University of Alberta

Expressed Sequence Tags (EST) analysis, annotation and immune gene identification from a spleen cDNA library in the duck (*Anas platyrhynchos*)

by

Jianguo Xia ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Master of Science
in
Microbiology and Biotechnology

Department of Biological Sciences

Edmonton, Alberta
Fall 2006

Your file Votre référence
ISBN: 978-0-494-22410-6
Our file Notre référence
ISBN: 978-0-494-22410-6

ABSTRACT

To identify immune genes in an Expressed Sequence Tag (EST) project from a duck spleen cDNA library, a high-throughput EST analysis and annotation pipeline was built using *Perl* programming language. 2,900 raw EST sequences were subjected to quality checking, data cleaning, clustering and assembly, which generated 1,885 unique sequences. These sequences were annotated based on homology detection and domain identification. 1,248 sequences had putative identities assigned and 208 were immune relevant. Gene Ontology (GO) and pathway annotations were performed. 425 pairs of homologous sequences between duck and chicken were compared within different GO categories. Homologous genes had regions sharing 87-94 percent identity. However, immune genes showed lower overall sequence identities, with immune surface receptors the least conserved. These EST were submitted to *NCBI dbEST*. 2,725 sequences were accepted. Over 120,000 clones from this duck spleen cDNA library were macroarrayed onto 14 nylon membranes and screened for specific immune genes.

Acknowledgements

I wish to thank my advisor, Dr. Katharine E. Magor, for her patience, advice, and guidance throughout my graduate studies. In the past two and half years, I learned not only the requisite knowledge and skills in Immunology and Genetics, but also how to work independently, think critically, write meticulously, and present effectively. These benefits are going to have a profound influence on my future career. I would also like to thank Dr. David Wishart and Dr. Warren J. Gallin for their encouragements and assistances in my design and programming the EST analysis pipeline. I thank our lab technician, Deb Moon, for her help in my cDNA library construction, arraying, and screening. Lastly, I want to thank my wife, Yanqing Du, for her unwavering love, understanding, and support throughout the process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Terms in Biology and Immunology:

AMP	antimicrobial peptides
AMV	avian myeloblastosis virus
APC	antigen presenting cells
BCR	B cell receptor
bp	base pairs
CDS	coding sequence
CFU	colony forming unit
CMI	cell mediated immunity
CRD	carbohydrate recognition domain
Cremp	complement regulatory membrane protein
CSF	colony stimulating factor
CTL	cytotoxic T lymphocytes
DC-SIGN	dendritic cell–specific intercellular adhesion molecule-3 grabbing nonintegrin
EC	enzyme commission
EGF	epidermal growth factor
ER	endoplasmic reticulum
EST	expressed sequence tag
GPCR	G-protein coupled receptors

HBV	hepatitis B virus
HIV	human immunodeficiency virus
Ig	immunoglobulin
IFN	interferon
IL	interleukins
IRF	interferon regulatory factors
ICSBP	interferon-consensus sequence binding protein
ISRE	interferon stimulated response element
JAK	Janus kinas
LPS	ipopolysaccharide
LRR	leucine-rich repeat
MAC	membrane attack complex
MHC	major histocompatibility complex
MIP	macrophage inflammatory proteins
NF-κB	nuclear factor kappa B
NK	natural killer cell
PAMP	pathogen-associated molecular patterns
PFU	plaque forming units
PGN	peptidoglycan
PRR	pattern recognition receptors
RAG	recombination activation gene
RANTES	regulated upon activation, normal T-cell expressed, and presumably

	secreted
RSS	recombination signal sequences
RTK	receptor tyrosine kinase
SSH	suppression subtraction hybridization
STAT	signal transducer and activator of transcription
TAP	transporter associated with antigen processing
TCR	T cell receptor
TLR	Toll-like receptor
TM	transmembrane
TNF	tumor necrosis factor

Terms in Bioinformatics and Computer Sciences:

AMOS	a modular open source assembler
BLAST	basic local alignment search tool
BLOSUM	block substitution matrix
CAP3	contig assembly program (the third version)
CGI	common gateway interface
COG	clusters of orthologous groups of proteins
CPAN	comprehensive Perl archive network
DAG	directed acyclic graphs
GO	gene ontology
GUI	graphical user interface
HMM	hidden Markov model

HSP	high scoring pair
IRIS	Immunogenetic Related Information Source
KEGG	Kyoto encyclopedia of genes and genomes
MIPS	Munich information center for protein sequences
OOP	object oriented programming
OS	operating system
PAM	point accepted mutations
Perl	practical extraction and report language
PRINTS	protein fingerprints
PSSM	position specific scoring matrixes
SGE	Java Sun grid engine
SMART	simple modular architecture research tool
SQL	structured query language
TGICL	TGI clustering tool
TIGR	The Institute for Genome Research
Uniprot	universal protein resources

Chapter 1

INTRODUCTION

1.1 Expressed Sequence Tag (EST) project

Despite the growing number of completely sequenced genomes, many species of medical and agricultural importance have not yet been prioritized for genomic sequencing. Maybe some day in the future, a technology breakthrough in developing faster and cheaper DNA sequencing will allow every researcher to have his favorite model organism completely sequenced; then using new algorithms developed in the field of bioinformatics, it will be possible to produce an accurate list of all its genes with functional annotations from the sequence data. However, until that day comes, EST projects remain the primary choice as a quick and relatively inexpensive way for gene discovery in the un-sequenced organisms. EST analysis and annotation, a highly structured and involved process, remains the primary means for carrying out such tasks.

EST projects refer to the systematic sequencing of cDNA clones from carefully constructed cDNA libraries usually in a high-throughput manner. They play a fundamental role in gene discovery from un-sequenced organisms. In addition, EST data are often used as landmarks in genome assembly, gene prediction, and gene variation studies. EST technology has become a powerful tool for genetic research since its introduction (Adams *et al.*, 1991).

1.1.1 Technology overview

The basic strategy in EST project is to select cDNA clones at random and then perform single pass automated sequencing from the clone ends. In most cases, there is no initial attempt to characterize the whole insert. Instead, ESTs are identified using only the partial sequence data by comparing them to other known sequences. Most ESTs are around 500~800 base pairs (bp) long based on the current sequencing technology.

Since cDNA can be directionally cloned into vectors, EST sequencing can be performed from one or both ends of the inserts. The 5' end sequence is derived from within the coding region and thus provides information about the protein encoded by the expressed gene. The 3' EST usually contains sequence from the untranslated region (3'-UTR) of the transcript and thus serves as a unique identifier for the gene, which allows differentiation between similar members in a gene family.

The features of EST technology make it possible to sequence a large number of cDNAs from any particular tissue type within a relatively short time period. As a result, EST projects are usually conducted in a large-scale and high-throughput manner in order to obtain a comprehensive sampling and more importantly, to reduce the cost per sequence. However, one inherent problem associated with this approach is its vulnerability to the "law of diminishing returns". Although very efficient at the beginning, the rate of gene discovery will actually diminish with continuing efforts to sequence more EST data. This is because transcripts are not present with equal abundance in any tissue type. Highly expressed genes will be represented multiple times, while rarely expressed genes will likely be missed during random sequencing. This problem can be partially ameliorated by either using new cDNA libraries prepared from different sources, or constructing pooled, normalized or subtracted libraries depleted of clones already sequenced (Bonaldo *et al.*, 1996).

1.1.2 Basic steps in EST analysis and resources

EST data is like a genetic gold mine. However, a highly disciplined approach is required to make it maximally productive. There are three basic steps in EST analysis: 1) base calling and data cleaning, 2) clustering and assembly, and 3) putative identity assignment. For high throughput functional analysis, Gene Ontology annotation and KEGG pathway assignment are also frequently performed. Other customized analyses such as evolutionary studies can also be done based on the biological questions of interest.

1.1.2.1 Base calling and sequence cleaning

Two stages are involved in this step. The first stage is to improve the base-calling accuracy and to truncate the low-quality segments. The second stage is to clean sequences by removing vector/adaptor sequences, contaminants, and low complexity sequences. Ribosomal and mitochondrial sequences can also be removed at this stage.

The raw EST data generated from automated DNA sequencers is a four-color chromatogram (trace file), as well as a text file of the sequence representing the machine's best interpretation. In general, the error rate for the first few hundred nucleotides is very low, but then errors occur more and more frequently due to poor gel resolutions or weaker signals. There are two popular base-calling programs available to help improve the base-calling accuracy based on specially developed probability models: the open source *phred* and the *KB^{TM} Basecaller* shipped with the ABI 3730 sequencing machine from *Applied Biosystems* (http://www.appliedbiosystems.com). *Phred* is probably the most popular base-calling program (Ewing and Green, 1998; Ewing *et al.*, 1998). It retrieves the sequence information from the chromatogram and assigns every base a quality score or phd value by multiplying the $\log_{10}$ of the error probability by -10. For example, the most commonly used criterion of "high quality bases" is phd value > 20, which means the error probability is 0.01 or the base calling accuracy is 99%. After calculating quality scores, regions with high error rate can then be truncated based on the parameter (default error rate 0.05).

After running *phred*, the sequences derived from trace files need to be cleaned of vector/adaptor sequences, sequences from *E. coli*, phage, and other common vector contaminants, short sequences, low complexity sequences such as polyA/T tails and repetitive DNA. Two popular software tools are available for this purpose. The *TIGR seqclean* (http://www.tigr.org/tdb/tgi/software) is a *Perl* script package used for automatic trimming of low quality and low complexity sequences. It also validates ESTs by screening for various contaminants based on user supplied contaminants databases. The other program is *Lucy* which offers similar functions (Chou and Holmes, 2001).

1.1.2.2 EST clustering and assembly

The primary purpose of EST analysis is to identify genes of interest from a large body of sequences. However, randomly sequenced EST data is inherently redundant in which highly expressed genes are represented multiple times. This not only distracts researchers' attention from identifying new genes but also causes unnecessary computational burden for subsequent EST analyses. The EST clustering and assembly procedure attempts to address this issue. The clustering algorithm tries to identify transcripts derived from the same gene and then puts them into the same clusters. The assembly algorithm will subsequently merge the sequences in each cluster to produce contigs as the theoretical sequence of the underlying transcript. This process usually produces longer and better quality sequences and also significantly reduces the size of the dataset for downstream analyses. The clusters formed can help detect splice variants by comparing sequence members within the same cluster.

An EST clustering program works by putting closely related transcripts into the same group. There are two approaches for EST clustering, supervised and un-supervised. In supervised clustering (such as the NCBI *UniGene* www.ncbi.nlm.nih.gov/UniGene), ESTs are classified with respect to a reference sequence (usually a full length mRNA or predicted CDS from genomic sequences); while in unsupervised clustering, there is no prior-knowledge incorporated, and only sequences that overlap with each other to a pre-defined level will be included within the same cluster. The program usually first performs a pair-wise comparison for all the sequences, assigns a similarity score for each of them, and then puts them into different clusters based on a cut-off value. The assembly algorithm works by first performing a multiple alignment for each cluster and then calculating a tentative consensus sequence based on a "majority rules" for each set of aligned bases.

For large scale assembly of shot-gun sequence data (> 500 sequences), three first-generation programs are still widely used: *phrap* (http://www.phrap.org), *TIGR Assembler* (Sutton *et al.*, 1995), and *CAP3* (Huang and Madan, 1999). *Phrap* was the first available, but the latter two are more accurate. More recent second-generation assembly programs such as *Celera Assembler* (Huson *et al.*, 2001) and *Arachne* (Batzoglou *et al.*, 2002) were mainly developed for super-scale (~ 1 million sequences) genome sequence assembly. The well-known *AMOS* (A Modular Open Source Assembler http://amos.sourceforge.net) project was launched in order to bring all these innovations for sequence assembly under the open source framework. However, all these programs were primarily developed for the assembly of shotgun genomic sequences. They are not ideal for assembling EST sequences.

EST assembly differs from genome assembly in several ways. For instance, genomes often contain large amounts of repetitive sequences, especially in the non-coding region, while EST sequences usually do not have such repeats. In addition, EST data also contain features like alternative splicing, RNA editing and other post-transcriptional modifications. All these differences make these genome assembly programs less applicable to EST assembly. The *TGI Clustering Tool (tgicl)* is a program specially tailored for large-scale EST assembly (Pertea *et al.*, 2003). It internally uses the NCBI program called *MEGABLAST* for calculating the sequence similarities to produce clusters, and then uses the *CAP3* program for subsequent assembly. The assembly file in ACE format generated by *CAP3* can be inspected using a graphical program *clview* also available at the TIGR (http://www.tigr.org/tdb/tgi/software).

1.1.3 Sequence annotation strategies

1.1.3.1 Homology detection

Homology detection is usually the primary means for sequence annotation. Functional information can be transferred if the corresponding homologs can be identified using a database search. Homologs are defined as sequences derived from a common ancestor. There are two types of homology: orthology and paralogy. Orthologous genes are homologous genes separated by speciation events, while paralogous genes are homologous genes separated by gene duplication events within the same species. In theory, orthologs will typically retain the same or similar functions as their ancestors, while the duplicates (paralogs) are free to mutate to acquire new functions. However, depends on the level of divergence and annotation, it is difficult to distinguish orthologs from paralogs.

In homology detection, the choice of a sequence database is very important. For homology searches, protein databases are usually chosen if the organism under study does not have a closely related species whose genome was completely sequenced and well annotated. Using protein sequence is much more sensitive at detecting distant homologies than using DNA sequence for two important reasons: firstly, in DNA sequences, there is extra noise from the degenerate third position in each codon; secondly, protein sequences are composed of 20 different amino acids, each with distinct structural and chemical characteristics. The greater variability in the protein alphabet over the DNA alphabet allows for a much finer assessment of the degree of similarity (i.e. the BLOSUM (Henikoff and Henikoff, 1992)) than could be done with just four nucleotides in DNA sequences.

Among the public protein sequence databases, *Swissprot* offers some of the best quality data because its contents are manually annotated and maintained by experts. However, using only *Swissprot* will miss a lot of potentially important genes due to its smaller size (222,289 entries on 30-May-2006). The *TrEMBL* database is a computer-aided supplement of *Swissprot*. It contains all the predicted proteins derived from DNA sequences (2,948,323 entries on 30-May-2006) that have not yet been integrated into *Swissprot* (O'Donovan *et al.*, 2002). Finally, the *Uniprot* database (Universal Protein Resource) is the world's most comprehensive protein sequence database (UniRef100 contains 3,511,676 entries on 30-May-2006). It is composed of sequences from *Swissprot, TrEMBL, PIR* and other sources (Apweiler *et al.*, 2004). In addition, a significant proportion of the sequences in *Uniprot* have been assigned GO terms (Camon *et al.*, 2004), and they also have EC (Enzyme Commission) numbers if the gene products have enzymatic functions. The later two sources of information constitute the basis for GO annotation and KEGG pathway assignment.

The search engine is another important issue to be considered. The three major players for large-scale sequence comparison are *BLAST* (Altschul *et al.*, 1997), *FASTA* (Pearson, 1990), and *HMMer* (Eddy, 1998). All of them are open source and freely available. There is also a commercial software *PatternHunter*^{TM} that claims to be faster and better than any of these tools (Ma *et al.*, 2002). Among the three open source programs, *BLAST* is the most widely used as it offers both speed and sensitivity. The other two are more sensitive but much slower (personal experience).

Database searches usually retrieve more than one hit. In some cases, multiple highly similar sequences are identified from the same species (gene families). It is difficult and sometimes arbitrary to select one hit as the ortholog to the query sequence. The top hit is not always the correct annotation or the most informative one. Therefore, it is often necessary to check and compare hits from different species to make a final judgment. In *BLAST* searching, many tutorials suggest that E values smaller than 10^{-4} are convincing and worth checking as a rule of thumb.

1.1.3.2 Motif identification

Using *BLAST* to search against protein databases is usually the primary means to identify query sequence homologs. However, many queries will still fail to yield any hits. Two possible reasons can account for this. The first one is that the query sequences have diverged beyond the point where any sequence homologs can be recognized by the search algorithm. The second one is that the query sequences represent novel genes and as a result, the databases do not contain their homologs. In either case, using a different search engine or changing the underlying databases will complement the previous search and increase the chances to obtain new information for the query sequences. One common approach is to conduct motif database searches.

Protein domains and protein sequence motifs represent smaller functional or structural subunits. They are often identified through multiple sequence or structure alignments. They are usually more conserved than the sequences themselves. If a domain within a protein is associated with certain functions, the harboring protein is much more likely to have such functions. A protein containing multiple domains might have no homologs that match its full length. However, if each domain or motif identified is known to be associated with some function, it is possible to hypothesize the general function for this protein.

There are as many domain/motif databases as there are sequence databases. Among the more popular ones are the *PFAM, COG, PROSITE*, *SMART* and *PRINTS* databases (Schultz *et al.*, 1998; Attwood, 2002; Tatusov *et al.,* 2003; Bateman *et al.*, 2004; Hulo *et al.*, 2004). Each database has its own format and associated scanning tools for domain/motif discovery. Among them, *HMMer* and *RPS-BLAST* are the most popular (Altschul *et al.*, 1997; Eddy, 1998). *HMMer* searches a query sequence against a database of profile hidden Markov models (HMM), while *RPS-BLAST* searches a query sequence against a database of position specific scoring matrixes (PSSM). The former is more sensitive at detecting closely related domains, but the later works for a broader range of evolutionary distances and is much faster.

Interproscan is so far the most comprehensive package for domain/motif identification. It packages and streamlines all the above mentioned databases to facilitate domain identification (Quevillon *et al.*, 2005).

1.1.4 Gene Ontology and KEGG pathway information

1.1.4.1 Gene Ontology (GO)

One of the major obstacles in conducting high throughput analysis in bioinformatics is the lack of a standard language in describing biological information. For instance, the functional annotations of a gene product are usually in the form of free text, which are quite meaningful for humans but very difficult for computer interpretation. Two major attempts to address this issue of a common language are Gene Ontology Annotation (GOA) (Ashburner *et al.*, 2000) and the Functional catalogue (Funcat) (Ruepp *et al.*, 2004) which is maintained at the Munich Information Center for Protein Sequences (MIPS) (Mewes *et al.*, 2002). Both schemas use a controlled vocabulary to describe the function, location or role of gene products in a species-independent way. It is possible to do mapping between these two vocabularies. The major difference is that the Funcat schema uses a strictly hierarchically structure, while the GO schema uses directed acyclic graphs (DAG), allowing more than one parent per child. For example, the "transmembrane receptor protein-tyrosine kinase" is a child of both "transmembrane receptor" and of "protein tyrosine kinase". Thus, through DAG structure, the GO annotated gene products can be analyzed from different perspectives. The schema offers this flexibility but brings with it the redundancy - members in one category might also be members of another group and will be analyzed more than once. Currently, the GO annotation system is more widely adopted and better supported.

GO tries to describe how a gene product behaves in a cellular context using three categories: 1) its molecular function, 2) the broad biological process it participates in, and 3) the cellular location it acts in. “GO slim” is a “slimmed-down” version of GO vocabulary. It contains a subset of GO terms providing a high-level overview of different categories without the detailed descriptions. Both GO slim and MIPS Funcat are widely used to present synoptic summaries for large sequence data sets (i.e. microarray probe sets, genome or EST data).

1.1.4.2 KEGG pathway

Sequence annotation is only the starting point to understand the functions of individual genes. GO annotation presents a high level static picture in the form of short synoptic phrases describing biological events. Living organisms are much more complex and dynamic. Metabolism and immune responses are only possible within a biological system. Moreover, metabolites, minerals, vitamins, and other small chemical compounds are also essential components other than just DNA and proteins. In this case, the Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling/metabolism pathways offer information from a new perspective. While GO tries to describe all genes and their products, the KEGG database focuses on enzymes, pathways, and interactions.

The KEGG database contains most of the known metabolic pathways and many of the known gene regulatory pathways in the form of 235 reference graphical diagrams (Kanehisa *et al.*, 2004). Given a list of enzymes (EC numbers) that are found in the gene catalog of an organism, the KEGG database can automatically compute and draw organism-specific pathways by mapping the enzymes to the diagrams in the database. For whole genome annotation, this offers another validation of the functional assignment by checking the pathway maps: the existence of a missing element indicates either the gene annotation is wrong or there exists an alternative pathway that uses different enzymes.

1.1.5 EST projects for immune gene discovery in other species

EST projects offer a way to quickly survey a large pool of gene transcripts. From the sequence tag data, many genes of interest can be identified. The immune-relevant genes, for example, are usually on the shortlist for their medical and agricultural importance. In fact, many EST projects as discussed below were initiated primarily for identifying immune genes.

In the shrimp EST project, a total of 2,045 clones were randomly sequenced from four cDNA libraries (two hemolymph and two hepatopancreas libraries), and 44 functional genes were identified to be immune relevant (Gross *et al.*, 2001). In the mosquito EST project, 5,925 clones were sequenced from a subtracted cDNA library (performed between normal and immune challenged hemocyte-like cell line), and 38 genes were identified to be potentially involved in immunity (Dimopoulos *et al.*, 2000). In the Tsetse fly EST project, 21,427 sequences were produced from a self-subtracted midgut cDNA library, 78 homologs were identified with known or putative immune functions (Lehane *et al.*, 2003). In a chicken EST project, 5,251 EST were sequenced from a T-cell enriched activated splenocyte cDNA library, and 80 immune-related genes were identified (Tirunagaru *et al.*, 2000). In addition, many immune genes were identified from EST projects developed for other purposes. For example, an *in silico* survey on 450, 000 publicly available chicken EST dataset generated from 64 cDNA libraries (derived from all tissue sources) identified 185 immune related sequences, 95 of which had not been reported before (Smith *et al.*, 2004).

1.2 Overview of the immune system

The main theme of my research project was to identify duck genes involved in immunity. An overview of the basic concepts, key players and important processes of the immune system is therefore warranted. The majority of this information is based on studies from human and mouse models, assuming the duck immune system is similar. This introduction will only focus on parts of the immune system that are relevant to the genes identified in our EST project.

1.2.1 Innate and adaptive immunity

No species could survive without some sort of immune system. In vertebrates, the immune system has evolved into an extremely complicated network with a wide variety of different molecules specialized for recognition, activation and regulation. The vertebrate immune system is divided into two subclasses: innate immunity and adaptive immunity. Innate immunity is phylogenetically conserved and appeared early during the evolution of multicellular organisms. It constitutes the most fundamental and essential part of the immune system. On the other hand, adaptive immunity is believed to appear only after the divergence of jawed fish, reviewed by (Hoffmann *et al.*, 1999). This concept has been challenged by recent studies suggesting a new form of adaptive immune system discovered in sea lampreys (Pancer *et al.*, 2004; Alder *et al.*, 2005).

1.2.1.1 Innate immunity

Innate immunity is the first-line of host defense, responsible for early detection and containment of pathogens. It is triggered immediately after detecting pathogen invasion. Accumulating evidence indicates that activation of innate immunity is a prerequisite to the induction of adaptive immunity, reviewed by (Banchereau and Steinman, 1998).

Pathogens are recognized through their evolutionarily conserved structures termed pathogen-associated molecular patterns (PAMP), such as peptidoglycan, lipopolysaccharide (LPS), flagella that are not present on the hosts, reviewed by (Medzhitov and Janeway, 1997). The receptors that recognize these PAMPs are called pattern recognition receptors (PRR). The most prominent PRR is the Toll-like receptor family. Other components of innate immunity include C-type lectins, complement proteins, antimicrobial peptides, cytokines, and molecules involved in inflammation and apoptosis.

1.2.1.2 Adaptive immunity

The major distinction between innate and adaptive immunity is the level of specificity in antigen recognition. The highly specific nature of adaptive immunity is enabled by the random generation of a large pool of recognition receptors (i.e. antibodies, T cell receptors) coupled with the ability to select and propagate the effective receptors and their originating cells (adaptive). While these receptors are highly specific, it takes several days for the effector cells undergoing proliferation and differentiation to take effect.

