A Large Database of Chicken Bursal ESTs as a Resource for the Analysis of Vertebrate Gene Function

Igor Abdrakhmanov,[1] Dmitry Lodygin,[1] Paul Geroth,[1] Hiroshi Arakawa,[1] Andy Law,[2] Jiri Plachy,[3] Berndt Korn,[4] and Jean-Marie Buerstedde[1,5]

[1]Heinrich-Pette-Institut for Experimental Virology and Immunology, Department of Cellular Immunology, 20251 Hamburg, Germany; [2]The Roslin Institute, Bioinformatics Group, Roslin, Midlothian EH25 9PS, UK; [3]Academy of Sciences of the Czech Republic, Institut of Molecular Genetics, 16637 Praha6, Czech Republic; [4]Deutsches Krebsforschunginstitut, Abteilung Molekulare Genomanalyse, Resourcen Zentrum des Deutschen Genomprojekts, D-69120 Heidelberg, Germany

Chicken B cells create their immunoglobulin repertoire within the Bursa of Fabricius by gene conversion. The high homologous recombination activity is shared by the bursal B-cell-derived DT40 cell line, which integrates transfected DNA constructs at high rates into its endogenous loci. Targeted integration in DT40 is used frequently to analyze the function of genes by gene disruption. In this paper, we describe a large database of >7000 expressed sequence tags (ESTs) from bursal lymphocytes that should be a valuable resource for the identification of gene disruption targets in DT40. ESTs of interest can be recognized easily by online BLAST or keyword searches. Because the database reflects the gene expression profile of bursal lymphocytes, it provides valuable hints as to which genes might be involved in B-cell–specific processes related to immunoglobulin repertoire formation, signal transduction, transcription, and apoptosis. This large collection of chicken ESTs will also be useful for gene expression studies and comparative gene mapping within the chicken genome project. Details of the bursal EST sequencing project and access to database search forms can be found on the DT40 web site (http://genetics.hpi.uni-hamburg.de/dt40.html).

[The sequence data described in this paper have been submitted to the GenBank data library under accession nos. AJ392050–AJ399459.]

The genomes of several model organisms have been sequenced completely (Tatusova et al. 1999) and the majority of all human genes are now represented in the public EST databases (Miller et al. 1997, 1999; Eckman et al. 1998). A challenge for the future will be the systematic analysis of the functions and interactions of the many novel genes. One approach, successfully applied to *Saccharomyces cerevisiae*, is the disruption of each open reading frame (ORF) by targeted integration of artificial gene constructs and the careful evaluation of gene disruption phenotypes (Winzeler et al. 1999).

Although a similar strategy is more difficult to pursue in vertebrates due to lower ratios of targeted to random integration, targeted gene disruption is used frequently to produce knockout mice. In the typical case, a heterozygous mutation is first introduced into murine embryonic stem (ES) cells by targeted integration of an artificial gene construct. The mutant cells contribute to the germline of chimeric mice after blastocyst injection, and further crossbreeding can give rise to homozygous mutant animals. This approach has great merit for the analysis of gene function in the whole animal (Smithies 1993), but it is technically demanding and expensive.

If the effects of gene inactivation can be studied in cell culture, the disruption of both gene copies in somatic cells is a valid alternative to the production of mutant mouse strains. A possible system for such studies is the chicken B-cell line, DT40, which integrates transfected DNA constructs at high ratios by homologous recombination (Buerstedde and Takeda 1991). DT40 has been used successfully for the genetic analysis of a variety of cell biology processes (Wang et al. 1996; Bezzubova et al. 1997; Fukagawa et al. 1999a; Kurosaki 1999; Takami et al. 1999). Gene disruption in DT40 is also possible, if the gene of interest is essential for cell proliferation, as conditional loss-of-function mutations can be generated using either cre-recombinase-mediated excision (Fukagawa et al. 1999b), tetracycline-regulated transcription (Wang et al. 1996), or tamoxifen-regulated protein transport (Fukagawa and Brown 1997).

Many features of the DT40 cell line are related to its derivation from an Avian Leukosis Virus-induced B-cell tumor. Chicken B cells develop a large repertoire

[5]Corresponding author.
E-MAIL buersted@genetics.hpi.uni-hamburg.de; FAX.
Article and publication are at www.genome.org/cgi/doi/10.1101/137900

10:2062–2069 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

of immunoglobulin genes by segmental gene conversion within the V-segment (Weill and Reynaud 1987). This occurs within the bursa of Fabricius, a special gut-associated lymphoid tissue of the bird located behind the cloaca. The strict compartmentalization of B cell development in the bursa of young chicks has been a powerful experimental advantage over other vertebrate species leading to numerous discoveries with regard to B cell characterization, repertoire development, and tumorigenesis.

To generate a useful resource for laboratories interested in early B cell development and/or the DT40 genetic system, we have created a large database of high-quality ESTs derived from purified bursal lymphocytes. This database can be analyzed as part of the DT40 web site either by the sequence homology search or by looking for keywords in annotations of the bursal ESTs.

