.
[3] V. Reinke, I.S. Gil, S. Ward, K. Kazmer, Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans, Development 131 (2004) 311–323.
[4] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, I. Herskowitz, The transcriptional program of sporulation in budding yeast, Science 23 (1998) 699–705.
[5] V. Reinke, H.E. Smith, J. Nance, J. Wang, C. Van Doren, R. Begley, S.J. Jones, E.B. Davis, S. Scherer, S. Ward, S.K. Kim, A global profile of germline gene expression in C. elegans, Mol. Cell 6 (2000) 605–616.
[6] H. Sang, Transgenic chickens—methods and potential applications, Trends Biotechnol. 12 (1994) 415–420.
[7] M. Naito, M. Sakurai, T. Kuwana, Expression of exogenous DNA in the gonads of chimaeric chicken embryos produced by transfer of primordial germ cell transfected in vitro and subsequent fate of the introduced DNA, J. Reprod. Fertil. 113 (1998) 137–143.
[8] T.S. Park, J.Y. Han, Derivation and characterization of pluripotent embryonic germ cells in chicken, Mol. Reprod. 56 (2000) 475–482.
[9] Y. Minakuchi, M. Ito, Y. Kohara, SPI: a tool for incorporating gene expression data into a four-dimensional database of Caenorhabditis elegans embryogenesis, Bioinformatics 20 (2004) 1097–1109.

[10] M. Primig, et al., GermOnline: a new cross-species community annotation database on germ-line development and gametogenesis, Nat. Genet. 35 (2003) 291–292.
[11] J.W. Wallis, et al., A physical map of the chicken genome, Nature 432 (2004) 761–764.
[12] V.G. Tirunagaru, L. Sofer, J. Cui, J. Burnside, An expressed sequence tag database of T-cell-enriched activated chicken splenocytes: sequence analysis of 5251 clones, Genomics 66 (2000) 144–151.
[13] P.E. Boardman, et al., A comprehensive collection of chicken cDNAs, Curr. Biol. 12 (2002) 1965–1969.
[14] S.J. Hubbard, et al., Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags, Genome Res. 15 (2005) 174–183.
[15] J. Stebler, D. Spieler, K. Slanchev, K.A. Molyneaux, U. Richter, V. Cojocaru, V. Tarabykin, C. Wylie, M. Kessel, E. Raz, Primordial germ cell migration in the chick and mouse embryo: the role of the chemokine SDF-1/CXCL12, Dev. Biol. 272 (2004) 351–361.
[16] J.H. Shin, H. Kim, D. Lim, M. Jeon, B.K. Han, T.S. Park, J.K. Kim, H.S. Lillehoj, B.W. Cho, K.Y. Han, Analysis of chicken embryonic gonad expressed sequence tags, Anim. Genet. 37 (2006) 85–86.
[17] J.Y. Han et al., Gene expression profiling of chicken primordial germ cell ESTs. Submitted for publication.
[18] B. Ewing, P. Green, Base-calling of automated sequencer traces using Phred. II. Error probabilities, Genome Res. 8 (1998) 186–194.
[19] D. Gordon, C. Abajian, P. Green, Consed—A graphical tool for sequence finishing, Genome Res. 8 (1998) 195–202.
[20] G. Pertea, X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, J. Quackenbush, TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets, Bioinformatics 19 (2003) 651–652.
[21] W.J. Kent, BLAT—the BLAST-like alignment tool, Genome Res. 12 (2002) 656–664.
[22] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[23] S. Audic, J.M. Claverie, The significance of digital gene expression profiles, Genome Res. 7 (1997) 986–995.
[24] E. Susko, A.J. Roger, Estimating and comparing the rates of gene discovery and expressed sequence tag (EST) frequencies in EST surveys, Bioinformatics 20 (2004) 2279–2287.
[25] C.E. Bonferroni, Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8 (1936) 3–62.