There are two arms of adaptive immunity: humoral immunity and cell-mediated immunity, organized around two classes of specialized immune cells, B cells and T cells, respectively. Humoral immunity is mediated by antibodies secreted from B cells. Antibodies can recognize and bind to the surface of invading microbes in the blood, labeling them for destruction. The cell-mediated immunity involves the activation of cytotoxic T cells, NK cells and macrophages. They are able to recognize and kill virus-infected cells or transformed cells (i.e. tumor cells).

1.2.2 Major immune cell types

In vertebrates, the immune responses are carried out by several specialized cell types derived from two distinct cell lineages. The myeloid lineage includes monocytes, macrophages, dendritic cells and neutrophils. The lymphoid lineage includes T lymphocytes, B-lymphocytes and natural killer (NK) cells. T and B-cells constitute the core of adaptive immunity with their antigen-specific receptors generated mainly through somatic gene rearrangement. Neutrophils, macrophages and NK cells are mainly involved in innate immunity using a variety of pattern recognition receptors encoded in the germline, reviewed by (Cooper *et al.*, 2001). Dendritic cells (DC), with their capacity for capturing, processing and presenting antigens, bridge innate immunity with adaptive immunity. The focus over the next few pages will be given to T cells, B cells, DCs and NK cells.

1.2.2.1 T lymphocytes

There are two subpopulations of T cells, CD8+ T cells and CD4+ T cells respectively, as identified by their surface markers. CD8+ T cells recognize antigens presented through class I major histocompatibility complex (MHC) molecules, and the CD4+ T cells recognize antigens presented via class II MHC molecules. The activation of T cells needs both the contact signal from the recognition of the MHC-antigen complex and an additional co-stimulatory signal only available through professional antigen presenting cell (APC) such as B cells, macrophages and DCs.

The CD8+ T cells develop into cytotoxic T cells (CTL), a major player in cell-mediated immunity (CMI). They recognize and lyse virus-infected or transformed cells. Naive CD4+ T helper cells (Th0) will further differentiate into either Th1 or Th2 subtypes, based on the cytokines profiles they secrete. Th1 cells produce interferon (IFN)-gamma, and interleukin (IL)-2 and tumor necrosis factor (TNF)-beta, which enhance the killing efficiency of CTL and macrophages. Th2 cells produce interleukin (IL)-4, 10 and 13, which are essential for B cell activation and differentiation, reviewed by (O'Garra and Arai, 2000).

1.2.2.2 B lymphocytes

In mammals, B cells are derived from the bone marrow. Naive B cells leave the bone marrow and migrate to the spleen and peripheral lymph nodes. B cells recognize antigens through their surface antibodies which is not MHC restricted. The activation of B cells is usually the result of reciprocal stimulation between Th2 and B cells that recognize the same antigen.

Some “super-antigens” can activate B cells by themselves. In most cases, however, the B cell activation requires the help from Th2 cells. Ligand binding will lead to the cross-linking of the B cell receptors, which provides the first required signal. The B cell will then ingest and present the antigens to Th2 cells via its MHC class II molecules, forming a conjugate with the Th2 cell. The Th2 cell reactivity toward the same antigen will activate itself and subsequently releases cytokines essential for B cell activation. The contact and mutual signaling between B and Th2 cells will lead to B cell activation, clonal expansion and differentiation into antibody-secreting plasma cells, reviewed by (Parker, 1993).

1.2.2.3 Dendritic cells (DC)

Perhaps one of the most important conceptual breakthroughs in the field of immunology during the past decade is the recognition and appreciation of the roles of DCs during immune responses. It is now widely accepted that DCs function as bridges between innate and adaptive immunity, reviewed by (Banchereau and Steinman, 1998).

Immature DCs are distributed in peripheral tissues, where they constantly monitor the surrounding environment through various PRRs expressed on their surface. Once they detect PAMPs or sense “danger signals” (i.e. necrotic cell debris or substances released from stressed cells like interferon-alpha, heat shock proteins) (Matzinger, 1994), a maturation process will begin, in which DCs gradually transform into potent antigen presenting cells. At the same time, they begin to migrate toward nearby lymph nodes where they meet and present the captured antigens to potential antigen-specific T cells to initiate adaptive immunity (Granucci *et al.*, 2001; Lindstedt *et al.*, 2002).

1.2.2.4 Natural killer (NK) cells

NK cells constitute an important arm of innate immunity. They serve as the first level of surveillance against cancer and viral infections. Unlike T or B cells, NK cells recognize the target cells by scanning the expression level of MHC class I molecules on cell surfaces. NK cells use two groups of receptors: inhibitory receptors prevent NK cell activation upon encountering normal expression levels of MHC class I, whereas, activating receptors recognize diverse ligands, reviewed by (Lanier, 2005). Since both virus-infected cells and cancer cells often have reduced expression of MHC class I, they become susceptible to killing by NK cells. The lectin families are involved in NK cell mediated recognition of MHC class I molecules, reviewed by (Lopez-Botet *et al.*, 1997).

1.2.3 Important molecules in innate immunity

1.2.3.1 C-type lectins

C-type lectins represent a large family of calcium-dependent carbohydrate-binding proteins. The common feature of this family is the highly conserved carbohydrate recognition domain (CRD) essential for ligand binding. C-type lectin family members are found to be involved in inflammation, viral infection, tumor immunity and many other immune functions, reviewed by (van Vliet *et al.*, 2005).

Members of the C-type lectin family are further divided into different subfamilies based on their functions or unique localizations. Major subfamilies include the endocytic receptor family, the collectin family, the selectin family and the lymphocyte lectin family. The endocytic receptor family members mediate endocytosis of bound ligands. The macrophage mannose receptor is a well-studied endocytic receptor. It recognizes Gram-positive and Gram-negative bacteria, yeasts, parasites and mycobacteria (Engering *et al.*, 1997). The collectin family members contain a collagen-like domain. They usually assemble into large oligomeric complexes to carry out their functions. Collectins bind preferentially to monosaccharide structures and/or lipid moieties on the surface of microorganisms, reviewed by (van de Wetering *et al.*, 2004). Selectin family members are mainly expressed in vascular endothelia and in circulating leukocytes. They are involved in selective cell adhesion during inflammation. DC-SIGN (CD209) molecule is an important C-type lectin member that functions both as an adhesion receptor and as an endocytic pathogen-recognition receptor. It is expressed primarily on DCs, participating in DC migration and T cell activation. DC-SIGN has been shown to serve as the portal for human immunodeficiency virus (HIV) infection (Geijtenbeek *et al.*, 2000). Identified on human and rodent lymphocytes, the lymphocyte lectin family members have received increasing attention in recent years. In particular, the mouse Ly49 family and human CD94/NKG2 family expressed on NK cells can trigger or inhibit cell lysis by NK cells through interacting with MHC class I molecules expressed on the target cells, reviewed by (Natarajan *et al.*, 2002).

1.2.3.2 Toll-like receptors (TLR)

TLRs form a family of ancient microbial PRRs highly conserved from fruit flies to humans. The extra-cellular domain of TLRs contains interspersed leucine-rich repeats (LRR) recognizing a variety of PAMPs, while the cytoplasmic TIR domain is homologous to the interleukin (IL)-1 receptor, which signals through the NF-kappa B pathway.

More than 10 members of TLR family have been identified so far, each recognizing a different microbial pattern. These pattern motifs include mannans in yeast cell walls, various bacterial cell wall components (i.e. lipopolysaccharide (LPS), peptidoglycan (PGN), lipopeptides), viral components (i.e. double-stranded RNA, deoxyoligonucleotides), *etc.* The recognition of these ligands usually triggers the NF-kappa B pathway through the TIR domain, leading to the production of anti-microbial peptides and cytokines such as interferons (IFN), which finally activate adaptive immunity, reviewed by (Takeda and Akira, 2005).

1.2.3.3 Complement proteins

The complement system is the main humoral arm of innate immunity. It plays an essential role in host defense against invading microbes during inflammation. The complement acts through a complex cascade involving over 20 serum glycoproteins that function either as enzymes or as binding agents.

The complement system can be activated in three different ways: 1) the classical pathway, 2) the lectin pathway, and 3) the alternative pathway. The classical pathway is activated by antibodies. The lectin pathway is activated by the mannose-binding protein that recognizes the mannose groups of microbial carbohydrates. The alternative pathway is activated by C3b binding to the cell walls and other surface components of foreign microbes. Activated complement proteins will initiate a cascade of enzymatic reactions leading to the lysis of pathogens or antibody-flagged cells by forming the membrane attack complex (MAC). Some complement factors can attract inflammatory cells through chemotaxis (i.e. C5a). Some can bind to the surface of the microbes and act as opsonins to promote phagocytosis (i.e. C3b).

1.2.3.4 Antimicrobial peptides (AMP)

AMPs form another diverse and evolutionarily ancient family of effector molecules in innate immunity. A large number of AMPs have been isolated from a variety of species, including insects, amphibians and mammals. AMPs are usually short peptides between 12-50 amino acids long. Based on their net charge, AMPs can be grouped into cationic and anionic peptides. Most AMPs are cationic peptides which are further divided according to their composition and structural features. For example, cecropins are linear amphipatic-helical peptides devoid of cysteine residues, while defensins are complex cysteine-rich peptides which usually consist of a beta-sheet, reviewed by (Sitaram and Nagaraj, 2002).

Three types of defensins have been identified in mammals, alpha-defensins, beta-defensins and theta (circular) minidefensins. All of them have a beta-sheet structure stabilized by three intramolecular cysteine disulphide bonds in different orders, reviewed by (Lehrer and Ganz, 2002). Defensins are usually synthesized as larger protein precursors, which are then cut to produce smaller active forms. Some defensins are secreted constitutively by epithelial cells, and some are brought to the site by leukocytes in response to pathogen invasions (Selsted and Ouellette, 1995). Based on the fact that defensins are cationic peptides while bacterial membranes usually have a surplus of negative charges, one of the proposed bactericidal mechanisms of defensins is that they penetrate and form lethal holes in the membrane of bacteria (Matsuzaki *et al.*, 1995).

1.2.4 Important molecules in adaptive immunity

1.2.4.1 Major histocompatibility complex (MHC)

Originally described as major transplantation antigens that determine the tissue compatibility between different individuals, MHC is now considered to be the foundation of adaptive immunity. With a few exceptions, the adaptive immune system only responds to environmental stimuli by recognizing antigens presented in the context of MHC molecules. For instance, T cells cannot recognize antigens in solution. They only respond to antigens presented in the context of self-MHC (thus termed "MHC restriction"). Although B cells can recognize antigens in their native forms, the initial activation of B cells requires accessory signals from T cells, since most antigens are T cell-dependent. This suggests if a microbe mutated its proteins such that none of its peptides can bind to MHC, it will not be able to trigger an adaptive immune response. This partially explains why MHC molecules identified so far among different species are usually polymorphic and polygenic, a clear sign of natural selection to maximize their antigen presenting capacities.

There are two types of MHC molecules with very similar three-dimensional structures. The most prominent is the peptide-binding platform in which two MHC-unique domains fold together to form a groove whose base is a beta-sheet with sides consisting of two alpha-helices. MHC class I is a heterodimer of membrane-bound alpha chain and non-covalently associated beta$_2$-microglobulin (β_2M). The alpha1 and alpha2 domains form the peptide-binding groove. β_2M is associated with the alpha3 domain and is essential for the stability and function of the MHC. MHC class I molecules are expressed constitutively on almost all nucleated cells. They interact with the CD8+ T cells, in which the T cell receptor (TCR) recognizes the specific peptide presented in the context of self-MHC, and the CD8+ molecule binds to the alpha3 domain of the MHC class I alpha chain (Tanabe *et al.*, 1992). MHC class II molecules are formed by non-covalently bonded heterodimer of alpha and beta chains of similar size. Peptide antigen is bound to the cleft formed between alpha1 and beta1 domain. The expression of MHC class II molecules is restricted to professional APC. MHC class II molecules interact with CD4+ T cells, in which the TCR binds the specific antigen-MHC complex and the CD4+ molecule binds to the beta2 domain of MHC class II beta chain (Konig *et al.*, 1992).

1.2.4.2 Immunoglobulins (Ig)

There are two forms of immunoglobulins (Ig): membrane bound and secreted. The former are B cell receptors (BCR) and the latter are the effector molecules in humoral immunity, commonly known as antibodies. Immunoglobulins are composed of two heavy chains and two light chains. Their antigen-binding sites are protein heterodimers encoded by somatically rearranged gene elements - V and J elements for light chains; V, D, and J elements for heavy chains.

In mammals, the diversity of antibodies is mainly the result of V(D)J rearrangements, in which the recombination activation gene (RAG) encoded recombinases randomly select and cleave recombinational signal sequences (RSS) flanking multiple V, D, J segments to produce wide varieties of antibodies that potentially recognize a wide variety of antigens. In birds, a different process named gene conversion works as the main mechanism in generating the Ig repertoire. In chickens, for example, there is only a single functional light chain V element, which recombines with a single J element early in B cell development. The rearranged VJ then undergoes an iterative homologous recombination process in which an array of nonfunctional pseudo-V elements, located upstream of the functional V segment, provide templates/donors for sequence transfer (Reynaud *et al.*, 1987).

Antigen binding will lead to cross-linking of BCR which triggers downstream signaling. Together with the help of T cells, B cells will be activated to initiate clonal expansion, proliferation and finally mature into antibody-secreting plasma cells. During the process, the antigen-binding sites will undergo point mutations to increase the affinities for the same epitope. They may further undergo class switching by rearranging the C elements to diversify their effector functions without changing the antigen specificity (Rothman *et al.*, 1989).

1.2.4.3 T cell receptors (TCR)

TCR molecules are structurally and functionally similar to immunoglobulins. TCR are heterodimers composed of alpha/beta or gamma/delta chains, each with a variable and a constant domain. In human and mouse, most T cells express alpha/beta TCR. A small proportion of T cells expressing gamma/delta TCR are mainly distributed in the skin and gut. These T cells recognize antigens expressed by injured epithelial cells and produce factors that affect wound repair (Jameson *et al.*, 2002). Like antibodies, the repertoire of TCR is also generated by VDJ rearrangements enabled by RAG encoded recombinases (Davis and Bjorkman, 1988), but there is no somatic hypermutaion or class switching taking place after antigen recognition.

1.2.5 Cytokines

Cytokines are a family of small secreted proteins involved in the regulation of immune responses, inflammation and hematopoiesis (development of blood cells). They are released by a variety of different cell types. Cytokines together with cytokine receptors form intricate networks to orchestrate the activities of target cells during immune responses. Many cytokines are found to share certain signaling properties with others, which sometimes blur the distinctions between them. For the convenience of this discussion, cytokines are somewhat arbitrarily divided into four partially overlapping groups: interleukins (IL), colony stimulating factors (CSF), inteferons (IFN) and chemokines.

1.2.5.1 Interleukins (IL)

As the name suggests, ILs mainly act as messengers between (inter-) leukocytes. They also play essential roles in immune cell activation, proliferation and differentiation. Different ILs are usually produced by different cell types, regulating different spectrum of target cells. Most ILs exert their effects on target cells by triggering the JAK-STAT signaling pathway.

IL-1 is mainly produced by activated macrophages. It is the major pro-inflammatory cytokine. IL-1 is very important in activating T helper cells, promoting B cell maturation, and regulating DC activation (Koide *et al.*, 1987). IL-

2 is an autocrine factor for both Th1 and NK cells. IL-4 is a Th2 autocrine growth factor. Binding of IL-4 to its receptors can induce Th0 cells to differentiate into Th2 cells. IL-4 can also inhibit the initiation of cytolytic function of CTL by altering the ratio of co-stimulatory ligands expressed on DC (King *et al.*, 2001). IL-12 is produced by activated DCs and macrophages. It can induce Th0 cells into Th1 cells. Thus, the cytokine milieu, especially the IL-4/IL-12 ratio is important in influencing the Th1/Th2 decision at early stages in immune responses (Lederer *et al.*, 1996).

1.2.5.2 Colony-stimulating factors (CSF)

CSFs are the major regulators for the production of two types of immune cells: the macrophages and granulocytes. There are three major types of CSF: macrophage-CSF (M-CSF), granulocyte-CSF (G-CSF), and granulocyte macrophage-CSF (GM-CSF). The functions of the first two are relatively lineage-specific, while the last acts at an earlier stage of lineage commitment, regulating the expansion and maturation of primitive hematopoietic progenitors, reviewed by (Barreda *et al.*, 2004). Binding of CSF to their receptors stimulates the growth and proliferation of hematopoietic cells into different blood cell types.

M-CSF is produced by diverse cell types, including fibroblasts, epithelial cells and macrophages. It is required for the growth, survival and differentiation of mononuclear cells. M-CSF receptor (MCSFR), now known as CSF1R, is primarily expressed on cells of the macrophage lineage. It mediates most of the biological effects of M-CSF. Ligand binding induces the receptor dimerization and transphosphorylation, triggering the receptor tyrosine kinase (RTK) signaling pathway (Reedijk *et al.*, 1992).

1.2.5.3 Interferons (IFN)

IFNs play a critical role in antiviral immunity. Two types of IFN are found in humans: type I and type II. IFN-alpha and IFN-beta belong to type I IFN, while IFN-gamma is type II IFN. Like ILs, IFNs also act through the JAK-STAT signaling pathway. They bind to their respective receptors, resulting in different, though partially overlapping cellular effects. The expression of IFN is regulated by a family of IFN regulatory factors (IRF).

Type I IFN mainly accounts for the antiviral activity of IFN. IFN-alpha is secreted by virus-infected cells. Plasmacytoid DC were found to be the major IFN-alpha producer after viral challenge (Kadowaki *et al.*, 2000). Both IFN-alpha and IFN-beta bind to the ubiquitously expressed type I IFN receptors. Ligand binding will induce an “antiviral state” in the cells, resulting in the inhibition of both viral replication and cell proliferation. Type I IFNs also enhance of the killing ability of NK cells, reviewed by (Le Page *et al.*, 2000). Type II IFN or IFN-gamma is mainly produced by activated T cells and NK cells. They also exhibit some antiviral activity. IFN-gamma can act directly on virus infected target cells, inhibiting the viral replication (Costa-Pereira *et al.*, 2002). However, their antiviral activity is significantly less potent compared to type I IFNs. The main function of IFN-gamma is controlling immune function. Binding of IFN-gamma to the target cells increases the surface expression of class I or class II MHC molecules, enhancing their abilities in antigen presentation, reviewed by (Boehm *et al.*, 1997).

1.2.5.4 Tumor necrosis factors (TNF)

Originally identified as lymphocyte and macrophage products which caused lysis of tumor cells, the TNF family, together with their receptors has now grown into a cytokine superfamily involved in inflammation, immunity, and apoptosis. TNF-alpha is produced by activated T cells, NK cells and macrophages. TNF-beta, also known as lymphotoxin-alpha, is expressed by NK cells, T cells and B cells. TNF receptors are mainly expressed on immune cell types, reviewed by (Locksley *et al.*, 2001).

Depending on the context, ligand binding can lead to opposite destinies of the target cell. In normal cells, TNF-alpha is mitogenic and can induce cell activation and proliferation, whereas in transformed cells, binding by TNF-alpha will lead to programmed cell death or apoptosis. Two adaptor molecules were identified to be pivotal for the signaling result - the TRAF (TNF receptor–associated factor) and the TRADD (TNF receptor-associated death domain) molecule. The former activates the NF-kappa B pathway, while the latter triggers the caspase pathway, reviewed by (Inoue *et al.*, 2000). Blocking the NF-kappa B signaling pathway was shown to result in normal cell apoptosis in response to TNF-alpha (Pimentel-Muinos and Seed, 1999).

1.2.5.5 Chemokines

Chemokines are known for their selective attraction and activation of leukocytes at the site of inflammation during immune responses. All chemokines share the same basic tertiary structural framework with at least three beta-sheets and a C-terminal alpha helix (Koopmann *et al.*, 1999). Most chemokines have at least four cysteines in conserved positions. Chemokines are divided into CXC and CC chemokines based on the position of two cysteines at their N-termini. They are adjacent in the CC family and are separated by a single amino acid in the CXC family. The chemokines discovered so far have outnumbered their receptors. Thus in general, many chemokines bind multiple receptors and many receptors have multiple chemokine ligands, creating the potential for combinatorial diversity in their functions, reviewed by (Kim, 2004). Most chemokines act via the G-protein-coupled receptors (GPCR) signaling pathway.

Macrophage inflammatory proteins (MIP)-1 alpha, MIP-beta and RANTES are important members of the CC chemokine family. They are structurally and functionally related pro-inflammatory cytokines. MIP-1 alpha and MIP-beta are both produced by macrophages, neutrophils, eosinophils and inflammatory fibroblasts. They have similar chemotactic activities on monocytes, B cells and DCs.

However, differential effects on T cells have been reported. MIP-1 alpha selectively attracted CD8+ T cells, while MIP-1 beta selectively attracted CD4+ T cell (Schall *et al.*, 1993; Taub *et al.*, 1993). RANTES, also known as CCL5, is produced by circulating T cells after being stimulated by TNF-alpha and IL-1. RANTES is chemotactic for T cells, eosinophils and basophils. Three RANTES receptors have been identified, CCR1, CCR3 and CCR5 (Bischoff *et al.*, 1993). It was reported that RANTES, MIP–1 alpha and beta played important roles in the suppression of HIV replication (Cocchi *et al.*, 1995). CCR7 is a chemokine receptor expressed on mature DCs, activated B cells, and T cells (Forster *et al.*, 1999). Two CCR7 ligands have been identified so far, CCL21 and CCL19. Both are produced by stromal cells in the T cell area of secondary lymphoid organs (Scandella *et al.*, 2004). They are the principle regulators of immune cell trafficking during primary immune responses.

1.2.6 Important immune signaling pathways

The most salient feature of the immune system is its numerous specialized receptors and effector molecules that are able to recognize and destroy pathogens and transformed cells. Compared to them, the signaling pathways involved in immune responses are much less function specific. Many of these signaling pathways also participate in other physiological or metabolic processes.

Except for steroids and nitric oxides (NO) that are able to diffuse into the cells and bind internal receptors, most signaling molecules (i.e. proteins) are too large or polar to pass the cellular membrane. They have to rely on surface receptors to enter into a cell's interior. Receptor mediated signaling pathways are essential for the regulation of immune responses. It is through these specific receptors and other accessory cues that cells receive and respond to environmental stimuli and coordinate with each other in a proper way to carry out immune defenses.

The signaling pathway works basically like a molecular circuit. Receptor molecules usually either possess themselves intrinsic ligand-triggered kinase activity or associate with adaptor proteins that provide the enzymatic activity. Ligand binding will cause conformational changes (i.e. cross-linking or oligomerization) in the receptor molecules, converting the corresponding protein kinases into an active state. Downstream protein phosphorylation usually triggers the release of second messengers (i.e. cAMP, IP3) to relay, and often amplify the signal further inside the cells. In most cases, the signaling cascade leads to the activation of transcription factors (i.e. NF-kappa B, STAT) which subsequently enter the cell nucleus to influence the expressions of target genes. Finally, protein phosphatases terminate the signal by removing the specific phosphoryl groups from the modified proteins.