RESULTS

Generation of Bursal EST Sequences

Bursal cells of the inbred CB strain were chosen as the source for the cDNA library because we believe that the transcript profile from primary cells is more interesting than that from the transformed DT40 cell line. In addition, the inbred genetic background facilitates the analysis of sequence variation due to the absence of allelic polymorphism. The cDNA was oligo-dT primed and directionally cloned into a plasmid to allow sequencing from the 5′ end of the cDNA inserts. About 55,000 primary clones were picked by a robot and transferred into microtiter plates as a permanent library stock. These ordered clones and filters representing the library can be requested from the German Genome Resource Center Primary Database in Berlin (http://www.rzpd.de/).

Plasmid templates from >10,000 arrayed clones were sequenced on an ABI377 sequencer using a primer that anneals in the polylinker upstream of the cDNA inserts. After rigorous control for vector sequence contamination, 7403 high-quality sequences of an average read length of 567-bp bases were incorporated into the bursal EST database. The proportion of full-length cDNA inserts was estimated by the analysis of ESTs derived from the following abundantly expressed genes (GenBank accession nos.: L00677, L08165, AJ004940, AF158370, L28704, X07265, Y00416, X92865, X62640, M24193) whose full-length coding sequence is known. About 15% of these sequences were full length and another 13% were nearly full length (<200 nucleotide cds sequence missing).

A Web Interface to the Bursal EST Database for BLAST and Keyword Searches

The bursal EST database provides a useful resource for the identification of chicken coding sequences for gene disruption and expression studies. However, tools are required to extract the information in a simple yet effective manner. To meet this need, we have provided several ways to query the data. First, an online BLAST homology search can be performed directly against the bursal EST sequences (http://genetics.hpi.uni-hamburg.de/cgi-bin/est-blast.cgi).

We also wanted to enable keyword searches of sequence annotations. This would permit the identification of whole classes of genes as defined by the keyword. Because it is difficult to create a new EST annotation system, we decided to base our annotations on information extracted from the BLAST search reports of our ESTs against the public databases. One of the advantages of this approach is that the annotations can be updated easily if new versions of the public databases are released. Each bursal EST was entered as a query sequence into the BLAST software of the GCG Wisconsin package (v. 10) and run under default conditions against the following databases: (1) our own bursal EST database (BLASTN), (2) the human EST database (BLASTN), (3) GenBank/EMBL (BLASTN), and (4) the PIR Protein database (BLASTX). All BLAST searches included filtering of low complexity sequences and regions homologous to the chicken repetitive sequence CR1 were masked prior to the search against our own bursal EST database. The following information was extracted for the top five hits of each BLAST report: (1) the name of the report reflecting the query sequence name and the searched database, (2) the rank of the hit, (3) the description of the homologous subject sequence, (4) the score, and (5) the expected probability. The extracted information was field separated and entered into a flat data file (blast_hit_file) with a single line corresponding to each hit in the BLAST search report. This blast_hit_file can be searched online for keywords using two forms. The more simple form (Fig. 1, http://genetics.hpi.uni-hamburg.de/estonline.html) allows only searches for a keyword or a list of keywords. This form also enables retrieval of BLAST reports and EST files. The more complex form (Fig. 2, http://genetics.hpi.uni-hamburg.de/estonlinecomplex.html) allows the user to enter two alternative search terms, as well as unwanted keywords and specifications of score and hit number limits. In addition, the searches can be restricted to particular databases. On submission of the form, formatted entries in the blast_hit_file containing hyperlinks to the original BLAST reports are returned.

The success of the keyword searches depends on how the homologous sequences were annotated in the public databases. However, this shortcoming is alleviated by redundancy, because each bursal EST is represented in the blast_hit_file by the descriptions of the five highest hits from three public databases. This in-

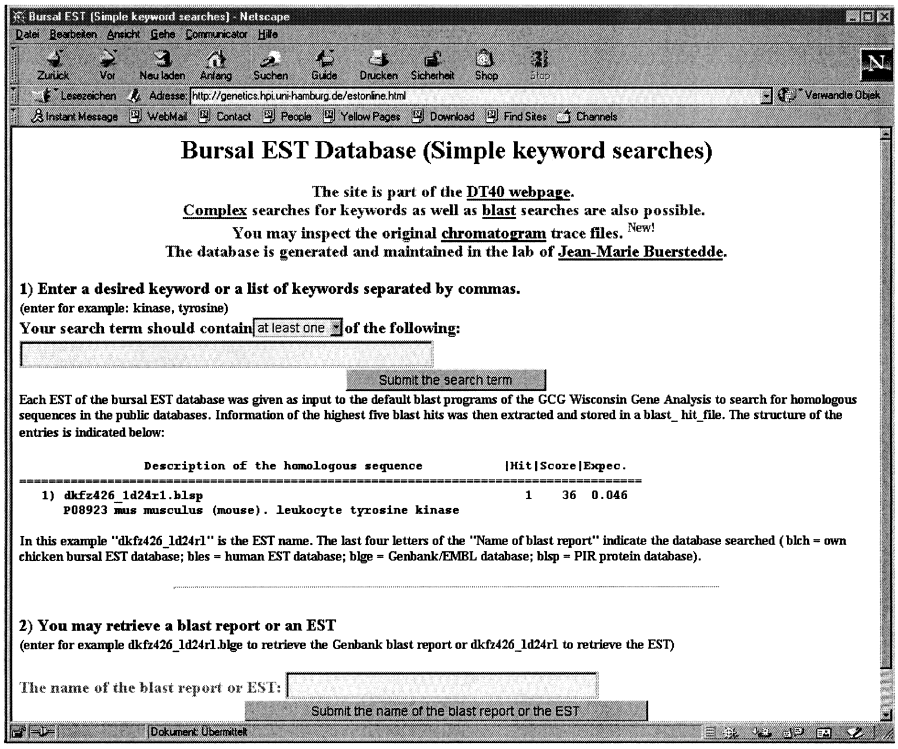

Figure 1 The form to search the BLAST hit file using a single keyword, or to retrieve the BLAST report or EST sequence.

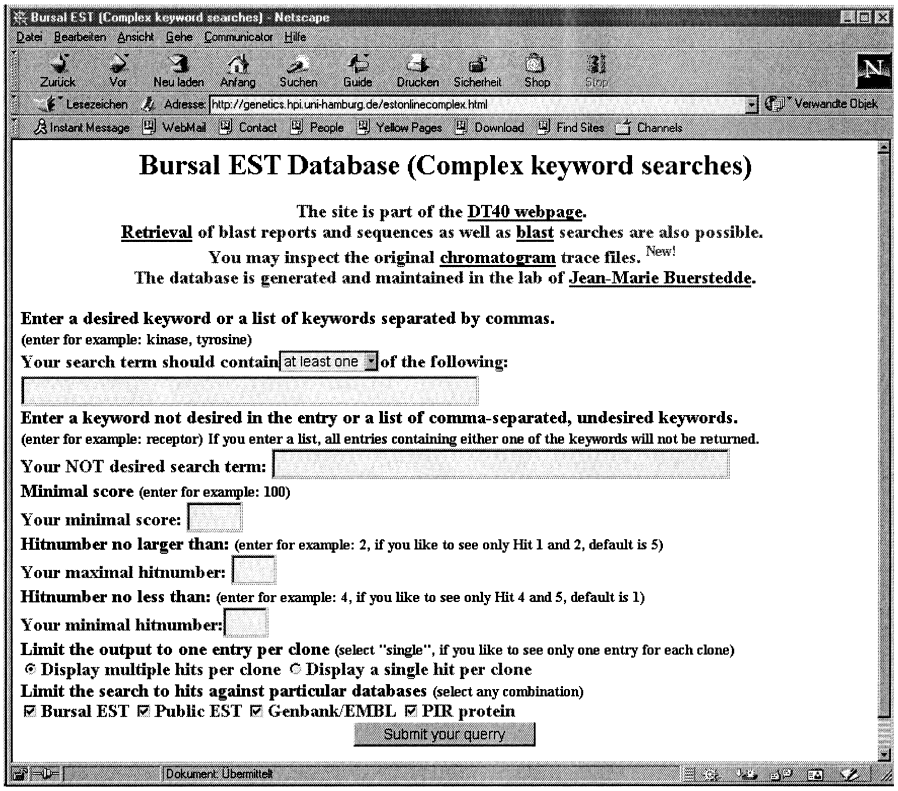

Figure 2 The form to search the BLAST hit file using complex search terms.

creases the chances that at least one of the entries matches the chosen keyword. In addition, it is not difficult to subsequently try a number of alternative keywords to enhance the success rate. Under certain conditions, the redundant presentation of multiple BLAST search hits for a single clone may be undesirable. If the "Display a single hit per clone" option is selected, only the first entry for each clone will be displayed.

Defining Sequence and Gene Clusters

The assumption was made that an overlap of 50 identical nucleotides for a query and a subject sequence in the bursal EST database is evidence for membership within the same sequence cluster. A program was written that grouped all sequences that fulfilled this requirement into the smallest possible number of different sequence clusters. The EST in the cluster that overlapped with the highest number of other ESTs was taken as the cluster-defining sequence. The clusters were then ordered according to descending size, and the following information was entered into a flat file (sequence_cluster_file) for each cluster-defining sequence and each public database: (1) the name of the BLAST report reflecting the name of the cluster-defining query sequence and the database searched, (2) the number of ESTs in the cluster, (3) the score, and (4) the description of the best-matched subject sequence in the public database.

To further normalize the database, we identified nonoverlapping clusters that seem to correspond to different parts of the same gene. We reasoned that the cluster defining ESTs should show a high BLAST score of >300 to the same refer-

ence sequence in the public databases. Clusters fulfilling these criteria were combined and information about the remaining gene clusters were entered into a flat file (gene_cluster_file) using the format described above for the sequence_cluster_file.

The sequence_cluster_files and gene_cluster_files can be searched online for keywords using a form (http://genetics.hpi.uni-hamburg.de/estonlineprivate.html) similar to the estonlinecomplex form. Whereas the sequence_cluster_file is searched by default, the gene_cluster_file is chosen for input, if the "Gene Cluster" option is selected.