1.2.6.1 G-protein coupled receptors (GPCR) pathway

GPCR proteins form a large and diverse receptor family. Their common features are seven transmembrane domains (7TM) and most of their intracellular pathways involving the activation of G-proteins, reviewed by (Kristiansen, 2004). G proteins are named as such because they bind guanine nucleotides: GDP-bound when it is inactive and GTP-bound when active. Ligand binding changes the GPCR conformation and activates the G protein which then detaches from the receptor to trigger downstream signaling cascades. Depending on the type of G protein activated, a variety of pathways can be initiated. The GPCRs are among the largest protein families known, members of which are involved in diverse activities including stimulus-response pathways (i.e. visual, olfactory receptors), behaviors (i.e. receptors for neurotransmitters in brain). GPCR pathways also play an important role in immune regulation. Many chemokine receptors signal through the GPCR pathway, reviewed by (Neves *et al.*, 2002)

1.2.6.2 Receptor tyrosine kinase (RTK) pathway

As the name suggests, the membrane receptors of this pathway usually possess intrinsic kinase activity in their cytoplasmic tail. They usually span the plasma membrane just once. Ligand binding of two adjacent receptors causes the formation of a homodimer, converting the receptor's protein kinase into active state. Subsequent phosphorylation cascades lead to the activation of Ras protein, a small G protein (they are monomeric compared to the heterotrimeric G protein of the GPCR pathway) serving as a common node in RTK signaling. Ras protein is a pro-oncogene. Mutations in Ras have been identified in many tumor types, reviewed by (Bos, 1989). The RTK pathway is involved in regulating a wide spectrum of activities such as cell proliferation and differentiation, apoptosis and metabolism. Several cytokines (i.e. growth factors, M-CSF) act through this pathway.

1.2.6.3 JAK-STAT pathway

The JAK-STAT pathway is the primary signaling mechanism for a wide array of cytokines and growth factors. JAK (Janus kinase) is a group of non-receptor tyrosine kinases. STAT (signal transducer and activator of transcription) serves both as a signal transducer in the cytoplasm and as a transcription factor in the nucleus. It regulates the expression of a large group of cytokines and growth factors. The membrane receptors for JAK-STAT pathways are single-pass transmembrane proteins that lack intrinsic kinase activity. Ligand binding will induce receptor dimerization, which then activates the associated JAK. JAK will phosphorylate certain tyrosine residues in one or another STAT, inducing them to form active dimers, which are subsequently translocated into the nucleus and affect the transcription of the target genes. GM-CSF, and the majority of IL use the JAK-STAT pathway, reviewed by (Schindler, 1999).

1.2.6.4 NF-kappa B pathway

Nuclear factor kappa B (NF-kappa B) is a central transcription factor involved in immunity. It is expressed in a wide variety of immune cells and has been linked to cellular transformation, proliferation, and apoptosis suppression. In the normal state, NF-kappa B resides in the cytosol bound to an inhibitor called I-kappa B. Upstream activation will lead to phosphorylation of I-kappa B, which will subsequently become ubiquitinated and destroyed by proteasomes, freeing the NF-kappa B. NF-kappa B will then enter the nucleus where it binds to the promoters/enhancers of more than 60 genes, reviewed by (Baeuerle and Baltimore, 1996). MHC proteins, Toll-like receptors, TNF-alpha, IL-1, 2, 6, 8, adhesion molecules, chemokines all act through this signaling pathway, reviewed by (Richmond, 2002).

1.2.7 Antigen processing and presentation

Antigen processing and presentation through MHC molecules is the first step towards initiating adaptive immune responses. Different sources of antigens are usually processed in different pathways and presented on different types of MHC molecules, resulting in different types of immune responses. The endogenous peptide antigens are presented via MHC class I molecules to CD8+ T cells, while the exogenous peptides antigens are presented through MHC class II molecules to CD4+ T cells. However, DCs are capable of cross priming in which the exogenous derived peptides can be processed and presented via the MHC class I pathway.

1.2.7.1 Endogenous antigen processing and presentation pathway

Endogenous antigens are antigens that have been generated within the cells such as viral peptides and tumor antigens. They are degraded in the cytosol by proteasomes.

These fragments are then transported by transporter associated with antigen processing (TAP) protein across the membrane of the endoplasmic reticulum (ER). Within the ER, the peptides, the newly synthesized MHC class I heavy chain and β_2M form a stable complex, which is then transported to the Golgi apparatus and finally displayed on the cell surface. If activated CD8+ T cells recognize these antigens, they will secrete toxins (i.e. perforins and granzymes) to cause lysis of the infected cells or use Fas ligand to induce apoptosis of the malignant cells, reviewed by (Andersen *et al.*, 2006).

1.2.7.2 Exogenous antigen processing and presentation pathway

Exogenous antigens are antigens that have entered the body from the outside. They are fragmented by proteases and stored in endosomes. The MHC class II alpha chain and beta chain, along with the invariant chain, are synthesized and assembled in the ER to form class II MHC complexes, which are subsequently transported to the Golgi apparatus to fuse with the endosome. Inside the endosome, the invariant chain is replaced with the peptide fragment. Finally, the complexes are transported and displayed on the cell surface. The professional APC are the main cell types that express MHC class II molecules. They can activate antigen-specific T helper cells (CD4+), which then secrete cytokines that further activate CTL or B cells.

1.2.7.3 Cross priming

Cell-mediated immunity (CMI) is the main mechanism of host defense to clear the body of transformed cells or viral infections. However, if viruses evolve mechanisms to either block the MHC class I processing pathway or simply avoid infecting antigen presenting cells, the host will fail to activate the CMI. The cross priming pathway complements the normal CMI processing pathway by allowing the exogenously derived antigens to be presented on both types of MHC molecules. Thus, CD8+ T cells can be activated either by endogenously generated antigens (direct priming) or antigens from exogenous sources (cross priming). This understanding has profound implications in vaccine development. For instance, in HIV or cancer vaccine development, to induce long-term CMI (i.e. antigen specific CTL and CD8+ memory T cells) is usually the ultimate goal. However, to establish long term CMI requires the presence of T helper cells. It is essential that both the T helper cell and the CTL reside on the same DC at the time of activation (Shedlock and Shen, 2003; Sun and Bevan, 2003). This implies that the same source of antigen needs to be presented via both MHC class I and class II molecules on the APC at the same time, which is only possible if cross-priming is allowed. DCs are the principle cell type capable of cross-priming, reviewed by (Behrens *et al.*, 2004). Without being actively infected themselves, DCs can process and present on MHC class I many different types of exogenously derived peptides, including immune complexes, antigens originally synthesized in other cells, antigens, or vaccines to initiate potent CTL responses, reviewed by (Heath *et al.*, 2004).

1.2.8 The avian immune system

Although the major body of the aforementioned knowledge on the immune system is based on human and mouse models, most mechanisms and features are equally applicable to birds. Like mammals, birds possess both innate and adaptive immunity, with the latter further divided into humoral and cell mediated immunity. However, the avian immune system differs from that of mammals in two major ways: the presence of a unique specialized gut-associated immune organ called bursa of Fabricius, and a process named gene conversion.

In mammals, B cells are produced and mature primarily in the bone marrow. While in birds, B cells are initially produced in the embryonic liver, yolk and bone marrow, then they move to the bursa of Fabricius to mature. Inside the bursa, the avian Ig locus undergoes a single recombination event involving a single functional V, D and J element. This is then diversified through rearrangement in pseudogenes by gene conversion, reviewed by (Lundqvist *et al.*, 2006). This process mainly takes place during the first six weeks in a bird's life (Mueller *et al.*, 1960).

1.3 The duck model

Ducks have traditionally been an important agricultural species, with more than 3 million metric tons of meat produced annually (http://www.fao.org). Due to the recent re-emergence of bird flu, ducks are in the spotlight for their roles in the dispersal and evolution of the influenza A viruses (Webster and Hulse, 2005). In addition, ducks are also an invaluable animal model for hepatitis B virus (HBV) infection, reviewed by (Schultz *et al.*, 2004). The study of duck immunity is important to both veterinary and human public health.

As the first bird genome ever sequenced (Hillier *et al.*, 2004), the chicken genome provides an invaluable resource for the study of bird evolution and immunity. In contrast, the genetic resource for duck is very limited in the public databases. By 2005, the NCBI *dbEST* database contained only 234 sequences from duck, *Anas platyrhynchos* (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).

1.3.1 Ducks and Influenza A viruses

As natural hosts of the influenza A viruses, the ducks are considered to be the prime sources of all human influenza pandemics, reviewed by (Webster *et al.*, 1992). Under normal conditions, these viruses are usually species-specific, showing a preferential binding to avian sialic acid receptors (Matrosovich *et al.*, 1999). However, the increasing number of confirmed cases of human avian influenza in recent years (http://www.who.int) showed that the bird flu viruses are evolving quickly. There is concern that they could acquire the ability to jump to humans and cause another pandemic.

Ducks are often asymptomatic carriers of the influenza viruses which are lethal to other domesticated poultry. Even the highly pathogenic H5N1 strain was found to be reverting to nonpathogenicity in ducks (Hulse-Post *et al.*, 2005), making ducks the "Trojan horses" for harboring and spreading the influenza viruses. Therefore, to understand the duck immune system at molecular level, particularly the identification of genes involved in host-virus interaction, holds the key to understand the evolution of the virus and to permit effective vaccination strategies.

1.3.2 Ducks and Hepatitis B viruses (HBV)

Ducks are also among a few animals suitable for the study of HBV infections. Unlike mice, ducks are natural hosts of duck hepatitis B viruses (DHBV) that belong to the same hepadnaviridae virus family as human HBV. These two viruses share many similarities with respect to virion ultrastructure, antigenic composition, genetic organization, and mechanism of replication (Yuasa *et al.*, 1991). In addition, like HBV-infected humans, DHBV-infected ducks exhibit similar symptoms such as persistent infection in young ducks and transient infection in adults (Jilbert *et al.*, 1998). As a result, ducks are often used as an animal model for the study of HBV infection and for testing new antiviral agents (Le Guerhier *et al.*, 2003; Thermet *et al.*, 2003; Foster *et al.*, 2005). The successful application and wide acceptance of this model is currently limited by a lack of markers/reagents to measure the immune response to DHBV.

1.4 Research objectives

The research objective of my lab is to understand duck immune system at molecular level. To this end, a cDNA library was constructed from a duck spleen and 3, 168 clones were randomly picked and sequenced. The objective of my research work was to perform immune discovery based on the resources. This was approached from three aspects: (i) to analyze the EST data to identify immune genes, (ii) to array the cDNA library and screen with homologous probes, and (iii) to generate subtracted probes to screen for immune genes in a high-throughout manner.

1.4.1 Objective 1: EST analysis, annotation, and immune gene identification

Over 3,000 sequences were produced from our EST project. A systematic and disciplined analysis as described in section **1.1** was needed. Basically, we wanted to process the EST data, cluster and assemble them, annotate the sequences, identify immune relevant genes, perform some custom analysis, and finally publish the EST data. Therefore, one of my research goals was to construct a high throughput EST analysis and annotation pipeline to address these requirements.

1.4.2 Objective 2: to array the cDNA library and screen it with homologous probes

Random sequencing can quickly generate a lot of EST fragments from which immune genes can be identified. However, gene discovery through this approach is limited by its random nature, namely, which gene can be discovered is a matter of pure chance. As a complementary strategy, it is also desirable to perform a more targeted search for specific genes of interest. By arraying this cDNA library onto membranes or chips, the clones can be screened with probes designed from homologous genes of other species. This approach will greatly enhance and extend the utility of this cDNA library by allowing further identification and characterization of genes of interest.

1.4.3 Objective 3: to generate subtracted probes to improve the rate of immune gene discovery

The rate of gene discovery by random sequencing is subject to the law of diminishing returns, which limits the scale of our EST project. Using an arrayed cDNA library allows two possible ways to improve the situation without the need to construct new libraries: first, several abundant genes (i.e. immunoglobulins) identified in the previous EST project can be flagged and excluded from the subsequent round of sequencing; second, the arrayed cDNA library can be screened with subtracted probes enriched for immune genes to improve the rate of immune gene discovery in a high throughput manner.

In our case, the probe pool can be generated via tissue subtractions between the duck spleen and brain. The rationale is that the spleen is an organ where immune cells abound, while the brain is usually considered an immune-privileged site depleted of conventional types of immune cells under normal conditions. Therefore, by performing subtraction between mRNAs extracted from the brain and the spleen tissue, most transcripts expressed by common house-keeping genes will be removed while the differentially expressed ones such as those from the immune cells will remain. This procedure could potentially yield a probe pool enriched for immune relevant sequences, which can then be labeled to screen the arrayed library.

Chapter 2

MATERIALS AND METHODS

2.1 cDNA library construction and EST sequencing

A cDNA library was constructed from a healthy male White Pekin duck (*Anas platyrhynchos*) in the lambda ZAP Express vector (Mesa *et al.*, 2004). Two-thirds of the library was taken through a single round of amplification by plating on host cells and eluted in SM buffer [0.1 M NaCl, 0.8 M magnesium sulfate, 0.5 M Tris (pH 7.5), 0.1% w/v gelatin] and stored at 4°C with added chloroform (0.3% v/v). A mass excision was conducted using an aliquot of the once-amplified library (1.5×10^7 phage) according to the lambda ZAP product manual (*Stratagene*). cDNA clones were recovered as bacterial colonies and picked by the *QPix* robot (*Genetix, UK*) into 96-well plates. Bacterial suspension was transferred from each plate to infect a deep-well block, and DNA was isolated using the microwave lysate method (Marra *et al.*, 1999) and sequenced using T3 forward primer on an ABI 377 automated sequencer (*PE Applied Biosystems Inc., Foster City, CA*).

2. 2 EST analysis and annotation

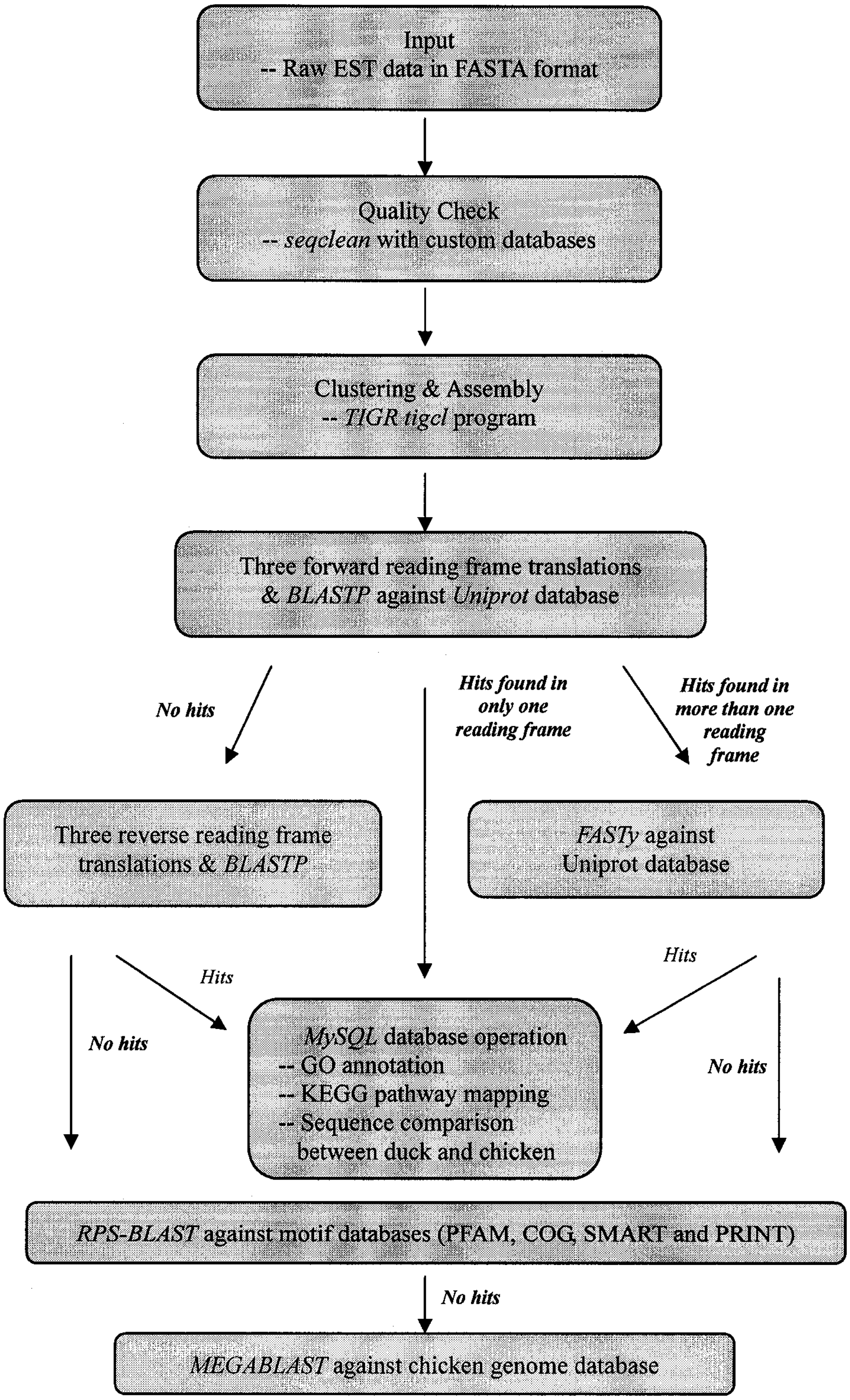
The pipeline for EST analysis and annotation was mainly implemented as a package of *Perl* scripts combined with the MySQL database. The whole analysis was performed on a laptop computer (OS Red Hat Linux 9.0 installed with *MySQL* 4.1, CPU Intel Pentium IV 1.4GHz, 768MB RAM).

Five major steps were carried out during the EST analysis: (i) a sequence quality check, (ii) EST clustering and assembly, (iii) a homology search against the *Uniprot* database, (iv) a motif database search, and (v) a chicken genome database search. In addition, three customized analyses were also performed: (i) Gene Ontology annotation, (ii) KEGG pathway assignment, and (iii) sequence comparison between homologous duck and chicken genes. **Figure 2.1** shows a diagram of the EST analysis pipeline. The main program used at each step is indicated in the corresponding box.

2.2.1 EST quality checking, clustering, and assembly

Sequence chromatographs were manually inspected during the EST sequencing process, which made it unnecessary to use *phred*. The input was the raw ESTs in FASTA format. In the first stage, a quality check was performed using the TIGR *seqclean* program against a custom sequence database (composed of pBK-CMV vector sequences, adaptor and linker sequences). The vector sequences together with low complexity sequences (repeats, polyA/T tails, *etc*) were trimmed. In addition, sequences calculated to be either of low quality (> 3% ambiguous base calls) or too short (< 100 base pairs after trimming) were also removed. The resulting sequences were then clustered and assembled using the *TIGR tigcl*

Figure 2.1 The EST analysis and annotation pipeline


Input
-- Raw EST data in FASTA format
Quality Check
-- seqclean with custom databases
Clustering & Assembly
-- TIGR tigcl program
Three forward reading frame translations
& BLASTP against Uniprot database
No hits
Hits found in only one reading frame
Hits found in more than one reading frame
Three reverse reading frame translations & BLASTP
FASTy against Uniprot database
Hits
No hits
Hits
No hits
MySQL database operation
-- GO annotation
-- KEGG pathway mapping
-- Sequence comparison between duck and chicken
RPS-BLAST against motif databases (PFAM, COG, SMART and PRINT)
No hits
MEGABLAST against chicken genome database

package with default parameters. In brief, the cleaned EST sequences were first compared to each other using a modified version of NCBI *MEGABLAST* program. Sequences highly similar to each other were put into the same cluster. An assembly process was subsequently performed for each cluster using the *CAP3* program. This process generated contigs based on multiple alignments of member sequences within each cluster. Sequences were left out as singlets if they contained many mismatches or introduced many gaps within the multiple alignments.

2.2.2 Homology detection and motif identification

Unique sequences generated after clustering and assembly (including contigs, singlets and unclustered sequences) were translated in three forward reading frames and then searched against a local *Uniprot* sequence database using the NCBI *BLASTP* program in three separate batches. The result returned from each batch was then parsed and compared with each other. Sequences with significant hits (E value $< 10^{-5}$) in more than one reading frame were considered to contain frame-shift errors. These sequences were searched against the same database using the *FASTy* program (Pearson *et al.*, 1997). To account for possible reversals during the cloning process, sequences without hits in the three forward reading frames were translated in three reverse reading frames and the same strategies were applied to search the *Uniprot* database. In the fourth stage, sequences without any matches were searched against a series of local motif databases (*PFAM, COG, SMART and PRINT*) using the NCBI *RPS-BLAST* program. Finally, all the remaining sequences were compared against the chicken genome database using the NCBI *MEGABLAST* program.

2.2.3 EST annotation and immune gene mining

The search for immune-relevant genes was mainly carried out by manual inspection of the returned *Uniprot* annotations for each unique sequence. The criteria for "immune relevance" are based on the *IRIS* (Immunogenetic Related Information Source) database (Kelley *et al.*, 2005). Basically, they include genes that have known or putative functions in innate or adaptive immunity, genes that participate in immune system development and maturation, genes that are induced by an immune modulator, genes that are expressed primarily in immune tissue, genes involved in immune signaling pathways, and genes whose products interact directly with pathogens.

2.2.4 GO annotation and pathway assignment

The majority of the *Uniprot* hits have GO terms assigned, which permits further GO annotation for the query sequence. The Gene Ontology data files were downloaded and installed on a local *MySQL* database according to the instructions at the EBI web site (http://www.ebi.ac.uk/GOA/). GO annotation was performed by mapping the accession number of *UniProt* hits to their associated GO terms in the *MySQL* database.

For KEGG pathway mapping, the EC numbers of the query sequences were primarily obtained by parsing the EC numbers from the definition line of their associated *Uniprot* hits. The final list of EC numbers was submitted to the KEGG pathway database (http://www.genome.jp/kegg/tool/search_pathway.html) using the chicken pathways as the references.

2.2.5 Sequence comparison between duck and chicken

In order to compare the sequence similarity between chicken and duck with regard to different functional groups and between immune and non-immune genes, the unique sequences were further divided into different categories based on GO classification. Only sequences that received hits in all three databases searches (*Uniprot* database, Gene Ontology database and chicken genome database) were analyzed this way. To avoid the skewing of the results by over-represented genes, the following sequences were removed from the dataset: ribosomal protein sequences, immunoglobulin sequences, and clusters with multiple contigs (only contig1 was used for each cluster). In addition, sequences with a total aligned length shorter than 100 base pairs were also excluded because most of these matches represented only short conserved functional domains with over 99% percent identity. The sequence similarities were calculated using percent identity scores of the high scoring pair (HSP) from the *BLASTN* reports. The average sequence similarity of each functional group between duck and chicken was calculated using the sum of all identical base pairs divided by the sum of the total aligned length for that group.

2.3 cDNA library arraying and screening

2.3.1 Library packaging and mass excision

An aliquot (0.5 μl) of the original, unamplified cDNA library (Mesa *et al.*, 2004) was packaged *in vitro* into lambda phage using *Gigapack* III Gold packaging extract (*Stratagen*). The resulting phage library was then titered by counting the plaque forming units (PFU) after infecting XL1-Blue MRF' host cells. The library was subsequently mass excised in the presence of the *ExAssist* helper phage in the host cells, which yielded the pBK-CMV phagemids. The phagemids were then used to infect *E coli* XLOLR cells, which produced the pBK-CMV plasmids with the inserts in the XLOLR cells. The plasmid library was titered by counting the colony-forming units (CFU) after growing XLOLR cells overnight on kanamycin-containing plates.

2.3.2 Library picking and arraying

The *E coli* XLOLR cells containing pBK-CMV plasmids were then plated onto Q-tray plates (25 x 25 cm, *Genetix*) with an average density ~ 2,000 colonies per Q-tray. These colonies were picked using *QBot* (*Genetix*) into 384-well plates in LB broth containing 5% glycerol and kanamycin (50 μg/ml), and incubated overnight. The bacterial liquid cultures were then printed onto duplicate sets (set A and B) of nylon membranes using *QPix* (*Genetix*) and grown for 16~18 hours. The colonies on the membranes were then lysed and cross-linked for hybridization. The Q-tray plates were stored in the cold room at 4 °C. The 384-well plates were stored in the freezer at –80 °C. The membranes were kept in a paper cassette and stored at room temperature in a dry drawer.

2.3.3 Library evaluations and screening

Since the immunoglobulins were found to be most abundantly expressed in the spleen cDNA library based on our EST data, the first round of library screening was carried out using Ig probes to check the overall quality of the arrayed library and also to flag them from subsequent sequencing. The ^{32}P-dCTP labeled Overgo probes (Ross *et al.*, 1999) were designed based on IgL constant region and IgM mu chain (**Table 2.1**). After pre-hybridizing for 2 hours, the membranes were then hybridized overnight, washed at high stringency and then exposed to X-ray film for 48 hours at –80 °C.

The second round of library screening was carried out to screen for six specific immune genes: CSFR1, beta-defensin, CCR7, CCL19, TLR7 and IFN-alpha using ^{32}P-dCTP labeled Overgo primers (**Table 2.1**). Positive clones were identified on the membranes, the corresponding bacterial cultures were retrieved from the 384-well plates and grown overnight in 2 mL liquid cultures, followed by DNA miniprep and sequencing to determine their identities.

Table 2.1 Overgo primers used for library screening

Primers	Ova (5'->3')	Ovb (5'->3')
IgL	TCCTGGGCCAGCCCAAGGTGTCTC	AAGACGTGGACAGTGGGAGACACC
IgM	TCACCAACGCCACCGTGGCCACCA	GGGAAGTTGACAATGGTGGTGGCC
CSFR1	GGAGCCCACACAAAGACCTACCTT	GCAGCAGATTTGGTCAAAGGTAGG
TLR7	CTGAATGCTCTGGGAAAGGTTGTC	AGGACAGCCAGTCTTTGACAACCT
CCR7	TTGAAGGCTGAGATGGTCTTGACC	ACAACAGCGTCATGCTGGTCAAGA
CCL19	CAGGAAGGTCCCAAATAAAGGCAA	CTTCTTCAGGGCCTAATTGCCTTT
Defensin	GGGAGATGTTCAACTCTAGTTCCC	ATGTACTCCTGCAGCAGGGAACTA
IFN-alpha	ATGCCTGGGCCATCAGCCCCACCA	TGTAGATGGCTGGTGGTGGTGGGG

2.4 SSH probe synthesis and library screening

One gram of frozen tissue from the duck spleen or brain was ground into a fine powder separately with mortar and pestle in liquid nitrogen. Using the *FastTrack2.0* mRNA extraction kit (*Invitrogen*), the sample was homogenized and digested for 60 minutes at 45 °C in the lysis buffer. mRNA was then extracted using oligo(dT) cellulose and eluted with elution buffer.

Suppression subtraction hybridization (SSH) was performed using the PCR-Select cDNA Subtraction Kit (*Clontech*) according to the manufacturer's instructions. In brief, the mRNA samples extracted from the duck spleen and brain tissue were reverse transcribed into double-stranded cDNAs using avian myeloblastosis virus (AMV) reverse transcriptase with poly(dT) as the primer. The cDNA pools were then digested with RsaI to generate short fragments. To screen for the genes that were differentially expressed in the spleen tissue, the cDNA (tester) from the spleen was further divided into two portions and each ligated with a different 40 bp DNA adaptor. The ligated tester cDNA pools were then subtracted twice against an excess of cDNA from the brain tissue (driver). Using primers complementary to the two adaptors, two rounds of PCR were then performed to amplify the cDNAs that had not hybridized to the driver sequences. In parallel, a reverse subtraction was also performed to amplify the genes differentially expressed in the brain tissue. The products of the major steps in this approach, including reverse transcription reaction, restriction digestion, ligation of adaptors, subtraction, and PCR amplification, were each examined by agarose gel electrophoresis. These subtracted cDNA probes were labeled with ^{32}P-dCTP by random priming and then hybridized with the arrayed cDNA library. Membranes 11A and 11B were chosen to test the synthesized SSH probes. Positive clones were retrieved from corresponding 384-well plates and then sequenced.

Chapter 3

RESULTS

3.1 EST analysis and annotation

3.1.1 Summaries from the EST analysis pipeline

In total, 3,168 clones were randomly selected and sequenced, which yielded 2,900 acceptable sequences. These sequences were sequentially processed through the EST analysis pipeline. The result from each major analysis step was summarized in **Table 3.1**. In brief, 96 sequences were removed during the quality check. The remaining 2,804 sequences were then subjected to clustering and assembly. 1,312 sequences were grouped into 210 clusters, generating 213 contigs and 180 singlets. For small clusters, usually one contig per cluster was identified. While big clusters, such as the immunoglobulin cluster, which contained 236 members for IgL and 178 for IgM, each yielded over 20 contigs. There were also some clusters that produced no contigs as determined by the clustering algorithm. Combined with 1,492 unclustered sequences, this EST library yielded 1,885 unique sequences. The redundancy of this EST library was thus estimated to be about 1.5: 1.

The 1,885 unique sequences were then searched against the *Uniprot* sequence database. In total, 1,248 sequences received significant hits (E value < 10^{-5}) and

Table 3.1 Summary from the EST analysis pipeline

Processing stages	Input	Data description	Number
Chromatogram check	3,168	Poor quality	268
Sequence quality check	2,900	Poor quality	96
Clustering and assembly	2,804	Clusters	213
		Singlets	180
		Unclustered	1,492
Sequence database search	1,885	Hits in one reading frame	1,011
		Frame shifts	185
		Hits in reverse strands	52
Motif databases search	637	Motif search hits	23
Chicken genome search	614	Chicken genome hits*	377
Remaining sequences	237	Novel *	237

* Note: These two groups of sequences do not have associated annotation information.

were used for annotations (including the GO and KEGG). Among them, 1,011 sequences received hits in only one reading frame, 185 were found to contain frame-shift errors, and 52 were found in the reverse strands. After this step, 637 sequences were left without any match. These sequences were further searched against a series of motif databases, which returned an additional 23 hits.

The remaining 614 sequences were then compared against the chicken genome database. 377 sequences received matches in the chicken genome database. Unfortunately, these hits contained only information on their chromosomal location without any functional annotation. 237 sequences did not receive any match after the final stage. Manual checking of these sequences indicated that their quality was acceptable, and they are likely to represent novel transcripts or 3' untranslated regions (3' UTR).

These EST sequences were deposited in the GenBank *dbEST* database under accession numbers DR763793 - DR766439 and DR783112 - DR783189, respectively.

3.1.2 Duck spleen gene expression profile

The 20 most abundant transcripts from the EST project are listed in **Table 3.2.** Since the library was not normalized prior to random selection of the clones, it reflected the baseline information for the duck spleen gene expression profile under

Table 3.2 The 20 most abundant ESTs in duck spleen.

Putative ID	Number of clones	Percentage (%)
Immunoglobulin light chain	236	8.4
Immunoglobulin heavy chain	178	6.3
Ferritin H subunit	60	2.1
60S ribosomal protein subunits	45	1.6
Elongation factor 1 alpha 1	41	1.5
Hemoglobin beta chain	29	1.0
40S ribosomal protein subunits	26	0.9
Cytochrome c oxidase subunit	22	0.8
Invariant chain (Ii)	21	0.75
Beta-actin	20	0.71
Hemoglobin alpha-A chain	19	0.68
MHC class II alpha chain	13	0.46
Receptor for activated C kinase	11	0.39
Immunoglobulin J chain	11	0.39
MHC class I alpha chain	10	0.36
Beta-2 microglobulin	8	0.29
Nucleophosmin	8	0.29
Tubulin alpha-2 chain	7	0.25
Thymosin beta 4	7	0.25
Ubiquitin	7	0.25

the normal conditions.

Not surprisingly, the most abundant transcripts encoded immunoglobulin molecules, including three types of heavy chain (mu, upsilon and alpha) and one type of light chain (lambda). MHC class I alpha chain was also highly expressed. Further examination of its 10 sequence members revealed that they all derived from the dominant MHC locus that is adjacent to TAP2 gene (Mesa *et al.*, 2004). Among them, 9 sequences were from the allele U*02, and 1 from allele U*03. MHC class II alpha chain was also abundantly expressed. Their cluster contained 13 members. Sequences derived from MHC class II beta chain did not appear in the shortlist. They formed two small clusters, with two members in each, and two unclustered sequences. Several other genes in the list were abundantly expressed in erythrocytes, including ferritin H, hemoglobin and alpha-globin.

3.1.3 Gene Ontology annotation and KEGG pathway assignment

The 1,248 unique sequences that received hits from the *Uniprot* database were used for Gene Ontology annotation. In summary, 968 sequences received matches for Biological Process, 957 for Molecular Function, and 850 for Cellular Component. Among them, 577 sequences received annotations in all three ontology fields. In each ontology field, sequences were put into more general groups using GO slim terms to provide an overview (**Figure 3.1**). In particular, the "cell communication"

Figure 3.1

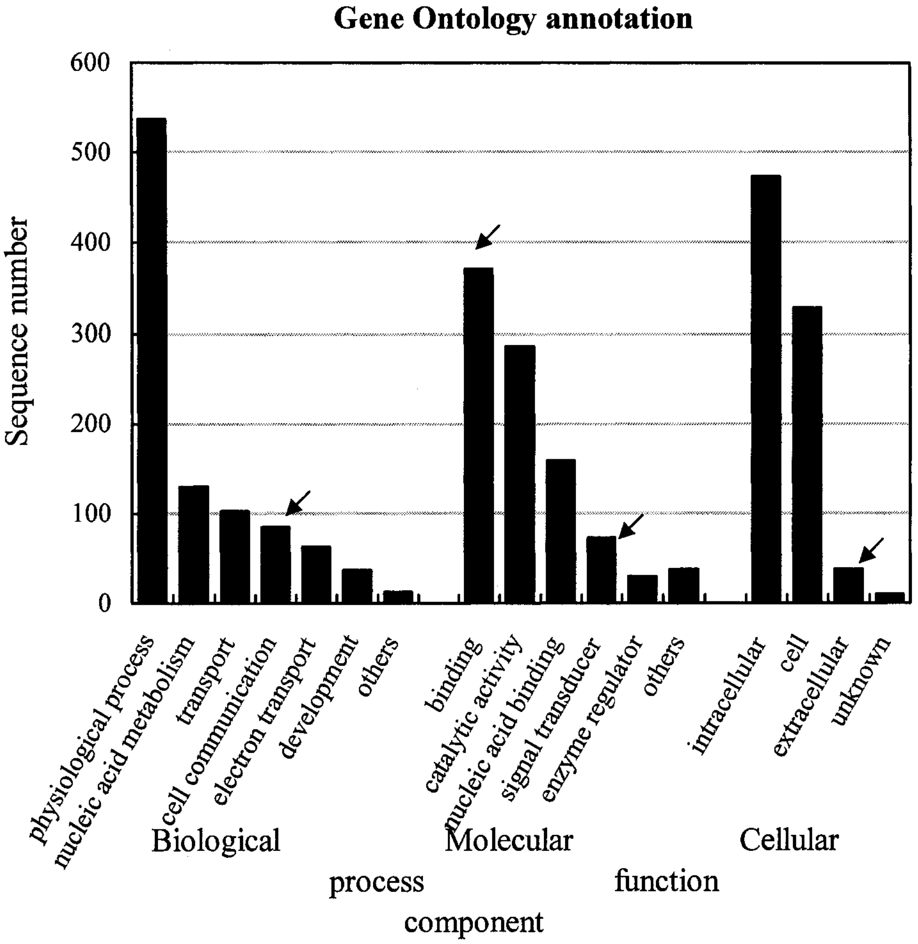


Figure 3.1 GO slim annotations on duck spleen EST. The Y-axis shows the number of sequences. The X-axis shows the three areas of the Gene Ontology. Within each area, sequences are further divided into subgroups at GO slim level to provide an overview. The arrows indicate the groups of sequences that are rich in immune related genes.

group in Biological Process, the “binding” and “signal transducer” groups in Molecular Function, and the “extracellular” group in Cellular Components were found to be enriched for immune genes.

Genes annotated as enzymes were further analyzed using the KEGG pathway if their corresponding EC numbers could be found in the *Uniprot*. In total, 100 unique sequences were found to be assigned with EC numbers. They were submitted to the KEGG pathway database and were mapped to 80 different pathways using the chicken pathways as reference. Most of them were metabolism pathways including Glycolysis, Citrate cycle, ATP synthesis, *etc.* 14 pathways were involved in immune responses **(Figure 3.2).**

3.1.4 Immune-relevant gene mining

Sequence annotations generated from the pipeline were inspected manually for genes involved in immunity. In total, 208 unique duck ESTs were considered to be immune relevant based on their similarity to genes from other species that are involved in immune processes. A list of selected immune-relevant ESTs is shown in **Table 3.3**. Seven categories were used to describe these immune relevant genes: (1) immune cell surface receptors, (2) lectin-like immunoreceptors, (3) cytokines and chemokines, (4) transcription factors involved in immune responses, (5) genes involved in antigen processing and apoptosis, (6) innate immune effectors, and (7) interferon-induced genes.

Figure 3.2

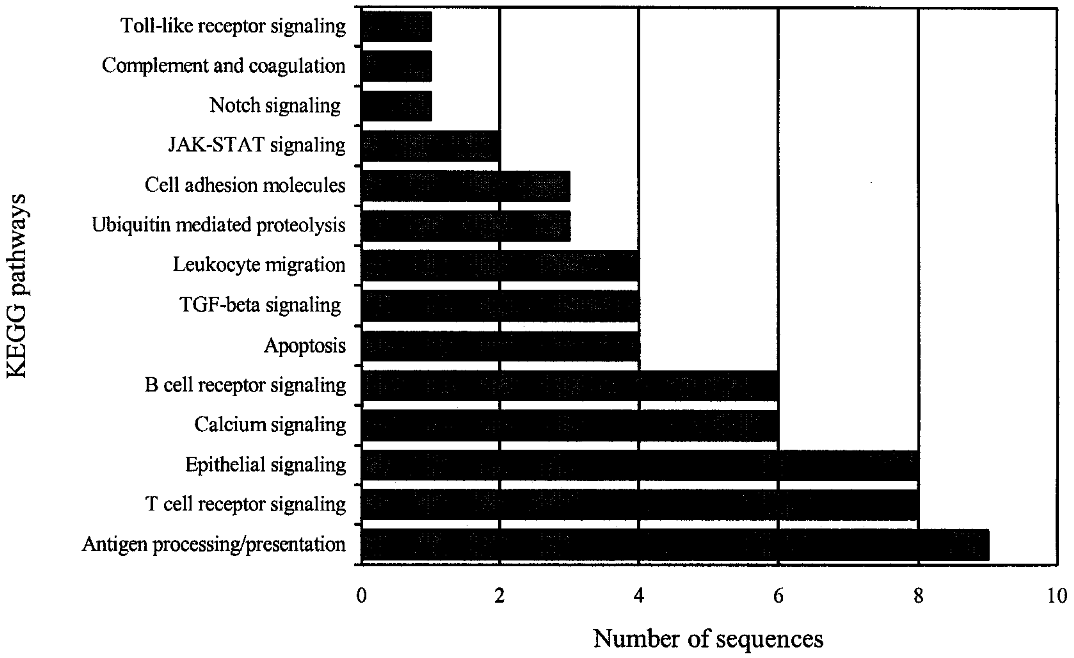


Figure 3.2 Duck genes involved in different immune system pathways. Y-axis indicates the name of the immune pathway. X-axis indicates the number of duck immune genes mapped to each pathway.

Table 3.3 Selected list of immune relevant genes

Duck ESTs		UniProt Matches				Gene Ontology Annotation			Chickenen ESTs	
GenBank ID	**Clone ID**	**Hits ID**	**Description**	**Organism**	**E-value**	**Process**	**Function**	**Component**	**Seq ID**	**Identities**
Immune cell surface receptors										
DR763981	2D8	P30414	NK cell tumor recognition protein	human	4e-28	protein folding	peptide binding	membrane	BU138593	156/180 (86%)
DR764012	2G8	Q8AV16	CD79B	chicken	1e-91	signal transduction	receptor	membrane	AJ443057	340/387 (87%)
DR766072	5C1	Q9TP70	MHC class II alpha chain	human	7e-69	immune response	protein binding	membrane	CK987448	60/67 (89%)
DR766184	6H1	P01887	Beta-2 microglobulin	mouse	8e-25	immune response	MHC class I receptor	membrane	AB178593	236/291 (81%)
DR766202	7B3	P15083	Polymeric Ig-receptor	rat	8e-10	N/A	N/A	membrane	XM_417977	139/164 (84%)
DR763813	10C4	P34902	common cytokine receptor	chicken	2e-67	cell proliferation	interleukin-2 binding	membrane	CV892851	346/390 (88%)
DR765298	27D7	Q99JA5	Lymphocyte antigen 6 complex	mouse	3e-20	N/A	N/A	membrane	CO505312	317/363 (87%)
DR765841	33B6	Q7Z2D4	CD82	human	5e-44	N/A	N/A	membrane	DR428189	436/503 (86%)
DR766103	5E12	Q6NSJ8	CSF2R	human	5e-27	signaling pathway	interleukin receptor	membrane	BU428299	349/403 (86%)
DR766115	5G1	Q6ZPH6	MKIAA1822 protein	mouse	5e-30	endocytosis	scavenger receptor	membrane	BU256170	263/295 (89%)
DR766245	7F3	Q86VW7	CSF1R	human	9e-16	phosphorylation	tyrosine kinase	membrane	BU432910	337/384 (87%)
DR764365	17B3	Q6YGU2	Toll-like receptor 2	rat	9e-06	receptor activity	N/A	membrane	XM_425176	165/200 (82%)
DR764879	22C1	Q9UGN4	CMRF35-H antigen	human	1e-13	cell adhesion	protein binding	membrane	NW_060341	72/84 (85%)
DR765080	24F4	P30533	Alpha-2-macroglobulin receptor	chicken	8e-74	cell proliferation	calcium ion binding	membrane	BU116249	557/621 (89%)

DR765308	27E6	Q8JHV3	endothelin receptor type B2	chicken	2e-40	GPCR signaling	protein binding	membrane	NW_060344	379/450 (84%)
DR765735	31H1	Q9DC83	Complement receptor Type 2	mouse	2e-09	B cell activation	protein binding	membrane	AJ727512	272/337 (80%)
DR765882	33F11	P16284	CD31	human	2e-09	cell recognition	protein binding	membrane	CK612996	303/334 (90%)
Lectin-like immuno-receptors										
DR783181	6D9	Q802S8	C-type lectin	chicken	1e-43	N/A	sugar binding	membrane	*N/A	*N/A
DR764124	14B7	Q9UMR7	C-type lectin	human	8e-9	immune response	sugar binding	membrane	*N/A	*N/A
DR783127	14D12	Q92478	C-type lectin	chicken	1e-43	N/A	sugar binding	membrane	*N/A	*N/A
DR783145	22G6	P37217	C-type lectin	chicken	1e-21	N/A	sugar binding	membrane	*N/A	*N/A
DR765730	31G3	Q5KU26	COLEC12	human	3e-14	phosphate transport	sugar binding	membrane	BU310599	324/383 (84%)
Cytokines and chemokines										
DR766004	4B6	O70460	CCL19	mouse	3e-13	chemotaxis	chemokine	extracellular	CF250494	323/379 (85%)
DR764376	17C4	P84444	CCL21	mouse	7e-09	chemotaxis	chemokine	extracellular	BG713657	200/228 (87%)
DR764436	18A2	O88803	Leukocyte chemotaxin 2	human	4e-34	chemotaxis	N/A	extracellular	BX272352	333/381 (87%)
DR764471	18D12	O55038	CXC13	mouse	1e-9	chemotaxis	chemokine	extracellular	BX278661	267/304 (87%)
DR765486	29G3	Q02960	MIF	chicken	8e-56	inflammation	cytokine	extracellular	BX270484	308/326 (94%)
Transcription factors involved in immune response										
DR766257	7G6	Q90871	ICSBP	chicken	6e-83	transcription regulation	DNA binding	cyto/nucl	AJ446205	578/647 (89%)
DR766345	8G9	P32519	Elf-1	human	4e-60	cytokine production	DNA binding	cyto/nucl	BU467484	584/631 (92%)
DR763870	10H12	Q922I8	Hcls1	mouse	1e-78	intracellular signaling	protein binding	N/A	AJ448508	562/642 (87%)
DR763973	12C9	Q91YV0	Tcf-4	mouse	5e-71	transcription regulation	DNA binding	cyto/nucl	BU266722	450/510 (88%)

DR763976	12D1	P25963	NF-kappaB inhibitor alpha	human	1e-45	apoptosis	TF binding	cytoplasm	AJ393886	460/509 (90%)
DR764044	13B8	Q80Z29	PBEF-1	rat	5e-11	cell proliferation	transferase	cyto/nucl	BU362525	470/502 (93%)
DR764083	13F3	P42224	STAT1	human	1e-92	caspase activation	transcription factor	cyto/nucl	BU144732	583/656 (88%)
DR764421	17G9	Q98TX7	IRF4	chicken	2e-47	T cell activation	DNA binding	cyto/nucl	CD218938	325/362 (89%)
DR764836	21G1	Q5EY36	Caterpiller 16.2	human	5e-67	transcription	ATP binding	cyto/nucl	BG711963	387/436 (88%)
Proteins involved in antigen processing and regulators of apoptosis										
DR765913	3A10	P43233	Cathepsin B	chicken	1e-102	proteolysiss	hydrolase	lysosome	BU491942	572/615 (93%)
DR783176	4D12	Q924M6	Apoptosis-inducing factor	rat	1e-56	apoptosis	oxidoreductase	mitochondrion	AJ729565	401/458 (87%)
DR766029	4F10	Q66K24	BNIP3	human	5e-47	apoptosis	protein binding	mitochondrion	CN218517	493/573 (86%)
DR766055	5A6	Q5EG05	CARD only protein	human	1e-07	apoptosis	caspase	intracellular	CD740540	123/149 (82%)
DR766371	9B4	Q90WU1	Caspase-8	chicken	1e-12	proteolysis	caspase	mitochondrion	BU307054	100/108 (92%)
DR764015	12G11	Q8NHS5	RNF151 protein	human	4e-26	apoptosis	metal ion binding	ubiquitin	BU474122	518/571 (90%)
DR764043	13B7	Q98TX3	Programmed cell death 4 protein	chicken	1e-50	N/A	protein binding	N/A	DR431508	313/361 (86%)
DR764288	16B6	Q07816	Apoptosis regulator BCL-X	chicken	4e-77	Apoptosis	protein binding	membrane	BU216611	323/372 (86%)
DR764444	18B2	Q8JGM8	BID	chicken	1e-30	apoptosis	death receptor binding	mitochondrion	AJ734385	109/124 (87%)
DR764820	21E9	O14727	Apoptotic protease activator	human	2e-12	apoptosis	caspase activator	cytosol	XM_416167	117/127 (92%)
DR764852	21H5	P25326	Cathepsin S	bovine	1e-69	proteolysis	hydrolase	lysosome	AJ719318	463/543 (85%)
DR765065	24E1	P04080	Cystatin B	human	1e-36	N/A	endopeptidase inhibitor	N/A	CO766372	246/273 (90%)
DR765125	25B8	Q80YE7	DAPK1	mouse	1e-51	apoptosis	ATP binding	cytoskeleton	AJ448497	379/424 (89%)
DR765256	26H4	Q8WUM4	Programmed cell death 6 protein	Human	1e-67	N/A	signal transducer	cytosol	AJ743022	467/501 (93%)

Innate immune effectors										
DR765554	2E4	Q9DEG0	Cremp	chicken	5e-17	N/A	N/A	N/A	BU212766	148/164 (90%)
DR763842	10F2	P10643	component C7	human	1e-75	complement activation	N/A	MAC	BU446496	269/304 (88%)
DR764075	13E7	P80391	Beta-defensin	turkey	2e-18	N/A	bactericidal	secreted	AY621318	68/76 (89%)
DR765706	31C5	Q8C669	Pellino protein homolog 1	mouse	2e-60	N/A	N/A	N/A	CO773876	412/429 (96%)
DR765888	33G5	P28799	Granulin precursor	human	1e-56	signal transduction	growth factor	N/A	BU297352	182/199 (91%)
Interferon inducible genes										
DR765199	26B8	P97371	PA28 alpha	mouse	3e-24	immune response	proteasome activator	cytosol	BM426178	179/221 (80%)
DR765660	30F9	Q9Z0E6	mGBP2	mouse	1e-37	immune response	nucleotide binding	N/A	CD734047	240/278 (86%)

Note: * N/A means either the chicken sequence is unavailable, or the two sequences cannot be aligned using *BLASTn* program.