The Gene Expression Profile of Bursal Lymphocytes Is Complex

The described estonlineprivate form is a convenient tool to analyze the complexity of the bursal EST database. By entering a first keyword like "dkfz," which is present in all entries, selecting the "Display a single hit per clone" and the "Sequence Cluster" or the "Gene Cluster" option, an ordered list of all sequence (5074) and gene (4999) clusters is returned. The most abundantly represented gene clusters in the bursal EST database (Table 1) originate from the chicken mitochondrial genome (113 sequences), the chicken beta actin message (58 sequences), and the chicken elongation factor 1 alpha (51 sequences). As expected for bursal cDNA library, both the immunoglobulin light (20 sequences) and heavy chain (20 sequences) are found among the 10 most frequent gene sequences in the database. With the exception of the mitochondrial genome gene cluster, none of the gene cluster presents >1% in the database and the 30 most frequent gene clusters represent <9% of all sequences. Thus, the gene expression profile of the bursal lymphocytes is unexpectedly heterogeneous (Bonaldo et al. 1996) given the fact that the cells have a well-defined cell-type and cell-stage specificity.

Homologs of the Bursal ESTs in the Public Databases

Bursal ESTs corresponding to the already sequenced chicken genes in the public databases can be distinguished by high BLAST scores, as the identity of the query and subject sequence is close to 100%. A total of 283 different gene clusters are returned if a BLAST score of ⩾300 is set as the limit and entries from the EMBL/GenBank or the PIR Protein databases containing ei-

Table 1. Representation of the Most Frequently Identified Genes in the Bursal EST Database

EST BLAST name	Number of clones	Score	Description of homologous subject sequence
dkfz426_25k2r1.blge	113	1477	Chicken mitochondrial genome
dkfz426_6b8rl.blge	51	698	Chicken elongation factor 1 alpha
dkfz426_3p7r1.blge	37	1100	Chicken beta-actin mRNA
dkfz426_26o9r1.bles	25	40	Human cDNA clone
dkfz426_20c2r1.blsp	23	41	Human hnRNP H mRNA
dkfz426_28h10r1.blge	21	311	Goose beta-actin mRNA
dkfz426_18n12r1.blge	20	1076	Chicken lg lambda light chain mRNA
dkfz426_22k18r1.blge	20	1550	Chicken mRNA for HS cognate 70kd protein
dkfz426_25n9r1.blge	20	1189	Chicken lg mu heavy chain mRNA
dkfz426_28p21.r1.blge	19	1322	Chicken DEAD-box RNA helicase mRNA
dkfz426_6e20r1.blge	19	1409	Chicken acidic ribosomal phosphoprotein P0
dkfz426_27o11r1.blge	18	803	Chicken ubiquitin 1 (Ubl) gene
dkfz426_13o10r1.blge	17	1178	Chicken nonhistone chromos. protein HMG-17
dkfz426_12c19r1.blge	15	1144	Chicken mRNA for B6.3 protein (Bu-1)
dkfz426_21g3r1.blge	15	1057	Human hnRNP B1 mRNA
dkfz426_14l24rl.bles	14	62	Human cDNA clone
dkfz426_1i11r1.bles	13	46	Human cDNA clone
dkfz426_18k14r1.bles	13	98	*Homo sapiens* o-glcnactransferase
dkfz426_18f10r1.blge	13	1350	Chicken mRNA for ribosomal protein L7a
dkfz426_17i21r1.blge	13	1513	Chicken MHC B complex protein (C12-3) mRNA
dkfz426_18c17r1.blge	13	105	Human prothymosin alpha mRNA
dkfz426_29p3r1.blge	13	892	Chicken mRNA for 90kDa heat shock protein
dkfz426_24m6r1.blge	12	597	*Homo sapiens* dead box, X isoform (DBX)
dkfz426_8o9r1.blge	12	1179	mRNA for chicken alpha-tubulin
dkfz426_23c24r1.blsp	11	377	Human ribosomal protein RS.40K
dkfz426_19b6r1.blge	11	1215	Chicken mRNA for ATF4
dkfz426_17a8r1.blge	11	194	Human chromosome 19
dkfz426_14o8r1.bles	11	40	Human chromosome 22
dkfz426_11d3r1.blge	11	545	Chicken nucleolar protein No38 (B23)
dkfz426_23k5r1.blsp	10	35	Mouse N-glycan alpha 2, 8-sialytransferase
dkfz426_11d20r1.blge	10	234	Human ribosomal protein L7 (RPL7) mRNA

ther the keywords "chicken" or "gallus" are requested by the estonlineprivate form.

The identification of new chicken homologs to already known genes in other species was the main purpose of the bursal EST database project. If we consider a BLAST score of 100 or higher as significant and exclude entries containing either of the keywords "chicken" or "gallus," 1682 of the total 4999 gene cluster are returned. Many of these gene clusters are probably new chicken orthologs of the homologous genes in other species. The gene clusters without high score matches in the public databases seem to represent poorly conserved cDNA sequences. Although some of the corresponding genes may be chicken- or avian-specific, it cannot be ruled out that significant homologies are present in parts of the transcripts that are not included in the EST.