* cyto/nucl: cytoplasm/nucleus; TF: transcription factor

3.1.4.1 Immune cell surface receptors

Members of this group are primarily expressed on the surface of different immune cell types. They participate in immune cell activation, proliferation, migration, and antigen recognition.

Clone 2D8 encoded a protein highly similar to the human NK cell tumor recognition protein. This molecule is a component of a putative tumor-recognition complex and participates in the protein folding process as a chaperone (Anderson *et al.*, 1993). Clone 2G8 was similar to the chicken CD79B (also known as B29). CD79B is a B-lineage specific member of the Ig superfamily. It is an essential member of a complex which is crucial for the correct assembly, transport and function of the B cell receptor (Wood *et al.*, 1993). Clone 5C1 was highly similar to the human MHC class II alpha chain. Clone 6H1 encoded β_2-microglobulin, the conserved part of the MHC class I molecule. Clone 7B3 was similar to the rat poly-Ig receptor. This receptor binds polymeric IgA and IgM at the basolateral surface of epithelial cells (Banting *et al.*, 1989). Clone 10C4 was similar to the human cytokine receptor common gamma chain which is a common component of the receptors for IL2, IL4, IL7, IL13, and IL21 (Nakarai *et al.*, 1994). The peptide predicted for clone 27D7 was very similar to the mouse lymphocyte antigen 6 complex. It participates in T cell activation and is up-regulated by TNF and IFN-gamma (Malek *et al.*, 1989). Clone 31H1 was similar to the complement receptor 2 (CR2/CD21), which is involved in B cell activation (Molina *et al.*, 1996). Clone

33B6 encoded a protein similar to the human CD82 antigen. It is a co-stimulatory protein for T cell activation (Lebel-Binay *et al.*, 1995). The peptides predicted for clone 5E12 was similar to the human colony stimulating factor 2 receptor beta chain (CSF2RB). It is also the common receptor beta chain for IL-3 and IL-5 (Hayashida *et al.*, 1990). CSF2RB is responsible for the common activities observed for these three cytokines. Clone 7F3 was similar to the human colony stimulating factor 1 receptor (CSF1R). CSF1R plays a critical role in regulating myeloid cell development, differentiation and expansion. CSF1R is also required for the optimal differentiation of DC *in vivo* (Macdonald *et al.*, 2005). Clone 5G1 encoded a protein fragment similar to the mouse KIAA (mKIAA) involved in receptor-mediated endocytosis (Okazaki *et al.*, 2003).

Clone 17B3 received a weak hit (9e-06) to the rat toll-like receptor 2 (TLR2). Alignment with TLR2 sequences from other species revealed that the predicted peptide contained a leucine rich repeat C-terminal (LRRCT) domain and a trans-membrane (TM) domain (**Figure 3.3**). TLR2 plays a central role in the recognition of Gram-positive bacteria and Mycobacteria. TLR2 knockout mice are selectively impaired in their ability to clear Gram positive bacteria (Underhill *et al.*, 1999). It constitutes a molecular link between microbial products, apoptosis, and host defense mechanisms (Aliprantis *et al.*, 1999).

Clone 22C1 encoded a protein similar to the human CMRF35. The CMRF family includes several Ig-like immunoregulatory signaling receptors primarily expressed

Figure 3.3

```
                  10        20        30        40        50        60        70
         ....|....|....|....|....|....|....|....|....|....|....|....|....|....|.
Duck     GTRALQALAATGCLRQGWPQHYACHSPPRYQGTLVRDVPASVLQCNRAAVLAPVCTALGLLCVAGAGWLV-
Chicken  .AG.V.....V.........G.V..A.A....V.L....T......P.........G.A.....V..G.L-
Rat      LLSFILERP.LVHVLVD..DS.L.D....LH.QRLQ.ARP...E.HQ..LVSG..C..-..LILLL.A.CY
Mouse    LLSFTMETP.LAQILVD..DS.L.D....LH.HRLQ.ARP...E.HQ..LVSG..C..-..LILLV.A.CH
Pig      FLSFT.GQQ.LAQVLSD..EN.L.D..SHVR.QR.Q.TRL.LTE.H.V..VSV..C..-F.LLLLT.A.CH
Human    FLSFT.EQQ.LAKVLID..AN.L.D..SHVR.QQ.Q..RL..SE.H.T.LVSGM.C..-F.LILLT.V.CH

                       LRRCT domain                              TM domain
```

Figure 3.3 Fragments of putative duck TLR2 aligned with TLR2 from other species. The gi numbers of the sequences used are chicken 50754697, rat 32737635, mouse 20140895, pig 35293573, and human 19718734. The multiple alignment was done by ClustalW with default parameters using the duck TLR2 as the reference sequence, and same residues in other sequences were replaced with dots. The boxes indicate the range of the predicted domains using the SMART program (http://smart.embl-heidelberg.de/). LRRCT: Leucine rich repeat C-terminal domain (residues 1-46), TM: transmembrane domain (residues 50-69).

on cells of the myeloid lineage (Clark *et al.*, 2001). This protein family includes the CLM receptors (CMRF-like receptors) in mouse (Chung *et al.*, 2003), and may include the novel Ig-like transcript (NILT) receptors in carp (Stet *et al.*, 2005). Clone 27E6 is similar to the chicken endothelin B receptor, a common receptor for endothelin 1, 2, and 3. The receptor functions via a typical GPCR pathway (Lecoin *et al.*, 1998). The protein predicted for clone 33F11 showed strong similarity to the human PECAM-1 (CD31). PECAM-1 is an adhesion molecule mainly expressed on platelets and at intercellular junctions of endothelial cells (Albelda *et al.*, 1991). It is implicated in leukocyte transendothelial migration.

3.1.4.2 Lectin- like immunoreceptors

Genes in this group encode a wide variety of calcium dependent carbohydrate-binding proteins, generally known as lectins. Many of them are implicated in the regulation of immune responses, reviewed by (Weis *et al.*, 1998).

Clones 6D9, 14D12 and 22G6 were found in the same cluster encoding C-type lectins. They showed similarity to the chicken gene 17.5, also known as Y-lec or CD69 (Bernot *et al.*, 1994; Rogers *et al.*, 2003). Mammalian CD69 is involved in lymphocyte proliferation and acts as a signal transmitting receptor for lymphocytes, NK cells, and platelets (Ziegler *et al.*, 1993). Clone 14B7 was similar to several lectins expressed on APC, including the DC immunoreceptor (DCIR) (Bates *et al.*, 1999). Clone 31G3 was similar to the human collectin subfamily member 12

(COLEC12), also known as the nurse cell scavenger receptor 2. COLEC12 recognizes oxidized phospholipids and helps remove damaged or apoptotic cells (Nakamura *et al.*, 2001).

3.1.4.3 Cytokines and chemokines

Genes in this group encoded proteins similar to cytokines and chemokines. They are involved in immune cell migration and intracellular communication between different cell types.

Clone 4B6 was similar to the mouse small inducible cytokine A19 (CCL19). CCL19 is secreted by DCs and is chemotactic for both näive CD4+ and CD8+ T cells. It also promotes encounters between DCs and circulating T cells (Ngo *et al.*, 1998). Clone 17C4 was similar to the mouse CCL21. CCL21 recruits T cells to the secondary lymphoid tissues (Gunn *et al.*, 1998). Both CCL19 and CCL21 are the ligands of the same receptor, CCR7 (Campbell *et al.*, 1998). Clone 18A2 received a strong hit to the human leukocyte chemotaxin 2 which is a positive regulator of chondrocyte proliferation (Yamagoe *et al.*, 1998). Clone 18D12 encoded a CXC chemokine similar to the mouse B lymphocyte chemoattractant (BLC, now called CXCL13). CXCL13 is strongly chemotactic for B cells. It directs the migration of B lymphocytes to follicles in secondary lymphoid organs (Gunn *et al.*, 1998). Clone 29G3 showed strong similarity to the chicken macrophage migration inhibitory factor (MIF). MIF is released primarily by myeloid cells in response to microbial stimuli, and promotes pro-inflammatory responses in both innate and adaptive immunity (Calandra and Roger, 2003).

3.1.4.4 Transcription factors involved in immune responses

Members of this group are involved in the regulation of immune gene expression. Some function by directly binding the promoter, enhancer or other regulatory regions of their target gene as transcription factors. Some affect the expression of their target genes indirectly by promoting or inhibiting the expression of other transcription factors.

Clones 7G6 and 17G9 encoded two different members of the IFN regulatory factor (IRF) family. Clone 7G6 was similar to the chicken interferon-consensus sequence binding protein (ICSBP, also known as IRF-8 for interferon regulatory factor 8). ICSBP was originally identified as a transcription factor recognizing the interferon stimulated response element (ISRE) located in the upstream regulatory region of IFN-gamma inducible MHC class I genes (Driggers *et al.*, 1990). ICSBP is involved in the differentiation of myeloid cells by driving their differentiation toward macrophages and plasmacytoid DCs (Tamura and Ozato, 2002; Tsujimura *et al.*, 2003). 17G9 encoded a protein similar to the chicken IFN regulatory factor 4 (IRF-4) (Matsuyama *et al.*, 1995). IRF-4 was shown to be able to bind the ISRE of the MHC class I promoter in vitro (Matsuyama *et al.*, 1995). The IRF-4 knockout mice were impaired in both B and T cell function (Mittrucker *et al.*, 1997). The protein encoded by clone 8G9 was similar to the human ETS-related transcription factor, Elf-1. Elf-1 is a transcription factor involved in the T cell receptor mediated trans-activation of HIV-2 gene expression (Leiden *et al.*, 1992). Clone 10H12 was similar to the mouse hematopoietic cell specific Lyn substrate 1 (Hcls1). Hcls1 is implicated in intracellular signal transduction for both clonal expansion and deletion in lymphoid cells (Taniuchi *et al.*, 1995). The peptide predicted for 12C9 was similar to the mouse immunoglobulin transcription factor 2 (also known as Tcf-4). Tcf-4 binds to the immunoglobulin enhancer Mu-E5/KE5-motif and also binds to the E-box that are present in the somatostatin receptor 2 initiator element (Pscherer *et al.*, 1996). Clone 12D1 showed significant identity to the human NF-kappa B inhibitor alpha which sequesters the NF-kappa B molecule in the cytosol and releases it after serine phosphorylation (Imbert *et al.*, 1996). Clone 13B8 encoded a protein similar to the pre-B cell colony enhancing factor (PBEF1). PBEF1 acts on early B-lineage precursor cells by enhancing the effect of IL-7 and stem cell factor (Samal *et al.*, 1994). Clone 13F3 encoded a protein similar to the human signal transducer and activator of transcription 1 (STAT1). STAT1 can be activated by IFN-alpha, IFN-gamma, IL6 and growth factors, and is considered to be a new molecular target for anti-inflammatory treatment, reviewed by (de Prati *et al.*, 2005). 21G1 was similar to the human caterpillar 16.2 related to the MHC class II transactivator CIITA (Harton *et al.*, 2002). Caterpillar 16.2 affects gene regulation through interfering with I kappa B degradation pathway and serves as a negative regulator of T cell activation (Conti *et al.*, 2005).

3.1.4.5 Proteins involved in antigen processing and apoptosis

Gene products in this group are part of the cellular machinery responsible for antigen processing and presentation. Some of them are also involved in apoptosis. Most gene products are enzymes such as proteases and kinases.

Clones 3A10, 23H6, and 21H5 showed strong similarities to several cathepsin family members. The cathepsin family is a group of lysosomal thiol proteases. Among them, 21H5 was most similar to the cathepsin S which is a key protease responsible for the proteolytic cleavage of MHC class II invariant chain, reviewed by (Hsing and Rudensky, 2005). Clone 24E1 was similar to the human cystatin B, which inhibits the protease activity of the members of the cathepsin family. Recent evidence showed it was also implicated in myoclonus epilepsy (Riccio *et al.*, 2001). Clone 25B8 was similar to a subunit of mouse death-associated protein kinase 1 (DAPK1). It is a pro-apoptotic serine/threonine protein kinase involved in both the Toll signaling pathway and the JAK-STAT pathway (**Figure 3.2**). Clone 4F10 encoded a protein similar to the human BNIP3. It belongs to the BCL-2 superfamily and is involved in T cell apoptosis (Lamy *et al.*, 2003). The gene product of clone 16B6 is similar to the chicken BCL-x involved in the repression of apoptosis (Boise *et al.*, 1993). Clone 9B4 encoded a protein similar to the human caspase-8. Caspase-8 participates in the caspase cascade responsible for death-receptor induced cell death, as well as antigen induced T cell proliferation, reviewed by (Newton and Strasser, 2003).

3.1.4.6 Innate immune effectors

Genes in this group mainly include members of the complement system and antimicrobial peptides (defensins). Clone 2E4 is most similar to the chicken membrane-associated complement regulatory membrane protein (Cremp) which protects cells against homologous complement (Inoue *et al.*, 2001). Clone 10F2 was similar to the human complement 7 (C7) precursors. It is a component of the membrane attack complex (MAC). C7 binds to C5b to form C5b-7 complex, serving as a membrane anchor (Thai and Ogata, 2004) in MAC. The peptide predicted for clone 13E7 is most similar to the turkey antimicrobial peptide THP1 precursor (beta-defensin). It is a cysteine-rich cationic low molecular weight peptide. *In vitro* experiments showed that it had bactericidal effects against both avian and human bacterial pathogens (Evans *et al.*, 1995). The protein predicted for clone 31C5 showed strong similarity to the mouse pellino protein homolog 1. Pellino functions as a scaffold protein involved in IL-1 signaling pathway (Jiang *et al.*, 2003). Clone 33G5 encoded a peptide similar to the human granulin identified in leukocytes (Bateman *et al.*, 1990).

3.1.4.7 IFN-inducible genes

Several aforementioned genes are under the influence of IFN, such as STAT1, ICSBP, IRFs, *etc.* These genes are well studied and their functions are discussed in their most appropriate functional groups. Genes included in this group mostly have rather unclear immune functions and their relevance to immunity is judged mainly by the observation that their gene expression was under the regulation of IFN. Clone 26B8 was similar to the mouse proteosome activator 28 (PA28) alpha. It is implicated in the proteasome assembly and efficient antigen processing (Yawata *et al.*, 2001). Clone 30F9 was similar to the mouse IFN induced guanylate binding protein-2 (mGBP2). These proteins are very abundant in IFN-gamma treated cells but their detailed functions are still unclear (Vestal *et al.*, 1998).

3.1.5 Sequence comparison between duck and chicken

In total, 684 unique sequences received matches in all three databases (the *Uniprot* database, the chicken genome database and the Gene Ontology database for molecular function), which permits further comparison between duck and chicken sequences based on GO categorizations. After removal of ribosomal sequences, immunoglobulin sequences, contig sequences and sequences with short aligned length (<100 base pairs), 425 sequences were left. Among them, 92 were immune-relevant.

The average sequence similarity between duck and chicken was 92.6% over a length of 90,947 aligned base pairs (average aligned length 214 bp per sequence). The average similarity of the immune relevant genes was calculated to be 91.1% over 19,246 aligned base pairs, as compared with 93.0% over 71,701 base pairs with the non-immune sequences. These genes were further divided into different functional groups based on GO slim categories. The average percent identities were calculated for each group (**Figure. 3.4**). Immune related genes, especially those involved in binding /recognition showed lower overall percent identities than other genes between chickens and ducks.

3.2 cDNA library arraying and screening

3.2.1 Library packaging and mass excision

To generate an unamplified cDNA library for arraying, an aliquot of the original library was repackaged and mass excised. The titre of the packaged phage library as evaluated by counting the PFU was about 1.2×10^3 / μl, with a total of 6.6×10^5 (final volume 550 μl). After mass excision, the titre of the excised plasmid library as evaluated by counting the CFU was about 1.2×10^2 / μl, with a total of 2.4×10^6 (final volume 20 ml).

3.2.2 Library picking and arraying

In order to generate an arrayed cDNA library for immune gene screening, 100 Q-trays were prepared and plated using 8~10 μl of the mass excised cDNA library (with densities about 1,800~2,200 colonies per Q-tray after 16~18 hours of incubation). Well-separated colonies were robotically picked into 357 384-well plates (1,000~1,400 colonies per Q-tray). The library was then printed onto

Figure 3.4

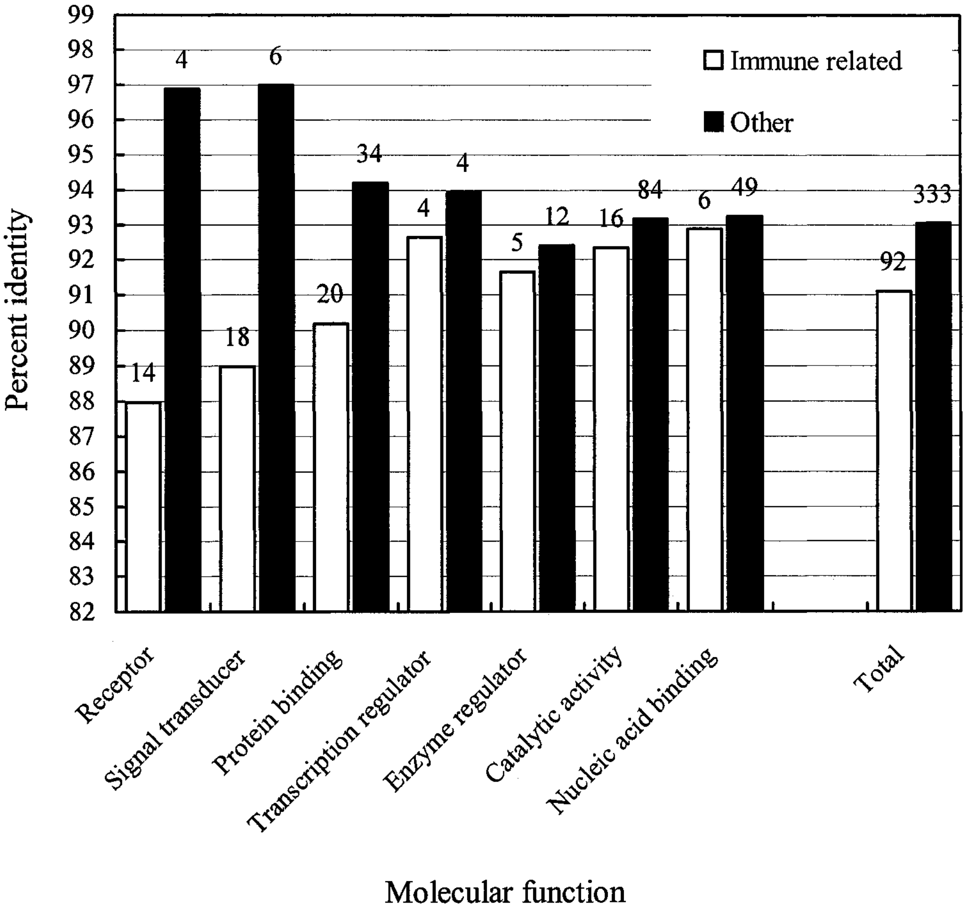


Figure 3.4 Sequence comparisons between duck and chicken in different functional groups. The sequences were classified based on selected categories in molecular function ontologies at the GO-slim level. The number of sequences included in each group is indicated on top of each histogram. The percent identity for each group was calculated using the sum of identical base pairs divided by the sum of base pairs that were aligned by the *BLASTn* program.

duplicate sets of nylon membranes (set A and B respectively, 14 membranes per set), with a density of 9,216 colonies (24×384-well plates) per membrane. Given that an average of ~25 wells in each plate failed to grow cells after overnight incubation, the total number of arrayed colonies was estimated about 1.28×10^5. Since there were also cells that failed to grow after they were printed on the membranes, the final number of arrayed clones suitable for hybridization was estimated to be around 1.2×10^5.

3.2.3 Library evaluation using Ig probes

To check the overall quality of the arrayed cDNA library, and to exclude the Ig clones in the subsequent screening, IgL and IgM *Overgo* probes were hybridized with the arrayed library. After hybridization, positive clones were found to be randomly distributed across all membranes with an average density of 4.88% (this number was obtained by counting one quarter of eight different membranes, the numbers of positive clones were 100, 121, 109, 135, 113, 94, 122, 106 each, with an average of 450 positive colonies per membrane).

3.2.4 Library screening for specific immune genes

To search for various immune related genes in the arrayed library, six pairs of ^{32}P-dCTP labeled *Overgo* probes for CSFR1, TLR7, CCR7, CCL19, beta-defensin, and

IFN-alpha were used to screen the membranes. In total, 75 positive clones were found and then sequenced for identification. The result was summarized in **Table 3.4.** The most abundant sequences were CSF1R transcripts. In total, 40 of them were identified. Among them, 32 were unique by examining the sequences at their 5' ends. Interestingly, the chicken CSF1R has not been identified in the NCBI database yet. The top hits for these 40 sequences were usually the chicken platelet-derived growth factor receptor beta, or PDGFRB (XP_414597), followed by CSF1R hits from other species. In mammals, PDGFRB is a paralog of CSF1R and they are located in tandem on the same chromosome (human on chromosome 5, mouse chromosome 18). Further examination revealed that the chicken PDGFRB is 9516 bp (XM_414596) long, which is approximately the combined length of human PDGFRB (NM_002609, 5718 bp) and CSF1R (NM_005211, 3985 bp). As indicated in **Figure 3.5**, the first section (before 5,500 bp) of the chicken PDGFRB was highly similar to PDGFRB of human and mouse, while the remaining part showed significant identity with the CSF1R sequences. This indicated that either these two genes were fused together in chicken, or there is a mistake in the annotation. Given the current annotation status of the chicken genome, the latter is likely to be the case.

The high similarity to the supposed chicken PDGFRB at the DNA level makes it possible to serve as a reference sequence to map our CSF1R sequences. In contrast, the similarity to mammalian CSF1R were only detectable at protein level and with

Table 3.4 Library screening for specific immune genes

Genes	Clones
CSF1R	18B20, 33G13, 44E6, 44E21, 44O24, 49A4, 50B13, 50N13, 70A9, 70N11, 80I2, 102H8, 105O4, 107M14, 112A6, 130J12, 136N9, 143N6, 149B15, 158E23, 166I22, 174F14, 195G15, 212B12, 226O20, 236K16, 242E8, 245H7, 275M11, 283N1, 287P14, 292I9, 292P3, 297C11, *6M18, *13G2, *29O14, *26A5, *42M9, *45B15
Beta-defensin	81F10, 99L9, 109B16, 122A14, 127M9, 140M7, 142F15, 151B6, 187E15, 198B17, 198J8, 213P10, 227C18, 239K22, 275K6, 296P5, *35B4, *45A21
CCL19	109O21, *17F21
TLR7	130K16
CCR7	140I24
IFN-alpha	N/A
[1]Alpha Integrin	183A9, 187J15
[2]IgL	158H1
[3]Unknown	56I19, 56F8, 89G18, 94F22, 112M7,*29G16

1. Alpha Integrin was found to have 10 consecutive nucleic acid matches to CCL19 probe.
2. Although IgL was found to have 12 exact nucleic acid matches to CSF1R probe, it is more likely that this was the result of contamination.
3. Unknown sequences received no hits in the database search and none of the probe sequences used was identified in them.

Figure 3.5

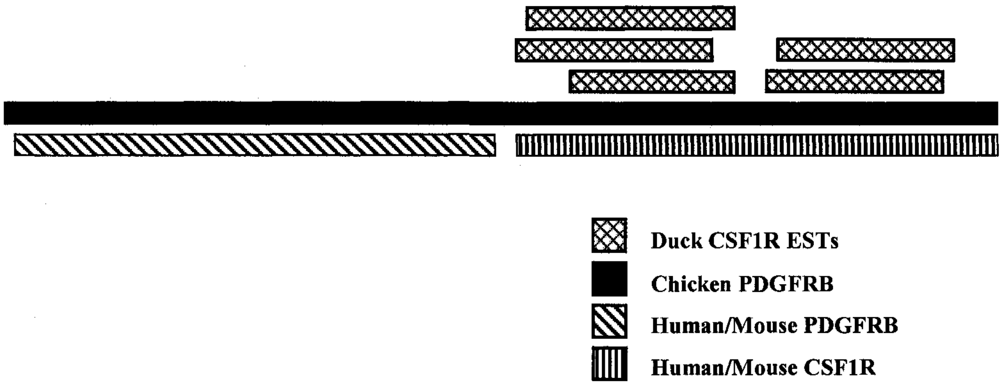


Figure 3.5 Schematic alignments of duck CSF1R ESTs, human/mouse PDGFRB and CSF1R genes, and chicken PDGFRB genes (in approximate proportion). Duck PDGFRB: XM_414596, human PDGFRB: NM_002609, mouse Pdgfrb: NM_008809; human CSF1R: NM_005211; mouse Csf1r: NM_001037859.

only a few exact matches. It is impossible to align with human and mouse CSF1R at the DNA level except for a very short fragment. The 40 CSF1R sequences were then clustered and assembled. The 3 contigs were mapped to the chicken PDGFRB at 80~90 % identity over 1000 bp with E values around 0.0. Not surprisingly, all of the sequences were mapped to the section of chicken PDGFRB after 5,500 bp (Clone *42M9 was the 5' most), which corresponds exactly to the section aligned with CSF1R from other species, further confirming that the chicken PDGFRB might be annotated incorrectly and contains both PDGFRB and CSF1R.

The second most abundant collection of sequences was the beta-defensin. In total, 18 of them were identified in the screening. Except for clone 151B6, which received a weak hit (E value = 0.042) to chicken beta-defensin, all other sequences received no significant hits (E value < 0.1) in the database search. Their identity was determined by identifying the probe sequence used for beta-defensin.

3.3 Library screening with SSH probes

In order to improve the rate of immune gene discovery, a SSH procedure was performed between the duck spleen and brain tissue. Membrane 11A was chosen to test the synthesized SSH probes enriched for spleen specific transcripts. The hybridization pattern showed much stronger background noise compared to that using Overgo probes. In total, 72 positive clones were identified. These clones were further sequenced for their identities. 50 of them generated good quality sequence and were annotated. To compare with the gene discovery efficiency by random sequencing, 24 clones were randomly picked and sequenced. 17 of them yielded good quality data and were annotated. The results were summarized in **Table 3.5** and **Table 3.6.**

As indicated from these two tables, clones that hybridized with SSH probes were mainly involved in protein synthesis and energy metabolism. In addition, they were severely biased toward transcripts of translation elongation factors which constituted almost half of the positive clones. Other sequence species were poorly represented. In contrast, sequences identified by random sequencing showed a much more even representation of different mRNA species.

Table 3.5 Library screening using SSH probes

Putative ID	**Clone #**
Translation elongation factor	22
60S ribosomal protein	3
40S ribosomal protein	2
Cytochrome subunits	2
Ferritin H	2
Ig light chain	2
Nucleophosmin	2
Beta actin	1
Ubiquitin	1
ATPase	1
Aldolase	1
Unknown	11

Table 3.6 Clones identified by random sequencing

Putative ID	**Clone #**
60S ribosomal protein	3
Beta actin	2
C-type lectin	2
Ferritin H	2
Ig heavy chain	2
Hemoglobin	1
Copine III	1
Unknown	4

Chapter 4

DISCUSSION

4.1 Evaluation of the EST analysis pipeline

Since the introduction of the concept of expressed sequence tagging, the number of EST projects has continued to increase every year. They have greatly accelerated the pace of gene discovery and contributed significantly to genome assembly and gene annotation. During the past decade, however, efforts and resources have gradually shifted from EST projects towards large-scale genome sequencing projects whose promises and challenges had drawn most of the public attention. Although EST projects have a longer history, programs and services available for EST analysis are limited compared to the ones devoted for genome sequencing. As a result, EST analysis and annotation still remains a relatively involved and onerous task.

After I surveyed the available resources and tested some of the software packages for EST analysis, I found either these programs did not meet our needs, or they required computing environments I could not meet. Therefore, I decided to build my own EST analysis pipeline around several basic programs mentioned in the Introduction (section **1.1**). This approach allowed me to tailor the program to our particular situation (i.e. limited computing power, customized analysis needs). In addition, I was also able to try some relatively new algorithms that were recently developed in the field.

Because of the limitations of the computing power, a significant portion of my efforts were spent on tailoring the program according to our particular situations. For example, since the cDNA fragments were directionally cloned into the vectors, the majority of ESTs should be on the forward strands of the clones. Therefore, instead of using *BLASTX* which first performs a six-frame translation and then searches the translation sets against the database, I first did a three-forward reading frame translation and then searched the database using *BLASTP*. Sequences with hits in more than one reading frame indicated the occurrence of frame-shift errors which are quite common in EST sequencing. They were further analyzed using the *FASTy* program, which was optimized for nucleotide sequences with frame shift errors (Pearson *et al.*, 1997). Only sequences that received no hits were subjected to reverse translation in case of reversal during the cloning process.

This annotation pipeline was a compromise between speed and sensitivity, with considerations to both the EST specific problems (i.e. frame shift errors) and our custom requirements (i.e. GO annotation for functional analysis in different categories). In our case, the motif search against a series of databases using *RPS-BLAST* did not add many new hits on top of the sequence search results. Since I was unable to run *Interproscan* which was suggested to be more sensitive, it might be possible that this program might yield more information. In the future, if the computational resources were not a serious limitation factor, this part could be replaced with a simple *BLASTX* followed by *Interproscan*. This simplification would greatly reduce the program complexity and improve its robustness and maintainability.

The *BLAST* results were parsed to return the five most "informative" hits for each query EST sequence. These five hits included the top hits, hits from human, mouse or rat, and chicken. Hits from other species were also included if they were within the top five *BLAST* hits and their annotations did not contain words like "hypothetical", "chromosome", or "unknown". The default tentative annotation was assigned by the top hit. This is neither always correct nor always the most informative one. Although a close neighbor to duck, the chicken genome is not as well annotated as that of human and mouse. Many top hits from chicken were annotated as "hypothetical protein". In these cases, manual inspection of all the five hits was done to choose the most informative annotation.

In addition, several customized analyses were also performed using *MySQL* database. Gene Ontology and KEGG pathway assignments are relatively simple operations. They were primarily based on the information from the *Uniprot* annotations. GO assignments were achieved by mapping the *Uniprot* accession numbers to the corresponding GO ID using three tables. KEGG pathway annotation was done by parsing the EC numbers for pathway database assignment using only one table. The sequence comparison between the chicken and duck required more complex operations. In total, nine tables were built to enable this fine-grained analysis.

4.2 Duck EST annotations and immune gene analysis

The original input was 2,900 raw EST sequences. After quality checking and data cleaning, they were subjected to clustering and assembly. This process attempts to reduce the data redundancy by putting all the ESTs derived from the same gene into one cluster and generating one contig to represent all of them (in case of inconsistency caused by alleles, gene rearrangements, somatic hypermutations, or sequence errors, more than one contig may be generated). With supervised clustering, it is possible to reach the ideal result. However, without reference sequences, the ESTs are clustered mainly based on their similarity to each other. Thus it is possible that two unique sequences might be derived from the same gene if there is no overlap between them. This process generated 1,885 unique sequences.

In the subsequent EST annotation via homology search, 1,284 of them received annotations. Finally, all ESTs (including both the annotated and un-annotated sequences) were submitted to the NCBI *dbEST*, and 2,725 were accepted. Prior to this project, the *dbEST* database contained only 234 duck sequence entries, thus our EST data add significantly to the genetic information available for this organism, allowing further study of avian immunology.

EST data also provided a coarse picture about the expression profile in the normal duck spleen. From our data, except for beta-actins and ribosomal sequences that were highly expressed on almost all tissue types, the most abundant transcripts were different types of Ig sequences, sequences for MHC assembly, and sequences that are abundantly expressed in erythrocytes like ferritin H and hemoglobin. These data faithfully reflected the functions of the spleen as both an immune organ and a blood reservoir.

After EST annotation, the results were mined for immune relevant genes, which yielded 208 unique genes involved in immunity. These immune-relevant genes are distributed through all levels of the immune system, including immune cell surface receptors, lectin-like immunoreceptors, cytokines and chemokines, transcription factors involved in immune responses, genes involved in antigen processing and regulators of apoptosis, and innate immune effectors. Compared to immune genes discovered in other EST projects (in shrimp 44 out 2,045, mosquito 38 out of 5,925, Tsetse fly 78 out of 215,427, chicken 80 out of 5,251), this number was very high. One major reason is that they mainly considered pathogen recognition molecules and immune effectors that directly participate in immune response, while we used a much broader definition of immune relevancy (Kelley *et al.*, 2005). The other contributing factor is that as vertebrates, ducks possess a much more complicated immune system than invertebrate organisms under study in those EST projects.

Comparison of homologous sequences between duck and chicken revealed the overall lower similarities of immune relevant genes than other non-immune ones. In each functional category of the immune genes, the average identity is also different. This indicates that while all parts of the immune system come under selection from pathogens, some components are under greater selection pressure and diverge more rapidly. As indicated in **Figure 3.3**, cell surface receptors diverge fastest. From the cell surface to the cytoplasm and into the cell nucleus, the percent identity scores increase steadily, reflecting different selection pressure imposed on these sequences. Possibly some parts of the cellular machinery are easier targets for pathogens while other parts (i.e. transcription factors) are under more functional constraints. The alternative might be the result of "neutral drift"- if the genes are not essential for the survival of the organism, they are under less selection pressure, thus accumulating more mutations than other more essential functional groups during the same time period.

Admittedly, the calculation of sequence similarities based on *BLAST* alignment is certainly an overestimate, since the *BLAST* program only tries to align the most conserved regions (local alignment) between the two sequences being compared. A better way would involve using the overall optimal alignment between the two homologous sequences (global alignment). However, given the incomplete nature of EST data, our *BLAST* approach is currently the only feasible way.

Based on the data collected from this study, there are significant sequence similarities (average 92.6%) over long stretches (average 214 bp per aligned sequences) between homologous sequences from duck and chicken. The majority of them shared regions with 90% nucleotide identity, with the average identity for immune genes about 91.1% and non-immune being about 93.0%, respectively. Therefore the vast majority of reagents and resources available for the chicken, including both the chicken EST libraries (Abdrakhmanov *et al.*, 2000; Tirunagaru *et al.*, 2000; Boardman *et al.*, 2002) and the chicken genome sequences (Hillier *et al.*, 2004) should be very useful for the identification and study of duck immunity. For example, probes could be designed based on the conserved regions of the chicken homologous genes to screen the arrayed cDNA library.

Several recent reports showed that it was possible to study gene expression using microarrays constructed from a closely related species (Cros *et al.*, 1999; Huang *et al.*, 2000; Medhora *et al.*, 2002; Moody *et al.*, 2002; Chalmers *et al.*, 2005). However, there is no consensus on the degree of sequence similarity required for successful cross-species hybridizations. A study using a human microarray to analyze pig immune gene expression, suggested that 84% similarity could produce consistent results (Moody *et al.*, 2002). Therefore, it is very likely that the spotted cDNA microarrays based on chicken sequences (Liu *et al.*, 2001; Cogburn *et al.*, 2003; Burnside *et al.*, 2005) could be used to study duck gene expression profiles during immune challenges. One risk associated with this approach is that some immune genes are evolving very fast (i.e. the C-type lectins) and they might be at the limit of the detection.

4.3 Evaluations of the arrayed cDNA library

The overall quality of the arrayed cDNA library was first evaluated by hybridizing the membranes with Ig probes. Positive clones were found to be randomly distributed across all 14 membranes with an average density about 4.88%. Compared to ~15% Ig transcripts based on the EST data (obtained by combining the sequence numbers of Ig light chain and heavy chain), this value based on membrane hybridization was probably an underestimate.

One possible reason is that since the cDNA library was not a full-length cDNA library, a significant proportion of the inserts will be only partial sequences of the corresponding genes. Therefore, it is likely that some probes will fail to detect some Ig clones on the membrane. In addition, the two methods used to determine clone identity (sequencing v.s. hybridization) do not have the same sensitivity and specificity. The former is the definitely the gold standard but laborious, while the later method often suffers from the effects such as washing/hybridization conditions, DNA secondary structures, steric hindrances, surface effects, which might interfere with the binding efficiency of the probes to their targets.

Further library screening using Overgo probes for several immune related genes was also quite successful. Among the six immune genes that were probed, five of them were identified by sequencing the positive clones. These included CSF1R, beta-defensin, TLR7, CCR7, and CCL19. Analysis of multiple clones corresponding to CSFR1 and beta-defensin showed that the majority of these clones were unique, reflecting the low redundancy of the arrayed library. In addition, my analysis also suggested that the current annotation for chicken PDGFRB might be incorrect. It seems to contain both the PDGERB and CSF1R genes in its sequence.

4.4 Problems associated with SSH-PCR generated probes

The attempt to use the SSH generated probes for library screening to improve the rate of immune gene discovery was unsuccessful. Two possible reasons can account for this failure. One possible reason might be the quality of the kit itself, since the control provided with the kit did not seem to work, which made it quite difficult to check the results to optimize the protocols in major steps.

The other cause of the failure was probably the tissue choice. Brain tissue might be so different from spleen tissue such that a lot of the differentially expressed genes under normal conditions are not immune related. Using the *Digital Differential Display* (DDD) techniques at NCBI *dbEST* (http://www.ncbi.nlm.nih.gov/UniGene/info_ddd.shtml) on published EST libraries, I conducted a virtual subtraction between a pool of human spleen and brain EST libraries randomly sequenced under normal conditions. The top ten differentially expressed genes were listed in **Table 4.1**.

Table 4.1 Virtual EST subtraction between human spleen and brain tissue. The top 10 statistically differentially expressed genes between human spleen and brain tissues as calculated with Digital Differential Display (DDD) at NCBI dbEST. The Library IDs used for brains were 18415, 18353, 18352, 18318, 18317, 13865, 18466, 18467, 1750, 16390, 16376, and 1749 with a total number of 163,729 EST. The Library IDs used for spleens were 18474 and 16431 with a total number of 33,359 EST.

	Abundance (per 10,000)		Differentially expressed
	Spleen	Brain	
Translation elongation factor	2131	342	1789
Biglycan (BGN)	1553	27	1526
Invariant chain	1313	113	1200
Ubiquitin C	1517	516	1001
Heat shock protein 70	1103	139	964
Beta actin	890	213	677
Granulin	713	41	672
MHC class I	495	1	494
Vimentin	522	49	473
Complement component 1	163	19	144

As the data suggested, many of these highly differentially expressed genes were not involved in immunity. As a result, the sequences remained after subtraction for these differentially expressed genes will still be significantly more abundant than the expression level of most immune related genes (usually less than one copy per 1,000). Moreover, the subsequent two rounds of PCR will amplify this gap "exponentially". In addition, the total transcripts under examination will also be greatly amplified, which requires more work to get a representative sampling. In microarray experiments, when the starting mRNA is very rare (i.e. samples from tumor biopsies), the common practice is to use T7 polymerase for linear amplification instead of PCR.

Most of the published applications of this technique were used to identify differentially expressed genes between the same cell line at different developmental stages or the same tissue type under different experimental conditions (i.e. normal vs. immune-challenged). The PCR-Select kit appears to be unsuitable for subtraction between two highly different tissue types such as spleen and brain.

The *GeneSeeker* kit (*Genetix*) (Rast *et al.*, 2000) can help ameliorate such problems. The kit is specially designed to identify the low prevalent and differentially expressed genes from the arrayed cDNA library. The protocol is essentially derived from the SSH protocol. The main idea is still subtraction followed by amplification. However, several steps are introduced (i.e. size selection, a limited cycle of linear amplification) to reduce hybridization background, and to prevent the subtracted probe pools from being dominated by a few highly abundant but differentially expressed transcripts.

4.5 Future directions

My research project focused on EST analysis and immune gene identification from our duck EST project. Significant progress was made in many aspects. However, from the perspective of looking for a more general solution for immune gene screening or EST analysis, there are several aspects that can be improved.

4.5.1 *In silico* immune gene screening

The main obstacle to immune gene screening is the lack of a standard definition for immune relevance. Even for the well-studied human or mouse genes, different researchers will probably produce different lists of immune genes based on their field of interest, although many should overlap. As a result, there is no gold standard to serve as a reference to make the classification. Currently, the solution is manual inspection on the annotated genes as was done in the immune gene identification from our EST data. However, by relaxing the requirements, several computational methods can be performed to greatly reduce the searching space before requiring human inspection. This issue can be tackled in three ways.

The first step involves building a dictionary of "common immunological terms", which can be collected from the public databases (i.e. NCBI *EntrezGene*), literature (i.e. *Journal of Immunology, Nature Immunology, and Immunobiology*), textbooks, *etc.* The annotated sequences can be screened for the presence of these terms and classified accordingly. The IRIS (for Immunogenetic Related Information Source) database created by Dr. John Trowsdale at the University of Cambridge has collected 1,562 immune genes from human (Kelley *et al.*, 2005), which provides a good starting point.

The second way is to harness the power of Gene Ontology annotations which enable functional sequence analysis in a high throughput manner. The GO schema organizes genes naturally into their respective biological niches and greatly reduces the complexity of the data. The GO annotation also captures basic characteristics for each group of immune genes. For example, immune cell surface receptors are usually expressed on the cell surface (component: membrane), function through binding their ligands (function: peptide/sugar binding) and then trigger the downstream reactions (process: receptor activity/signal transduction); in contrast, cytokines and chemokines are usually secreted to the outside (component: extracellular) to recruit their target cells (process: chemotaxis). Several ontology groups are more enriched with immune related genes (**Figure 3.1**). This suggests another way to automate immune gene screening based on their GO annotation. By fine-tuning different query combinations or using more specific GO terms, one can quickly narrow down the search space and identify genes acting at different level of immunity.

The last one is to explore the protein domains/motifs that are often associated with immune functions (i.e. Ig superfamily (IgSF) domain, leucine rich repeats (LRR), death domain (DD), TIR domain).

These three aspects can be combined using a simple voting system with different weights associated with different sources of evidence. If a sequence was annotated (i.e. by *BLAST* search), with GO term assigned and/or domains identified (i.e. through *InterproScan*), a score can be calculated based on the evidence collected. A final ranked list can be produced and ready to be inspected by the researcher.

4.5.2 Ideas to improve the EST analysis pipeline

The EST analysis pipeline I built for our duck EST project could potentially serve as a general analysis and annotation platform for similar tasks. To make it more useful to bench biologists with limited bioinformatics skills, the pipeline can be improved in at least three ways.

The first is to improve its user interface. Currently, the program is run from a command line on a Linux terminal, which requires a basic understanding of both the Linux system and the command (shell) language. By building a graphical user interface (GUI), this program can be made more user-friendly and intuitive to use. There are two ways to achieve this, one is to build a desktop GUI (i.e. using *Java Swing*) while the other is to build a web-based service. Given the resource intensive-nature of the task, the latter is a more attractive solution. With a web service, users do not have to worry about software installation and the requirement for a powerful computer. In addition, web-based services have the advantage of centralized management, which makes the program much easier to maintain, update and upgrade.

The second is to improve the program performance by adopting distributed computing techniques. Even with a powerful computer such as a server machine, the local *BLAST* and *Interproscan* is still a very resource-intensive and time consuming task, especially when several users submit large numbers of sequences at the same time. A common approach in this situation is to split the job into a set of sub-jobs and distribute them to different computers (computer nodes). Each computer only works on its own portion of the task. Finally, all the finished jobs are assembled together and ready for the next step. The key to implement a distributed computing environment is a batch queuing system or a job scheduler. There are several popular packages available. Among them, the Java Sun Grid Engine (SGE, http://www.sun.com/software/gridware/), Condor (http://www.cs.wisc.edu/condor/), the Portable Batch System (PBS http://www.openpbs.org/), and the LSF Parallel System (http://www.platform.com/) are most popular.

The third way to improve this analysis pipeline is to add some new features. The pipeline only provides starting materials for more in-depth analyses by researchers. Accordingly, this program can be enhanced and extended to provide more functions.

For example, if the genome sequence of the organism under study or its closely related species is available, to enable chromosome mapping for the selected EST will be very desirable (i.e. in the form of NCBI *Map Viewer*, http://www.ncbi.nlm.nih.gov/mapview/static/MVstart.html). In addition, it is often the case that the primary purpose of the EST project is to identify certain genes of interest or to look for novel members of gene families (i.e. TLR, lectin family, CHIR), rather than just sequencing and meta-analysis of the EST data. Therefore, to enable a *BLAST*-based search engine coupled with a more sensitive domain identification tool (i.e. *HMMer*) against researcher's local database will be very valuable (i.e. in the form of the proWeb tool, *Do-It-Yourself WU-BLAST* http://www.proweb.org/Tools/WU-blast.html). This way, new sequences produced can be compared with all others immediately, and the database is maintained and curated by the researcher's lab and annotated incrementally.

4.6 Conclusions

As the natural host of Influenza A virus and an important animal model for Hepatitis B infection, the duck receives increasing attention from immunologists, virologists, and epidemiologists. However, the knowledge and reagents for studying duck immunity are seriously lacking. My research project and other ongoing programs in our lab represent an important attempt toward elucidating the immune system of this organism.

In conclusion, my research project was focused on EST analysis and immune gene identification. To this end, a multi-pronged approach was adopted. Firstly, I analyzed ~3,000 sequences generated from our EST project and identified 208 immune relevant genes. During the process, I built a high-throughput EST analysis pipeline for EST assembly, annotation, and publication. To facilitate further immune gene discovery, over 120,000 cDNA clones were arrayed, and screened for several specific immune genes. Analyzing 425 pairs of homologous genes between chicken and duck showed high sequence similarity between these two species, suggesting that chicken resources (such as EST libraries and microarrays) could potentially be used for the study of duck immunity.

APPENDIX A

A short discussion on *Linux, Perl,* and *MySQL* in high-throughput sequence analysis

A.1 The Linux operating system

The Linux (http://www.linux.org) is an open source Unix-style operating system originally developed by Linus Torvalds in 1991. The operating system is comprised of three components: the kernel, the shell, and the applications. The kernel is the heart of the operating system. It controls all the accesses to the computer resources (i.e. CPU, memory, file system, *etc.*). The shell or command interpreter acts as an interface between the kernel and the user at the terminal. The applications or utilities are tools for other operations (i.e. text editors, internet browsers, databases). Linux is free, open source, highly customizable and offers excellent support for software development and programming. These features have made Linux operating system the *de facto* platform for bioinformatics research. Most bioinformatics software tools were implemented under the Linux platform or otherwise have a Linux compatible version.