A minority of the bursal ESTs without significant match in the public database may represent genes for which the vertebrate orthologs have not been isolated. Good candidates for these are ESTs that are most homologous to nonvertebrate genes in the public databases. If we enter, for example, 100 as minimal score and "Caenorhabditis elegans" as the search term in the estonlineprivate form and limit the search to the PIR Protein database, 120 gene cluster entries are returned. Although some of the listed ESTs have significant matches in the public EST or GenBank databases, others may provide first evidence for the presence of *C. elegans* orthologs in vertebrates.

Identification of Candidate Genes Involved in Immunoglobulin Repertoire Development

The bursal EST database greatly facilitates the identification of disruption target genes in DT40, but it is beyond the scope of this paper to deal with all potential applications in detail. We therefore decided to limit the discussion to the topic of homologous recombination. Chicken B cells generate their immunoglobulin gene repertoire to a large extent by gene conversion (Weill and Reynaud 1987) and it is tempting to speculate that homologous recombination factors are overexpressed in these cells. In addition, one of the authors recently showed that a high rate of somatic hypermutation in immunoglobulin genes coexists with gene conversion in bursal B cells (H. Arakawa, in prep.).

Genes of interest can be identified in the bursal EST database by entering the coding sequence of a known homolog into the BLAST search form (http://genetics.hpi.uni-hamburg.de/cgi-bin/est-blast.cgi). If the cDNA of the chicken *RAD54* gene (Bezzubova et al. 1997; GenBank accession no. U92461.gb_ov) is used as input, the BLAST search returns three entries (dkfz426_24c4r1.dat, dkfz426_12j22r1.dat, and dkfz426_8i24r1.dat) representing *RAD54* ESTs. Although these direct BLAST searches are convenient to find homologs of a single defined sequence, BLAST searches with more than a few query sequences are cumbersome.

The alternative strategy is to search the EST annotations for keywords. This has an advantage in that ESTs corresponding to whole classes of functionally related genes can be retrieved. The estonlinecomplex form is better suited for these searches than the estonlineprivate form due to higher redundancy of the annotations. Perhaps the most critical factor is the choice of the right search term. In our search for recombination factors, an obvious first choice is "recombination." However the term "repair" should also be considered as there is a broad overlap between recombination and DNA repair functions and "repair" is used frequently in gene annotations of the public databases. Other more specific keywords are "MSH" to identify homologs of the MSH mismatch repair genes in the database and "terminal deoxynucleotidyl transferase" to identify homologs of the TdT. Table 2 compares the results when either one of these keywords is entered as the search term and a score limit of 100 and the "Display a single hit per clone" option is selected in the estonlinecomplex form. As expected, the "repair" search is most inclusive returning most of the entries of the "recombination"and the "MSH" searches as well as six additional "repair" search-specific entries. The entries missed in the "repair" search, but returned in the either the "recombination" or "MSH" searches are ESTs homologous to the meiosis-specific *MSH4* and the *S. pombe rec14* genes as well as two entries (dkfz426_25l2r1 and dkfz426_29b23) due to accidental matching of the "MSH" keyword. Expression of the chicken *MSH4* gene in bursal lymphocytes is interesting, as it suggests that the *MSH4* might be involved in immunoglobulin gene conversion as well as meiotic recombination.

The "terminal deoxynucleotidyl transferase" search returned only a single EST (dkfz426_15m16r1). Inspection of the dkfz426_15m16.blch BLAST search report reveals however that there are two overlapping ESTs (dkfz426_20p2r1 and dkfz426_26c3r1) in the bursal EST database. Together, these sequences define a gene encoding a protein with significant homology to the beta polymerase as well as to the TdT. Given the structural homology to a gap-filling polymerase and an error-prone nucleotide transferase, we consider it a prime candidate for the polymerase implicated in immunoglobulin somatic hypermutation (Brenner and Milstein 1966). The gene might also be involved in immunoglobulin gene conversion. These hypothesis are now being tested by gene disruptions in DT40 and murine ES cells.

DISCUSSION

We present a large database of >4000 chicken EST se-

Table 2. Examples of BLAST Hit Searches Using Different Keywords

Results of "recombination" keyword search		
EST BLAST name	**Score**	**Description of homologous subject sequence**
dkfz426_8i24r1.blge	815	*Gallus gallus* putative recombination factor GdRAD54
dkfz426_12d20r1.blsp	131	*Schizosaccharomyces pombe* meiotic recombination protein rec14
dkfz426_12j22r1.blge	1119	*G. gallus* putative recombination factor GdRAD54
dkfz426_24c4r1.blge	896	*G. gallus* putative recombination factor GdRAD54
Results of "repair" keyword search		
EST BLAST name	**Score**	**Description of homologous subject sequence**
dkfz426_214r1.blsp	251	*G. gallus* dna-repair protein complementing xp-a cells homolog
dkfz426_7j15r1.blsp	311	*Homo sapiens* dna mismatch repair protein msh2
dkfz426_7m3r1.blsp	103	yeast dna repair protein snm1
dkfz426_8i24r1.bles	105	*H. sapiens* dna-repair protein rad54
dkfz426_12b3r1.blge	117	*Mus musculus* dsb repair protein
dkfz426_12j22r1.blsp	237	yeast rad54 dna repair protein
dkfz426_15d15r1.bles	111	*H. sapiens* mismatch repair protein msh2
dkfz426_15i15r1.bles	113	*H. sapiens* mismatch repair protein msh2
dkfz426_16a7r1.blsp	159	*M. musculus* dna-repair protein xrcc1
dkfz426_18h22r1.blsp	198	*H. sapiens* mismatch repair protein msh2
dkfz426_21b1r1.blge	291	*M. musculus* mismatch repair protein msh6
dkfz426_22c10r1.blsp	185	*G. gallus* dna-repair protein complement, xp-a
dkfz426_22h7r1.blsp	127	*M. musculus* dna-repair protein xrcc1
dkfz426_23g23r1.blge	293	*M. musculus* mismatch repair protein msh6
dkfz426_24c4r1.blsp	166	yeast rad54 dna repair protein
Results of "MSH" keyword search		
EST BLAST name	**Score**	**Description of homologous subject sequence**
dkfz426_1p7r1.blge	281	human MSH4
dkfz426_7j15r1.blsp	311	*H. sapiens* dna mismatch repair protein msh2
dkfz426_15d15r1.blge	111	*H. sapiens* mismatch repair protein MSH2
dkfz426_15i15r1.blge	113	*H. sapiens* mismatch repair protein MSH2
dkfz426_18h22r1.bles	198	*H. sapiens* mismatch repair protein MSH2
dkfz426_21b1r1.blge	361	*H. sapiens* mismatch repair protein MSH6
dkfz426_23g23r1.blge	363	*M. musculus* mismatch repair protein MSH6
dkfz426_25l2r1.blge	137	human serine hydroxymethyltransferase (HUMSHTA)
dkfz426_29b23r1.blge	200	human tyrosine phosphatase (HUMSHPTP1A)
Results of "terminal deoxinucleotidyl transferase" keyword search		
EST BLAST name	**Score**	**Description of homologous query sequence**
dkfz426_15m16r1.blsp	110	clawed frog terminal deoxynucleotidyltransferase

quence clusters from bursal lymphocytes. ESTs corresponding to genes of interest in the database can be identified online by entering query sequences for BLAST searches or by searching annotations of the ESTs for keywords.

The chicken genome project has been lagging behind similar efforts in other model organisms. This has complicated genetic work in DT40 because the cDNA sequences of most potential knockout targets had to be isolated by cross-hybridization or reverse PCR. The presented bursal EST database will improve this situation, because many candidates for gene disruption can now be identified by online homology searches. By reflecting the gene expression profile of bursal lymphocytes, the database also provides clues as to which genes might fulfill B-cell–specific functions. Although only partial cDNA sequences are present in the database, the cDNA library clones can be requested free of restrictions and the known sequence can be extended with internal sequence derived primers. Alternatively, missing 5′ and 3′ sequences can be obtained by RACE. Once the cDNA sequence of the target gene is known, the design of the knockout constructs is often straightforward for DT40, because chicken introns are relatively

small and the genomic locus can be amplified and cloned using cDNA derived primers (Bezzubova et al. 1997).

The bursal EST database will also contribute to genome mapping efforts in the chicken. Only a few gene-specific markers are currently available (Groenen et al. 1998) and a better coverage of the genome is highly desirable for the better characterization of quantitative trait loci (QTLs). The mapping of conserved sequence markers is particularly advantageous, as available comparative mapping data suggest a surprisingly high conservation between the chicken and the human genome (Burt et al. 1999). Many of the newly identified ESTs are putative orthologs of known human genes whose map positions are known. Mapping of these ESTs as well as others that are homologous to mapped human sequence tagged sites (STS) will reveal the chromosomal synteny regions in the chicken.

Finally, the ESTs can be used for gene expression studies. It is planned to combine the bursal EST clusters with other chicken gene sequences in the database to derive a collection of unique cDNA gene sequences. A filter or DNA chip containing these sequences will allow one to quantitate the expression profile of thousands of different genes in a single experiment (Bowtell 1999). It might, for example, be possible to pinpoint disease resistance genes by comparing gene expression patterns for chicken lines differing in their susceptibility to pathogens. In addition, the availability of a unique gene array will facilitate the analysis of DT40 mutants displaying changes in their overall protein expression (Takami et al. 1997). These studies can be extended to transcription factor mutants to identify the target genes regulated by these factors.

Although the bursal EST database most likely includes any gene abundantly expressed in bursal lymphocytes, low abundant genes are less likely presented. Further sequencing of randomly selected clones will return fewer and fewer new sequences, as the cDNA library was not normalized and a redundancy level of ~50% is already attained. An additional shortcoming of the current bursal cDNA library is the low frequency of full-length clones. We are now planning to construct a new normalized bursal cDNA library using the biotinylated cap trapper method (Carninci and Hayashizaki 1999). Sequences derived from this new library should complement the content of the present bursal EST database and advance our long-term goal to identify all genes expressed in bursal B cells.