A.2 *Perl* and *Bioperl*

Perl (Practical Extraction and Report Language) is a high-level programming language originally developed by Larry Wall in 1987. *Perl* is an interpreted language and can run on any computer installed with a *Perl* interpreter for that particular operating system (Linux and Mac OS X come with *Perl* installed, Windows users can download and install *ActivePerl* from http://www.activestate.com). Although *Perl* is not a full-fledged programming language like *Java* and *C++*, it is extremely powerful for text processing. Since most sequence data is in text format, *Perl* is usually a natural choice for sequence manipulation and analysis.

Bioperl (http://www.bioperl.org) is a collection of *Perl* modules that facilitate the development of *Perl* scripts for bioinformatics applications. With *Bioperl*, many routine and tedious tasks such as sequence retrieval, sequence comparison and data parsing can be reduced to several lines of commands. *Bioperl* greatly accelerates project development, reduces the time spent in debugging, and results in a more maintainable code base. *Bioperl* can be freely obtained and installed using CPAN (Comprehensive *Perl* Archive Network http://search.cpan.org/dist/bioperl).

A.3 Structured Query Language *(SQL)* and *MySQL*

High throughput sequence analysis usually entails managing a large amount of data. Therefore, it is often necessary to use a database for data storage, retrieval and analysis. Relational databases and the query language *SQL* are a commonly used solution in such cases. The most popular open source relational database is probably

MySQL (http://www.mysql.com), which is free for academic users.

The *Perl DBI* modules provide the ability to automate database operations using Perl scripts. In addition, the *phpMyAdmin* (http://www.phpmyadmin.net), offers an excellent graphical user interface for managing and browsing *MySQL* tables and displaying query results via a web browser. These tools combined together make *MySQL* administration much easier and more productive.

A.4 Using *Perl* for sequence analysis

I chose *Perl* as the programming language to build the pipeline for EST analysis. An alternative choice is *Java*. In addition to its power in text manipulations, there are other good reasons in favor of *Perl* with regard to this particular project. First, for small to medium project, programming with *Perl* is usually faster than *Java* due to the much stricter typing of the latter; Secondly, *BioPerl* is currently more comprehensive and better supported than *BioJava*; Thirdly, *Perl's* learning curve is very shallow and a lot of tasks can be done with a little knowledge on *Perl* language, while programming with *Java* usually requires a good understanding on object oriented programming (OOP); Lastly, *Perl* is also the *de facto* programming language for CGI (common gateway interfaces) scripts used for generating dynamic web pages, which makes it easy to port the *Perl* scripts into web service.

I wrote numerous Perl scripts (available by request from the author by contacting kmagor@ualberta.ca) to reformat data as inputs for sequence analysis software, and to parse and extract the most essential data from the software's outputs. Several *BioPerl* modules, such as *Bio::SeqIO* for sequence manipulation, *Bio::SearchIO* for *BLAST* result parsing, and *DBI::mysql* for interfacing with *MySQL* database were often used. Among these scripts I wrote, several of them may be of more general usage: the script that reads all the raw sequence files in a folder (with or without subfolders) and merges them into a single multi-FASTA file, the BLAST parser that returns the top five most informative hits, the script that automatically loads the downloaded Gene Ontology files into the *MySQL* database, the script that automatically maps the *Uniprot* accession numbers to the corresponding GO terms, the script that expands the clusters generated by *CAP3* into individual ESTs, and the script that formats the EST data into *dbEST* publication format, *etc*. Some more complicated analyses such as the homologous gene comparisons between the duck and chicken within different GO categories proved to be very difficult to put into a script and run in an automatic way. They were basically achieved by issuing *SQL* commands and examining the returned results interactively.

Several *Shell* (bash) commands also prove quite useful in text manipulation, such as using *cat* to merge files, using *grep* for regular expression, and using *more/head/tail* to check the content of large files without loading the whole documents into the memory.

BIBLIOGRAPHY

Abdrakhmanov, I., D. Lodygin, P. Geroth, H. Arakawa, A. Law, J. Plachy, B. Korn and J. M. Buerstedde (2000). "A large database of chicken bursal ESTs as a resource for the analysis of vertebrate gene function." Genome Res **10**(12): 2062-9.

Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno and et al. (1991). "Complementary DNA sequencing: expressed sequence tags and human genome project." Science **252**(5013): 1651-6.

Albelda, S. M., W. A. Muller, C. A. Buck and P. J. Newman (1991). "Molecular and cellular properties of PECAM-1 (endoCAM/CD31): a novel vascular cell-cell adhesion molecule." J Cell Biol **114**(5): 1059-68.

Alder, M. N., I. B. Rogozin, L. M. Iyer, G. V. Glazko, M. D. Cooper and Z. Pancer (2005). "Diversity and function of adaptive immune receptors in a jawless vertebrate." Science **310**(5756): 1970-3.

Aliprantis, A. O., R. B. Yang, M. R. Mark, S. Suggett, B. Devaux, J. D. Radolf, G. R. Klimpel, P. Godowski and A. Zychlinsky (1999). "Cell activation and apoptosis by bacterial lipoproteins through toll-like receptor-2." Science **285**(5428): 736-9.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Andersen, M. H., D. Schrama, P. Thor Straten and J. C. Becker (2006). "Cytotoxic T cells." J Invest Dermatol **126**(1): 32-41.

Anderson, S. K., S. Gallinger, J. Roder, J. Frey, H. A. Young and J. R. Ortaldo (1993). "A cyclophilin-related protein involved in the function of natural killer cells." Proc Natl Acad Sci U S A **90**(2): 542-6.

Apweiler, R., A. Bairoch and C. H. Wu (2004). "Protein sequence databases." Curr

Opin Chem Biol **8**(1): 76-80.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Attwood, T. K. (2002). "The PRINTS database: a resource for identification of protein families." Brief Bioinform **3**(3): 252-63.

Baeuerle, P. A. and D. Baltimore (1996). "NF-kappa B: ten years after." Cell **87**(1): 13-20.

Banchereau, J. and R. M. Steinman (1998). "Dendritic cells and the control of immunity." Nature **392**(6673): 245-52.

Banting, G., B. Brake, P. Braghetta, J. P. Luzio and K. K. Stanley (1989). "Intracellular targetting signals of polymeric immunoglobulin receptors are highly conserved between species." FEBS Lett **254**(1-2): 177-83.

Barreda, D. R., P. C. Hanington and M. Belosevic (2004). "Regulation of myeloid development and function by colony stimulating factors." Dev Comp Immunol **28**(5): 509-54.

Bateman, A., D. Belcourt, H. Bennett, C. Lazure and S. Solomon (1990). "Granulins, a novel class of peptide from leukocytes." Biochem Biophys Res Commun **173**(3): 1161-8.

Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats and S. R. Eddy (2004). "The Pfam protein families database." Nucleic Acids Res **32**(Database issue): D138-41.

Bates, E. E., N. Fournier, E. Garcia, J. Valladeau, I. Durand, J. J. Pin, S. M. Zurawski, S. Patel, J. S. Abrams, S. Lebecque, P. Garrone and S. Saeland (1999). "APCs express DCIR, a novel C-type lectin surface receptor containing an immunoreceptor tyrosine-based inhibitory motif." J Immunol

163(4): 1973-83.

Batzoglou, S., D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. P. Mesirov and E. S. Lander (2002). "ARACHNE: a whole-genome shotgun assembler." Genome Res **12**(1): 177-89.

Behrens, G., M. Li, C. M. Smith, G. T. Belz, J. Mintern, F. R. Carbone and W. R. Heath (2004). "Helper T cells, dendritic cells and CTL Immunity." Immunol Cell Biol **82**(1): 84-90.

Bernot, A., R. Zoorob and C. Auffray (1994). "Linkage of a new member of the lectin supergene family to chicken Mhc genes." Immunogenetics **39**(4): 221-9.

Bischoff, S. C., M. Krieger, T. Brunner, A. Rot, V. von Tscharner, M. Baggiolini and C. A. Dahinden (1993). "RANTES and related chemokines activate human basophil granulocytes through different G protein-coupled receptors." Eur J Immunol **23**(3): 761-7.

Boardman, P. E., J. Sanz-Ezquerro, I. M. Overton, D. W. Burt, E. Bosch, W. T. Fong, C. Tickle, W. R. Brown, S. A. Wilson and S. J. Hubbard (2002). "A comprehensive collection of chicken cDNAs." Curr Biol **12**(22): 1965-9.

Boehm, U., T. Klamp, M. Groot and J. C. Howard (1997). "Cellular responses to interferon-gamma." Annu Rev Immunol **15**: 749-95.

Boise, L. H., M. Gonzalez-Garcia, C. E. Postema, L. Ding, T. Lindsten, L. A. Turka, X. Mao, G. Nunez and C. B. Thompson (1993). "bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death." Cell **74**(4): 597-608.

Bonaldo, M. F., G. Lennon and M. B. Soares (1996). "Normalization and subtraction: two approaches to facilitate gene discovery." Genome Res **6**(9): 791-806.

Bos, J. L. (1989). "ras oncogenes in human cancer: a review." Cancer Res **49**(17): 4682-9.

Burnside, J., P. Neiman, J. Tang, R. Basom, R. Talbot, M. Aronszajn, D. Burt and J. Delrow (2005). "Development of a cDNA array for chicken gene expression analysis." BMC Genomics **6**(1): 13.

Calandra, T. and T. Roger (2003). "Macrophage migration inhibitory factor: a regulator of innate immunity." Nat Rev Immunol **3**(10): 791-800.

Camon, E., M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez and R. Apweiler (2004). "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology." Nucleic Acids Res **32**(Database issue): D262-6.

Campbell, J. J., E. P. Bowman, K. Murphy, K. R. Youngman, M. A. Siani, D. A. Thompson, L. Wu, A. Zlotnik and E. C. Butcher (1998). "6-C-kine (SLC), a lymphocyte adhesion-triggering chemokine expressed by high endothelium, is an agonist for the MIP-3beta receptor CCR7." J Cell Biol **141**(4): 1053-9.

Chalmers, A. D., K. Goldstone, J. C. Smith, M. Gilchrist, E. Amaya and N. Papalopulu (2005). "A *Xenopus tropicalis* oligonucleotide microarray works across species using RNA from *Xenopus laevis*." Mech Dev **122**(3): 355-63.

Chou, H. H. and M. H. Holmes (2001). "DNA sequence quality trimming and vector removal." Bioinformatics **17**(12): 1093-104.

Chung, D. H., M. B. Humphrey, M. C. Nakamura, D. G. Ginzinger, W. E. Seaman and M. R. Daws (2003). "CMRF-35-like molecule-1, a novel mouse myeloid receptor, can inhibit osteoclast formation." J Immunol **171**(12): 6541-8.

Clark, G. J., B. Cooper, S. Fitzpatrick, B. J. Green and D. N. Hart (2001). "The gene encoding the immunoregulatory signaling molecule CMRF-35A localized to human chromosome 17 in close proximity to other members of the CMRF-35 family." Tissue Antigens **57**(5): 415-23.

Cocchi, F., A. L. DeVico, A. Garzino-Demo, S. K. Arya, R. C. Gallo and P. Lusso (1995). "Identification of RANTES, MIP-1 alpha, and MIP-1 beta as the major HIV-suppressive factors produced by CD8+ T cells." Science **270**(5243): 1811-5.

Cogburn, L. A., X. Wang, W. Carre, L. Rejto, T. E. Porter, S. E. Aggrey and J. Simon (2003). "Systems-wide chicken DNA microarrays, gene expression profiling, and discovery of functional genes." Poult Sci **82**(6): 939-51.

Conti, B. J., B. K. Davis, J. Zhang, W. O'Connor, Jr., K. L. Williams and J. P. Ting (2005). "CATERPILLER 16.2 (CLR16.2), a novel NBD/LRR family member that negatively regulates T cell function." J Biol Chem **280**(18): 18375-85.

Cooper, M. A., T. A. Fehniger and M. A. Caligiuri (2001). "The biology of human natural killer-cell subsets." Trends Immunol **22**(11): 633-40.

Costa-Pereira, A. P., T. M. Williams, B. Strobl, D. Watling, J. Briscoe and I. M. Kerr (2002). "The antiviral response to gamma interferon." J Virol **76**(18): 9060-8.

Cros, N., J. Muller, S. Bouju, G. Pietu, C. Jacquet, J. J. Leger, J. F. Marini and C. A. Dechesne (1999). "Upregulation of M-creatine kinase and glyceraldehyde3-phosphate dehydrogenase: two markers of muscle disuse." Am J Physiol **276**(2 Pt 2): R308-16.

Davis, M. M. and P. J. Bjorkman (1988). "T-cell antigen receptor genes and T-cell recognition." Nature **334**(6181): 395-402.

de Prati, A. C., A. R. Ciampa, E. Cavalieri, R. Zaffini, E. Darra, M. Menegazzi, H. Suzuki and S. Mariotto (2005). "STAT1 as a new molecular target of anti-inflammatory treatment." Curr Med Chem **12**(16): 1819-28.

Dimopoulos, G., T. L. Casavant, S. Chang, T. Scheetz, C. Roberts, M. Donohue, J. Schultz, V. Benes, P. Bork, W. Ansorge, M. B. Soares and F. C. Kafatos (2000). "Anopheles gambiae pilot gene discovery project: identification of mosquito innate immunity genes from expressed sequence tags generated from immune-competent cell lines." Proc Natl Acad Sci U S A **97**(12): 6619-24.

Driggers, P. H., D. L. Ennist, S. L. Gleason, W. H. Mak, M. S. Marks, B. Z. Levi, J. R. Flanagan, E. Appella and K. Ozato (1990). "An interferon gamma-regulated protein that binds the interferon-inducible enhancer element of major histocompatibility complex class I genes." Proc Natl Acad Sci U S A **87**(10): 3743-7.

Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-63.

Engering, A. J., M. Cella, D. Fluitsma, M. Brockhaus, E. C. Hoefsmit, A. Lanzavecchia and J. Pieters (1997). "The mannose receptor functions as a high capacity and broad specificity antigen receptor in human dendritic cells." Eur J Immunol **27**(9): 2417-25.

Evans, E. W., F. G. Beach, K. M. Moore, M. W. Jackwood, J. R. Glisson and B. G. Harmon (1995). "Antimicrobial activity of chicken and turkey heterophil peptides CHP1, CHP2, THP1, and THP3." Vet Microbiol **47**(3-4): 295-303.

Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res **8**(3): 186-94.

Ewing, B., L. Hillier, M. C. Wendl and P. Green (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Res **8**(3): 175-85.

Forster, R., A. Schubel, D. Breitfeld, E. Kremmer, I. Renner-Muller, E. Wolf and M. Lipp (1999). "CCR7 coordinates the primary immune response by establishing functional microenvironments in secondary lymphoid organs." Cell **99**(1): 23-33.

Foster, W. K., D. S. Miller, C. A. Scougall, I. Kotlarski, R. J. Colonno and A. R. Jilbert (2005). "Effect of antiviral treatment with entecavir on age- and dose-related outcomes of duck hepatitis B virus infection." J Virol **79**(9): 5819-32.

Geijtenbeek, T. B., D. S. Kwon, R. Torensma, S. J. van Vliet, G. C. van Duijnhoven, J. Middel, I. L. Cornelissen, H. S. Nottet, V. N. KewalRamani, D. R. Littman, C. G. Figdor and Y. van Kooyk (2000). "DC-SIGN, a dendritic cell-specific HIV-1-binding protein that enhances trans-infection of T cells." Cell **100**(5): 587-97.

Granucci, F., C. Vizzardelli, E. Virzi, M. Rescigno and P. Ricciardi-Castagnoli (2001). "Transcriptional reprogramming of dendritic cells by differentiation stimuli." Eur J Immunol **31**(9): 2539-46.

Gross, P. S., T. C. Bartlett, C. L. Browdy, R. W. Chapman and G. W. Warr (2001). "Immune gene discovery by expressed sequence tag analysis of hemocytes and hepatopancreas in the Pacific White Shrimp, Litopenaeus vannamei, and the Atlantic White Shrimp, L. setiferus." Dev Comp Immunol **25**(7): 565-77.

Gunn, M. D., V. N. Ngo, K. M. Ansel, E. H. Ekland, J. G. Cyster and L. T. Williams (1998). "A B-cell-homing chemokine made in lymphoid follicles activates Burkitt's lymphoma receptor-1." Nature **391**(6669): 799-803.

Gunn, M. D., K. Tangemann, C. Tam, J. G. Cyster, S. D. Rosen and L. T. Williams (1998). "A chemokine expressed in lymphoid high endothelial venules promotes the adhesion and chemotaxis of naive T lymphocytes." Proc Natl Acad Sci U S A **95**(1): 258-63.

Harton, J. A., M. W. Linhoff, J. Zhang and J. P. Ting (2002). "Cutting edge: CATERPILLER: a large family of mammalian genes containing CARD, pyrin, nucleotide-binding, and leucine-rich repeat domains." J Immunol **169**(8): 4088-93.

Hayashida, K., T. Kitamura, D. M. Gorman, K. Arai, T. Yokota and A. Miyajima (1990). "Molecular cloning of a second subunit of the receptor for human granulocyte-macrophage colony-stimulating factor (GM-CSF): reconstitution of a high-affinity GM-CSF receptor." Proc Natl Acad Sci U S A **87**(24): 9655-9.

Heath, W. R., G. T. Belz, G. M. Behrens, C. M. Smith, S. P. Forehan, I. A. Parish, G. M. Davey, N. S. Wilson, F. R. Carbone and J. A. Villadangos (2004). "Cross-presentation, dendritic cell subsets, and the generation of immunity to cellular antigens." Immunol Rev **199**: 9-26.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci U S A **89**(22): 10915-9.

Hillier, L. W., W. Miller, E. Birney, W. Warren, R. C. Hardison, C. P. Ponting, P. Bork, D. W. Burt, M. A. Groenen, M. E. Delany, J. B. Dodgson, A. T. Chinwalla, P. F. Cliften, S. W. Clifton, K. D. Delehaunty, C. Fronick, R. S. Fulton, T. A. Graves, C. Kremitzki, D. Layman, V. Magrini, J. D. McPherson, T. L. Miner, P. Minx, W. E. Nash, M. N. Nhan, J. O. Nelson, L. G. Oddy, C. S. Pohl, J. Randall-Maher, S. M. Smith, J. W. Wallis, S. P. Yang, M. N. Romanov, C. M. Rondelli, B. Paton, J. Smith, D. Morrice, L. Daniels, H. G. Tempest, L. Robertson, J. S. Masabanda, D. K. Griffin, A. Vignal, V. Fillon, L. Jacobbson, S. Kerje, L. Andersson, R. P. Crooijmans, J. Aerts, J. J.

van der Poel, H. Ellegren, R. B. Caldwell, S. J. Hubbard, D. V. Grafham, A. M. Kierzek, S. R. McLaren, I. M. Overton, H. Arakawa, K. J. Beattie, Y. Bezzubov, P. E. Boardman, J. K. Bonfield, M. D. Croning, R. M. Davies, M. D. Francis, S. J. Humphray, C. E. Scott, R. G. Taylor, C. Tickle, W. R. Brown, J. Rogers, J. M. Buerstedde, S. A. Wilson, L. Stubbs, I. Ovcharenko, L. Gordon, S. Lucas, M. M. Miller, H. Inoko, T. Shiina, J. Kaufman, J. Salomonsen, K. Skjoedt, G. K. Wong, J. Wang, B. Liu, J. Wang, J. Yu, H. Yang, M. Nefedov, M. Koriabine, P. J. Dejong, L. Goodstadt, C. Webber, N. J. Dickens, I. Letunic, M. Suyama, D. Torrents, C. von Mering, E. M. Zdobnov, K. Makova, A. Nekrutenko, L. Elnitski, P. Eswara, D. C. King, S. Yang, S. Tyekucheva, A. Radakrishnan, R. S. Harris, F. Chiaromonte, J. Taylor, J. He, M. Rijnkels, S. Griffiths-Jones, A. Ureta-Vidal, M. M. Hoffman, J. Severin, S. M. Searle, A. S. Law, D. Speed, D. Waddington, Z. Cheng, E. Tuzun, E. Eichler, Z. Bao, P. Flicek, D. D. Shteynberg, M. R. Brent, J. M. Bye, E. J. Huckle, S. Chatterji, C. Dewey, L. Pachter, A. Kouranov, Z. Mourelatos, A. G. Hatzigeorgiou, A. H. Paterson, R. Ivarie, M. Brandstrom, E. Axelsson, N. Backstrom, S. Berlin, M. T. Webster, O. Pourquie, A. Reymond, C. Ucla, S. E. Antonarakis, M. Long, J. J. Emerson, E. Betran, I. Dupanloup, H. Kaessmann, A. S. Hinrichs, G. Bejerano, T. S. Furey, R. A. Harte, B. Raney, A. Siepel, W. J. Kent, D. Haussler, E. Eyras, R. Castelo, J. F. Abril, S. Castellano, F. Camara, G. Parra, R. Guigo, G. Bourque, G. Tesler, P. A. Pevzner, A. Smit, L. A. Fulton, E. R. Mardis and R. K. Wilson (2004). "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution." Nature **432**(7018): 695-716.

Hoffmann, J. A., F. C. Kafatos, C. A. Janeway and R. A. Ezekowitz (1999). "Phylogenetic perspectives in innate immunity." Science **284**(5418): 1313-8.

Hsing, L. C. and A. Y. Rudensky (2005). "The lysosomal cysteine proteases in MHC class II antigen presentation." Immunol Rev **207**: 229-41.

Huang, G. S., S. M. Yang, M. Y. Hong, P. C. Yang and Y. C. Liu (2000). "Differential gene expression of livers from ApoE deficient mice." Life Sci **68**(1): 19-28.

Huang, X. and A. Madan (1999). "CAP3: A DNA sequence assembly program." Genome Res **9**(9): 868-77.

Hulo, N., C. J. Sigrist, V. Le Saux, P. S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. De Castro, P. Bucher and A. Bairoch (2004). "Recent improvements to the PROSITE database." Nucleic Acids Res **32**(Database issue): D134-7.

Hulse-Post, D. J., K. M. Sturm-Ramirez, J. Humberd, P. Seiler, E. A. Govorkova, S. Krauss, C. Scholtissek, P. Puthavathana, C. Buranathai, T. D. Nguyen, H. T. Long, T. S. Naipospos, H. Chen, T. M. Ellis, Y. Guan, J. S. Peiris and R. G. Webster (2005). "Role of domestic ducks in the propagation and biological evolution of highly pathogenic H5N1 influenza viruses in Asia." Proc Natl Acad Sci U S A **102**(30): 10682-7.

Huson, D. H., K. Reinert, S. A. Kravitz, K. A. Remington, A. L. Delcher, I. M. Dew, M. Flanigan, A. L. Halpern, Z. Lai, C. M. Mobarry, G. G. Sutton and E. W. Myers (2001). "Design of a compartmentalized shotgun assembler for the human genome." Bioinformatics **17 Suppl 1**: S132-9.

Imbert, V., R. A. Rupec, A. Livolsi, H. L. Pahl, E. B. Traenckner, C. Mueller-Dieckmann, D. Farahifar, B. Rossi, P. Auberger, P. A. Baeuerle and J. F. Peyron (1996). "Tyrosine phosphorylation of I kappa B-alpha activates NF-kappa B without proteolytic degradation of I kappa B-alpha." Cell **86**(5): 787-98.

Inoue, J., T. Ishida, N. Tsukamoto, N. Kobayashi, A. Naito, S. Azuma and T. Yamamoto (2000). "Tumor necrosis factor receptor-associated factor (TRAF) family: adapter proteins that mediate cytokine signaling." Exp Cell Res **254**(1): 14-24.

Inoue, N., A. Fukui, M. Nomura, M. Matsumoto, Y. Nishizawa, K. Toyoshima and T. Seya (2001). "A novel chicken membrane-associated complement regulatory protein: molecular cloning and functional characterization." J Immunol **166**(1): 424-31.