METHODS

Construction of the cDNA Library

The bursas of a 2-wk-old chicken of the inbred CB strain were excised, and a cell suspension was prepared. Although the bursa consists mainly of B cells, Ficoll gradient centrifugation was used to remove contaminating epithelial and red blood cells. Cytoplasmatic polyA$^+$ RNA from the isolated lymphocytes that should consist of >90% B cells was purified using RNA easy and Oligotex mRNA purification systems (Qiagen, Germany) according to the manufacturer's instructions. The mRNA was reverse-transcribed using a poly-dT primer, size-selected for fragments >500 bp, ligated to a linker, and directionally cloned into the *Not*I/*Sal*I sites of the pSPORT1 plasmid (GIBCO BRL). The plasmids were transformed into DH10B cells. Restriction enzyme digests of 96 transformed clones indicate that the average cDNA insert size was ~1.3 kb.

Sequencing and Sequence Analysis

Cultures were grown from clones of the ordered library and plasmid DNA was isolated using the QIAprep 96 Turbo miniprep kit (Quiagen, Germany) according to the manufacturer's instructions. Sequencing reactions were performed using half of the standard volume of BigDye terminator sequencing kits (Perkin-Elmer) and a primer (5′-AAAGCTGGTACGCCTG CAG-3′) hybridizing upstream of the *Sal*I site in the pSPORT1 polylinker. Given the directional cloning of the cDNA, this produces reads into the 5′ end of the cDNA inserts. The sequencing reactions were analyzed on a 377 ABI Sequencer. The sequence trace files were imported into the Staden Package and processed using the pregap4 program with the ALF/ABI to SCF Conversion, Estimate Base Accuracies, Initialize Experiment Files, Augment Experiment Files, Uncalled Base Clip, Trace Quality Clip, Sequencing Vector Clip, Cloning Vector Clip, Screen For Unclipped Vector, and Interactive clipping options enabled. Interactive clipping was necessary to check the overall quality of the sequence chromatograms and to adjust the boundaries of the sequences to be retained. It was also used to remove stretches of polymerized linker sequence at the beginning of the cDNA insert. If the sequence ended in a polyA tail, all but the last four A nucleotides of the polyA tail were removed. About 15% of the clone inserts began with a long polyA tail stretch that precluded unambiguous reading of further sequence due to polymerase slippage errors. About 5% of the sequences displayed ambiguous signals suggesting that they represented either a mixture of clones or sequencing reactions. Approximately 7% of the clones contained no insert and another 3% of the sequences were discarded due to overall poor quality. The remaining "good sequences" as marked by the pregap4 program were extracted from the experiment files and imported into the Wisconsin gene analysis package. All individual ESTs were then combined into a single bursal EST database using the gcgtoblast function of the Wisconsin package.

Database Presentation and Web Site Design

The programming language Perl was used to write the three flat data files (blast_hit_file, sequence_cluster_file, and gene_cluster_file) and the cgi-scripts. All scripts are available upon request. HTML forms allowing specification of keywords, scores, and other useful parameters were placed on an Apache web server.

Sequence and Gene Clustering

To avoid false clusters due to overlaps within the chicken repetitive sequence CR1, a Java program was written that masks all stretches of 12 nucleotides in the ESTs that are identical to the CR1 consensus sequence (Stumph et al. 1984). Subsequently, the BLAST search results of each EST against

the EST database were searched for evidence of stretches of >50 nucleotide overlap with other ESTs.

To order cluster-defining sequences according to cluster size, a list was made containing the names of all bursal ESTs in descending order according to the number of significant matches with other sequences in the EST database. The subject sequences of the BLAST hits that fulfilled the 50-nucleotide overlap criterium constitute the initial sequence clusters of each listed sequence. Each sequence in the list was then checked to see whether any members of its cluster overlapped secondary ESTs which themselves were not already members of the cluster. These secondary ESTs were added as new members to the cluster. The names of the original cluster members that were still present in the list were subsequently removed from the list. The whole process was repeated until no new members could be added to the clusters as defined by the remaining sequences in the list. At this stage, the sequence list was ordered according to descending cluster size and used to write the entries of the sequence_cluster_file.

The sequence clusters were further normalized under the assumption that cluster-defining sequences that were most homologous to the same subject sequence in the public databases with a BLAST score of >300 corresponded to the same gene. Reading the sequence_cluster_file line by line, the accession number of the subject sequence was associated with the name of the cluster defining sequence if the BLAST score exceeded 300. At the same time, it was checked whether the accession number found was already associated with the name of a sequence recorded earlier in the sequence_cluster_file. If this was the case, the cluster sizes in the entries of the earlier sequence were augmented and the entries of the later sequence were deleted. The result of these operations was the gene_cluster_file.