Jameson, J., K. Ugarte, N. Chen, P. Yachi, E. Fuchs, R. Boismenu and W. L. Havran (2002). "A role for skin gammadelta T cells in wound repair." Science **296**(5568): 747-9.

Jiang, Z., H. J. Johnson, H. Nie, J. Qin, T. A. Bird and X. Li (2003). "Pellino 1 is required for interleukin-1 (IL-1)-mediated signaling through its interaction with the IL-1 receptor-associated kinase 4 (IRAK4)-IRAK-tumor necrosis factor receptor-associated factor 6 (TRAF6) complex." J Biol Chem **278**(13): 10952-6.

Jilbert, A. R., J. A. Botten, D. S. Miller, E. M. Bertram, P. M. Hall, J. Kotlarski and C. J. Burrell (1998). "Characterization of age- and dose-related outcomes of duck hepatitis B virus infection." Virology **244**(2): 273-82.

Kadowaki, N., S. Antonenko, J. Y. Lau and Y. J. Liu (2000). "Natural interferon alpha/beta-producing cells link innate and adaptive immunity." J Exp Med **192**(2): 219-26.

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno and M. Hattori (2004). "The KEGG resource for deciphering the genome." Nucleic Acids Res **32**(Database issue): D277-80.

Kelley, J., B. de Bono and J. Trowsdale (2005). "IRIS: a database surveying known human immune system genes." Genomics **85**(4): 503-11.

Kim, C. H. (2004). "Chemokine-chemokine receptor network in immune cell trafficking." Curr Drug Targets Immune Endocr Metabol Disord **4**(4): 343-61.

King, C., R. Mueller Hoenger, M. Malo Cleary, K. Murali-Krishna, R. Ahmed, E. King and N. Sarvetnick (2001). "Interleukin-4 acts at the locus of the antigen-presenting dendritic cell to counter-regulate cytotoxic CD8+ T-cell responses." Nat Med **7**(2): 206-14.

Koide, S. L., K. Inaba and R. M. Steinman (1987). "Interleukin 1 enhances T-dependent immune responses by amplifying the function of dendritic cells." J Exp Med **165**(2): 515-30.

Konig, R., L. Y. Huang and R. N. Germain (1992). "MHC class II interaction with CD4 mediated by a region analogous to the MHC class I binding site for CD8." Nature **356**(6372): 796-8.

Koopmann, W., C. Ediriwickrema and M. S. Krangel (1999). "Structure and function of the glycosaminoglycan binding site of chemokine macrophage-inflammatory protein-1 beta." J Immunol **163**(4): 2120-7.

Kristiansen, K. (2004). "Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function."

Pharmacol Ther **103**(1): 21-80.

Lamy, L., M. Ticchioni, A. K. Rouquette-Jazdanian, M. Samson, M. Deckert, A. H. Greenberg and A. Bernard (2003). "CD47 and the 19 kDa interacting protein-3 (BNIP3) in T cell apoptosis." J Biol Chem **278**(26): 23915-21.

Lanier, L. L. (2005). "NK cell recognition." Annu Rev Immunol **23**: 225-74.

Le Guerhier, F., A. Thermet, S. Guerret, M. Chevallier, C. Jamard, C. S. Gibbs, C. Trepo, L. Cova and F. Zoulim (2003). "Antiviral effect of adefovir in combination with a DNA vaccine in the duck hepatitis B virus infection model." J Hepatol **38**(3): 328-34.

Le Page, C., P. Genin, M. G. Baines and J. Hiscott (2000). "Interferon activation and innate immunity." Rev Immunogenet **2**(3): 374-86.

Lebel-Binay, S., C. Lagaudriere, D. Fradelizi and H. Conjeaud (1995). "CD82, member of the tetra-span-transmembrane protein family, is a costimulatory protein for T cell activation." J Immunol **155**(1): 101-10.

Lecoin, L., T. Sakurai, M. T. Ngo, Y. Abe, M. Yanagisawa and N. M. Le Douarin (1998). "Cloning and characterization of a novel endothelin receptor subtype in the avian class." Proc Natl Acad Sci U S A **95**(6): 3024-9.

Lederer, J. A., V. L. Perez, L. DesRoches, S. M. Kim, A. K. Abbas and A. H. Lichtman (1996). "Cytokine transcriptional events during helper T cell subset differentiation." J Exp Med **184**(2): 397-406.

Lehane, M. J., S. Aksoy, W. Gibson, A. Kerhornou, M. Berriman, J. Hamilton, M. B. Soares, M. F. Bonaldo, S. Lehane and N. Hall (2003). "Adult midgut expressed sequence tags from the tsetse fly Glossina morsitans morsitans and expression analysis of putative immune response genes." Genome Biol **4**(10): R63.

Lehrer, R. I. and T. Ganz (2002). "Defensins of vertebrate animals." Curr Opin Immunol **14**(1): 96-102.

Leiden, J. M., C. Y. Wang, B. Petryniak, D. M. Markovitz, G. J. Nabel and C. B.

Thompson (1992). "A novel Ets-related transcription factor, Elf-1, binds to human immunodeficiency virus type 2 regulatory elements that are required for inducible trans activation in T cells." J Virol **66**(10): 5890-7.

Lindstedt, M., B. Johansson-Lindbom and C. A. Borrebaeck (2002). "Global reprogramming of dendritic cells in response to a concerted action of inflammatory mediators." Int Immunol **14**(10): 1203-13.

Liu, H. C., H. H. Cheng, V. Tirunagaru, L. Sofer and J. Burnside (2001). "A strategy to identify positional candidate genes conferring Marek's disease resistance by integrating DNA microarrays and genetic mapping." Anim Genet **32**(6): 351-9.

Locksley, R. M., N. Killeen and M. J. Lenardo (2001). "The TNF and TNF receptor superfamilies: integrating mammalian biology." Cell **104**(4): 487-501.

Lopez-Botet, M., M. Carretero, J. Perez-Villar, T. Bellon, M. Llano and F. Navarro (1997). "The CD94/NKG2 C-type lectin receptor complex: involvement in NK cell-mediated recognition of HLA class I molecules." Immunol Res **16**(2): 175-85.

Lundqvist, M. L., D. L. Middleton, C. Radford, G. W. Warr and K. E. Magor (2006). "Immunoglobulins of the non-galliform birds: antibody expression and repertoire in the duck." Dev Comp Immunol **30**(1-2): 93-100.

Ma, B., J. Tromp and M. Li (2002). "PatternHunter: faster and more sensitive homology search." Bioinformatics **18**(3): 440-5.

Macdonald, K. P., V. Rowe, H. M. Bofinger, R. Thomas, T. Sasmono, D. A. Hume and G. R. Hill (2005). "The Colony-Stimulating Factor 1 Receptor Is Expressed on Dendritic Cells during Differentiation and Regulates Their Expansion." J Immunol **175**(3): 1399-405.

Malek, T. R., K. M. Danis and E. K. Codias (1989). "Tumor necrosis factor synergistically acts with IFN-gamma to regulate Ly-6A/E expression in T lymphocytes, thymocytes and bone marrow cells." J Immunol **142**(6): 1929-36.

Marra, M. A., T. A. Kucaba, L. W. Hillier and R. H. Waterston (1999). "High-throughput plasmid DNA purification for 3 cents per sample." Nucleic Acids Res **27**(24): e37.

Matrosovich, M., N. Zhou, Y. Kawaoka and R. Webster (1999). "The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distinguishable properties." J Virol **73**(2): 1146-55.

Matsuyama, T., A. Grossman, H. W. Mittrucker, D. P. Siderovski, F. Kiefer, T. Kawakami, C. D. Richardson, T. Taniguchi, S. K. Yoshinaga and T. W. Mak (1995). "Molecular cloning of LSIRF, a lymphoid-specific member of the interferon regulatory factor family that binds the interferon-stimulated response element (ISRE)." Nucleic Acids Res **23**(12): 2127-36.

Matsuzaki, K., K. Sugishita, N. Fujii and K. Miyajima (1995). "Molecular basis for membrane selectivity of an antimicrobial peptide, magainin 2." Biochemistry **34**(10): 3423-9.

Matzinger, P. (1994). "Tolerance, danger, and the extended family." Annu Rev Immunol **12**: 991-1045.

Medhora, M., M. Bousamra, 2nd, D. Zhu, L. Somberg and E. R. Jacobs (2002). "Upregulation of collagens detected by gene array in a model of flow-induced pulmonary vascular remodeling." Am J Physiol Heart Circ Physiol **282**(2): H414-22.

Medzhitov, R. and C. A. Janeway, Jr. (1997). "Innate immunity: the virtues of a nonclonal system of recognition." Cell **91**(3): 295-8.

Mesa, C. M., K. J. Thulien, D. A. Moon, S. M. Veniamin and K. E. Magor (2004). "The dominant MHC class I gene is adjacent to the polymorphic TAP2 gene in the duck, *Anas platyrhynchos*." Immunogenetics **56**(3): 192-203.

Mewes, H. W., D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd and B. Weil (2002). "MIPS: a database for genomes and protein sequences." Nucleic Acids Res **30**(1): 31-4.

Mittrucker, H. W., T. Matsuyama, A. Grossman, T. M. Kundig, J. Potter, A. Shahinian, A. Wakeham, B. Patterson, P. S. Ohashi and T. W. Mak (1997).

"Requirement for the transcription factor LSIRF/IRF4 for mature B and T lymphocyte function." Science **275**(5299): 540-3.

Molina, H., V. M. Holers, B. Li, Y. Fung, S. Mariathasan, J. Goellner, J. Strauss-Schoenberger, R. W. Karr and D. D. Chaplin (1996). "Markedly impaired humoral immune response in mice deficient in complement receptors 1 and 2." Proc Natl Acad Sci U S A **93**(8): 3357-61.

Moody, D. E., Z. Zou and L. McIntyre (2002). "Cross-species hybridisation of pig RNA to human nylon microarrays." BMC Genomics **3**(1): 27.

Mueller, A. P., H. R. Wolfe and R. K. Meyer (1960). "Precipitin production in chickens. XXI. Antibody production in bursectomized chickens and in chickens injected with 19-nortestosterone on the fifth day of incubasion." J Immunol **85**: 172-9.

Nakamura, K., H. Funakoshi, K. Miyamoto, F. Tokunaga and T. Nakamura (2001). "Molecular cloning and functional characterization of a human scavenger receptor with C-type lectin (SRCL), a novel member of a scavenger receptor family." Biochem Biophys Res Commun **280**(4): 1028-35.

Nakarai, T., M. J. Robertson, M. Streuli, Z. Wu, T. L. Ciardelli, K. A. Smith and J. Ritz (1994). "Interleukin 2 receptor gamma chain expression on resting and activated lymphoid cells." J Exp Med **180**(1): 241-51.

Natarajan, K., N. Dimasi, J. Wang, R. A. Mariuzza and D. H. Margulies (2002). "Structure and function of natural killer cell receptors: multiple molecular solutions to self, nonself discrimination." Annu Rev Immunol **20**: 853-85.

Neves, S. R., P. T. Ram and R. Iyengar (2002). "G protein pathways." Science **296**(5573): 1636-9.

Newton, K. and A. Strasser (2003). "Caspases signal not only apoptosis but also antigen-induced activation in cells of the immune system." Genes Dev **17**(7): 819-25.

Ngo, V. N., H. L. Tang and J. G. Cyster (1998). "Epstein-Barr virus-induced molecule 1 ligand chemokine is expressed by dendritic cells in lymphoid tissues and strongly attracts naive T cells and activated B cells." J Exp Med

188(1): 181-91.

O'Donovan, C., M. J. Martin, A. Gattiker, E. Gasteiger, A. Bairoch and R. Apweiler (2002). "High-quality protein knowledge resource: SWISS-PROT and TrEMBL." Brief Bioinform **3**(3): 275-84.

O'Garra, A. and N. Arai (2000). "The molecular basis of T helper 1 and T helper 2 cell differentiation." Trends Cell Biol **10**(12): 542-50.

Okazaki, N., R. Kikuno, R. Ohara, S. Inamoto, H. Aizawa, S. Yuasa, D. Nakajima, T. Nagase, O. Ohara and H. Koga (2003). "Prediction of the coding sequences of mouse homologues of KIAA gene: II. The complete nucleotide sequences of 400 mouse KIAA-homologous cDNAs identified by screening of terminal sequences of cDNA clones randomly sampled from size-fractionated libraries." DNA Res **10**(1): 35-48.

Pancer, Z., C. T. Amemiya, G. R. Ehrhardt, J. Ceitlin, G. L. Gartland and M. D. Cooper (2004). "Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey." Nature **430**(6996): 174-80.

Parker, D. C. (1993). "T cell-dependent B cell activation." Annu Rev Immunol **11**: 331-60.

Pearson, W. R. (1990). "Rapid and sensitive sequence comparison with FASTP and FASTA." Methods Enzymol **183**: 63-98.

Pearson, W. R., T. Wood, Z. Zhang and W. Miller (1997). "Comparison of DNA sequences with protein sequences." Genomics **46**(1): 24-36.

Pertea, G., X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai and J. Quackenbush (2003). "TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets." Bioinformatics **19**(5): 651-2.

Pimentel-Muinos, F. X. and B. Seed (1999). "Regulated commitment of TNF receptor signaling: a molecular switch for death or activation." Immunity **11**(6): 783-93.

Pscherer, A., U. Dorflinger, J. Kirfel, K. Gawlas, J. Ruschoff, R. Buettner and R. Schule (1996). "The helix-loop-helix transcription factor SEF-2 regulates the activity of a novel initiator element in the promoter of the human somatostatin receptor II gene." Embo J **15**(23): 6680-90.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler and R. Lopez (2005). "InterProScan: protein domains identifier." Nucleic Acids Res **33**(Web Server issue): W116-20.

Rast, J. P., G. Amore, C. Calestani, C. B. Livi, A. Ransick and E. H. Davidson (2000). "Recovery of developmentally defined gene sets from high-density cDNA macroarrays." Dev Biol **228**(2): 270-86.

Reedijk, M., X. Liu, P. van der Geer, K. Letwin, M. D. Waterfield, T. Hunter and T. Pawson (1992). "Tyr721 regulates specific binding of the CSF-1 receptor kinase insert to PI 3'-kinase SH2 domains: a model for SH2-mediated receptor-target interactions." Embo J **11**(4): 1365-72.

Reynaud, C. A., V. Anquez, H. Grimal and J. C. Weill (1987). "A hyperconversion mechanism generates the chicken light chain preimmune repertoire." Cell **48**(3): 379-88.

Riccio, M., R. Di Giaimo, S. Pianetti, P. P. Palmieri, M. Melli and S. Santi (2001). "Nuclear localization of cystatin B, the cathepsin inhibitor implicated in myoclonus epilepsy (EPM1)." Exp Cell Res **262**(2): 84-94.

Richmond, A. (2002). "Nf-kappa B, chemokine gene transcription and tumour growth." Nat Rev Immunol **2**(9): 664-74.

Rogers, S., I. Shaw, N. Ross, V. Nair, L. Rothwell, J. Kaufman and P. Kaiser (2003). "Analysis of part of the chicken Rfp-Y region reveals two novel lectin genes, the first complete genomic sequence of a class I alpha-chain gene, a truncated class II beta-chain gene, and a large CR1 repeat." Immunogenetics **55**(2): 100-8.

Ross, M. T., S. LaBrie, J. McPherson and V. P. Stanton (1999). "Current Protocols in Human Genetics." John Wiley and Sons **33**(18): 5.6.1-5.6.5.

Rothman, P., S. C. Li and F. W. Alt (1989). "The molecular events in heavy chain class-switching." Semin Immunol **1**(1): 65-77.

Ruepp, A., A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G. Mannhaupt, M. Munsterkotter and H. W. Mewes (2004). "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes." Nucleic Acids Res **32**(18): 5539-45.

Samal, B., Y. Sun, G. Stearns, C. Xie, S. Suggs and I. McNiece (1994). "Cloning and characterization of the cDNA encoding a novel human pre-B-cell colony-enhancing factor." Mol Cell Biol **14**(2): 1431-7.

Scandella, E., Y. Men, D. F. Legler, S. Gillessen, L. Prikler, B. Ludewig and M. Groettrup (2004). "CCL19/CCL21-triggered signal transduction and migration of dendritic cells requires prostaglandin E2." Blood **103**(5): 1595-601.

Schall, T. J., K. Bacon, R. D. Camp, J. W. Kaspari and D. V. Goeddel (1993). "Human macrophage inflammatory protein alpha (MIP-1 alpha) and MIP-1 beta chemokines attract distinct populations of lymphocytes." J Exp Med **177**(6): 1821-6.

Schindler, C. (1999). "Cytokines and JAK-STAT signaling." Exp Cell Res **253**(1): 7-14.

Schultz, J., F. Milpetz, P. Bork and C. P. Ponting (1998). "SMART, a simple modular architecture research tool: identification of signaling domains." Proc Natl Acad Sci U S A **95**(11): 5857-64.

Schultz, U., E. Grgacic and M. Nassal (2004). "Duck hepatitis B virus: an invaluable model system for HBV infection." Adv Virus Res **63**: 1-70.

Selsted, M. E. and A. J. Ouellette (1995). "Defensins in granules of phagocytic and non-phagocytic cells." Trends Cell Biol **5**(3): 114-9.

Shedlock, D. J. and H. Shen (2003). "Requirement for CD4 T cell help in generating functional CD8 T cell memory." Science **300**(5617): 337-9.

Sitaram, N. and R. Nagaraj (2002). "Host-defense antimicrobial peptides:

importance of structure for activity." Curr Pharm Des **8**(9): 727-42.

Smith, J., D. Speed, A. S. Law, E. J. Glass and D. W. Burt (2004). "In-silico identification of chicken immune-related genes." Immunogenetics **56**(2): 122-33.

Stet, R. J., T. Hermsen, A. H. Westphal, J. Jukes, M. Engelsma, B. M. Lidy Verburg-van Kemenade, J. Dortmans, J. Aveiro and H. F. Savelkoul (2005). "Novel immunoglobulin-like transcripts in teleost fish encode polymorphic receptors with cytoplasmic ITAM or ITIM and a new structural Ig domain similar to the natural cytotoxicity receptor NKp44." Immunogenetics **57**(1-2): 77-89.

Sun, J. C. and M. J. Bevan (2003). "Defective CD8 T cell memory following acute infection without CD4 T cell help." Science **300**(5617): 339-42.

Sutton, G., O. White, M. Adams and A. Kerlavage (1995). "TIGR assembler: a new tool for assembling large shotgun sequencing projects." Genome Science and Technology(1): 9-19.

Takeda, K. and S. Akira (2005). "Toll-like receptors in innate immunity." Int Immunol **17**(1): 1-14.

Tamura, T. and K. Ozato (2002). "ICSBP/IRF-8: its regulatory roles in the development of myeloid cells." J Interferon Cytokine Res **22**(1): 145-52.

Tanabe, M., S. Karaki, M. Takiguchi and H. Nakauchi (1992). "Antigen recognition by the T cell receptor is enhanced by CD8 alpha-chain binding to the alpha 3 domain of MHC class I molecules, not by signaling via the cytoplasmic domain of CD8 alpha." Int Immunol **4**(2): 147-52.

Taniuchi, I., D. Kitamura, Y. Maekawa, T. Fukuda, H. Kishi and T. Watanabe (1995). "Antigen-receptor induced clonal expansion and deletion of lymphocytes are impaired in mice lacking HS1 protein, a substrate of the antigen-receptor-coupled tyrosine kinases." Embo J **14**(15): 3664-78.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin and D.

A. Natale (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**: 41.

Taub, D. D., K. Conlon, A. R. Lloyd, J. J. Oppenheim and D. J. Kelvin (1993). "Preferential migration of activated CD4+ and CD8+ T cells in response to MIP-1 alpha and MIP-1 beta." Science **260**(5106): 355-8.

Thai, C. T. and R. T. Ogata (2004). "Complement components C5 and C7: recombinant factor I modules of C7 bind to the C345C domain of C5." J Immunol **173**(7): 4547-52.

Thermet, A., C. Rollier, F. Zoulim, C. Trepo and L. Cova (2003). "Progress in DNA vaccine for prophylaxis and therapy of hepatitis B." Vaccine **21**(7-8): 659-62.

Tirunagaru, V. G., L. Sofer, J. Cui and J. Burnside (2000). "An expressed sequence tag database of T-cell-enriched activated chicken splenocytes: sequence analysis of 5251 clones." Genomics **66**(2): 144-51.

Tsujimura, H., T. Tamura and K. Ozato (2003). "Cutting edge: IFN consensus sequence binding protein/IFN regulatory factor 8 drives the development of type I IFN-producing plasmacytoid dendritic cells." J Immunol **170**(3): 1131-5.

Underhill, D. M., A. Ozinsky, A. M. Hajjar, A. Stevens, C. B. Wilson, M. Bassetti and A. Aderem (1999). "The Toll-like receptor 2 is recruited to macrophage phagosomes and discriminates between pathogens." Nature **401**(6755): 811-5.

van de Wetering, J. K., L. M. van Golde and J. J. Batenburg (2004). "Collectins: players of the innate immune system." Eur J Biochem **271**(7): 1229-49.

van Vliet, S. J., E. van Liempt, E. Saeland, C. A. Aarnoudse, B. Appelmelk, T. Irimura, T. B. Geijtenbeek, O. Blixt, R. Alvarez, I. van Die and Y. van Kooyk (2005). "Carbohydrate profiling reveals a distinctive role for the C-type lectin MGL in the recognition of helminth parasites and tumor antigens by dendritic cells." Int Immunol **17**(5): 661-9.

Vestal, D. J., J. E. Buss, S. R. McKercher, N. A. Jenkins, N. G. Copeland, G. S. Kelner, V. K. Asundi and R. A. Maki (1998). "Murine GBP-2: a new IFN-gamma-induced member of the GBP family of GTPases isolated from macrophages." J Interferon Cytokine Res **18**(11): 977-85.

Webster, R. and D. Hulse (2005). "Controlling avian flu at the source." Nature **435**(7041): 415-6.

Webster, R. G., W. J. Bean, O. T. Gorman, T. M. Chambers and Y. Kawaoka (1992). "Evolution and ecology of influenza A viruses." Microbiol Rev **56**(1): 152-79.

Weis, W. I., M. E. Taylor and K. Drickamer (1998). "The C-type lectin superfamily in the immune system." Immunol Rev **163**: 19-34.

Wood, W. J., Jr., A. A. Thompson, J. Korenberg, X. N. Chen, W. May, R. Wall and C. T. Denny (1993). "Isolation and chromosomal mapping of the human immunoglobulin-associated B29 gene (IGB)." Genomics **16**(1): 187-92.

Yamagoe, S., Y. Kameoka, K. Hashimoto, S. Mizuno and K. Suzuki (1998). "Molecular cloning, structural characterization, and chromosomal mapping of the human LECT2 gene." Genomics **48**(3): 324-9.

Yawata, M., S. Murata, K. Tanaka, Y. Ishigatsubo and M. Kasahara (2001). "Nucleotide sequence analysis of the approximately 35-kb segment containing interferon-gamma-inducible mouse proteasome activator genes." Immunogenetics **53**(2): 119-29.

Yuasa, S., R. C. Cheung, Q. Pham, W. S. Robinson and P. L. Marion (1991). "Peptide mapping of neutralizing and nonneutralizing epitopes of duck hepatitis B virus pre-S polypeptide." Virology **181**(1): 14-21.

Ziegler, S. F., F. Ramsdell, K. A. Hjerrild, R. J. Armitage, K. H. Grabstein, K. B. Hennen, T. Farrah, W. C. Fanslow, E. M. Shevach and M. R. Alderson (1993). "Molecular characterization of the early activation antigen CD69: a type II membrane glycoprotein related to a family of natural killer cell activation antigens." Eur J Immunol **23**(7): 1643-8.