ACKNOWLEDGMENTS

J.-M.B. acknowledges the kind advice and encouragement of Kay Foerger on his first journey into the world of programming and the UNIX operating system. The authors also thank Sabine Henze for technical help with the construction of the cDNA library, Florian Prill for the Java program to mask CR1 sequences, and Olga Bezzubova, Christine Laker, Klaus Harbers, Gustavo Martinez, and Frank Schnieders for helpful comments. This work was supported by the DFG grant Bu 631/2-1.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Bezzubova, O., Silbergleit, A., Takeda, S., and Buerstedde, J.M. 1997. Reduced X-ray resistance and homologous recombination frequencies in a RAD54−/− mutant of the chicken DT40 cell line. *Cell* **89:** 185–193.

Bonaldo, M.F., Lennon, G., and Soares, M.B. 1996. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res.* **6:** 791–806.

Bowtell, D.D. 1999. Options available—from start to finish—for obtaining expression data by microarray *Nat. Genet.* **21:** 25–32.

Brenner, S. and Milstein, C. 1966. Origin of antibody variation. *Nature* **211:** 242–243.

Buerstedde, J.M. and Takeda, S. 1991. Increased ratio of targeted to random integration after transfection of chicken B cell lines. *Cell* **67:** 179–188.

Burt, D.W., Bruley, C., Dunn, I.C., Jones, C.T., Ramage, A., Law, A.S., Morrice, D.R., Paton, I.R., Smith, J., Windsor, D., et al. 1999. The dynamics of chromosome evolution in birds and mammals. *Nature* **402:** 411–413.

Carninci, P. and Hayashizaki, Y. 1999. High-efficiency full-length cDNA cloning. *Methods Enzymol.* **303:** 19–44.

Eckman, B.A., Aaronson, J.S., Borkowski, J.A., Bailey, W.J., Elliston, K.O., Williamson, A.R., and Blevins, R.A. 1998. The Merck Gene Index browser: An extensible data integration system for gene finding, gene characterization and EST data mining. *Bioinformatics* **14:** 2–13.

Fukagawa, T. and Brown, W.R. 1997. Efficient conditional mutation of the vertebrate CENP-C gene. *Hum. Mol. Genet.* **6:** 2301–2308.

Fukagawa, T., Pendon, C., Morris, J., and Brown, W. 1999a. CENP-C use necessary but not sufficient to induce formation of a functional centromere. *EMBO J.* **18:** 4196–4209.

Fukagawa, T., Hayward, N., Yang, J., Azzalin, C., Griffin, D., Stewart, A.F., and Brown, W. 1999b. The chicken HRPT gene: A counter selectable marker for the DT40 cell line. *Nucleic Acids Res.* **27:** 1966–1969.

Groenen, M.A., Crooijmans, R.P., Veenendaal, A., Cheng, H.H., Siwek, M., and van der Poel, J.J. 1998. A comprehensive microsatellite linkage map of the chicken genome. *Genomics* **49:** 265–274.

Kurosaki, T. 1999. Genetic analysis of B cell antigen receptor signaling. *Annu. Rev. Immunol.* **17:** 555–592.

Miller, G., Fuchs, R., and Lai, E. 1997. IMAGE cDNA clones, UniGene clustering, and AceDB: An integrated resource for expressed sequence information. *Genome Res.* **7:** 1027–1032.

Miller, R.T., Christoffels, A.G., Gopalakrishnan, C., Burke, J., Ptitsyn, A.A., Broveak, T.R., and Hide, W.A. 1999. A comprehensive approach to clustering of expressed human gene sequence: The sequence tag alignment and consensus knowledge base. *Genome Res.* **9:** 1143–1155.

Smithies, O. 1993. Animal models of human genetic diseases. *Trends Genet.* **9:** 112–116.

Stumph, W.E., Hodgson, C.P., and Tsai, M.J. 1984. Genomic structure and possible retroviral origin of the chicken CR1 repetitive DNA sequence family. *Proc. Natl. Acad. Sci.* **81:** 6667–6671.

Takami, Y., Takeda, S., and Nakayama, T. 1997. An approximately half set of histone genes is enough for cell proliferation and a lack of several histone variants causes protein pattern changes in the DT40 chicken B cell line. *J. Mol. Biol.* **265:** 394–408.

Takami, Y., Kikuchi, H., and Nakayama, T. 1999. Chicken histone deacetylase-2 controls the amount of the IgM H-chain at the steps of both transcription of its gene and alternative processing of its pre-mRNA in the DT40 cell line. *J. Biol. Chem.* **274:** 23977–23990.

Tatusova, T.A. and Ostell, J.A. 1999. Complete genomes in WWW entrez: Data representation and analysis. *Bioinformatics* **15:** 536–543.

Wang, J., Takagaki, Y., and Manley, J.L. 1996. Targeted disruption of an essential vertebrate gene: ASF/SF2 is required for cell viability. *Genes & Dev.* **10:** 2588–2599.

Weill, J.C. and Reynaud, C.A. 1987. The chicken B cell compartment. *Science* **238:** 1094–1098.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285:** 901–906.

Received February 24, 2000; accepted in revised form October 26, 2000.