UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

MULTIPLE SPECIES COMPARATIVE ANALYSIS OF HUMAN CHROMOSOME 22 BETWEEN MARKERS D22S1687 AND D22S419 AND GENE EXPRESSION PROFILING IN ZEBRAFISH

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

CHYAU-YUEH CHRISTOPHER LAU

Norman, Oklahoma

2006

UMI Number: 3207898

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

MULTIPLE SPECIES COMPARATIVE ANALYSIS OF HUMAN CHROMOSOME 22 BETWEEN MARKERS D22S1687 and D22S419 AND GENE EXPRESSION PROFILING IN ZEBRAFISH

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY

_____
Dr. Bruce A. Roe, Chair

_____
Dr. Paul F. Cook

_____
Dr. Ann H. West

_____
Dr. George Richter-Addo

_____
Dr. Tyrell Conway

Acknowledgements

Glory to Him who is able to do immeasurably more than all we ask or imagine, and who along with His son, gave us all things.

I would like to express my sincere gratitude to my major professor, Dr Bruce Roe, for all that he taught me, and for instilling in me a meticulous approach and critical thinking in science, on top of the motivations, advises and support he gave me through the years. I wish to thank my advisory committee members, Dr. Paul F. Cook, Dr. Ann H. West, Dr. George Richter-Addo and Dr. Tyrell Conway for their continued advice, participation and encouragement. My special thanks go to Dr Han Wang who had given me access to his laboratory and zebrafish facility, and introducing me to the various techniques pertaining to the zebrafish experiments.

I also would like to extend my thanks to all members of Dr. Roe's laboratory, for their advice, support and help. Special thanks to Dr. Fares Najar who spent innumerable hours on my analysis and whose insights and humor are invaluable; Trang Do and Ahn Do who first trained me when I started work; Shelly K. Oommen, Jiangfeng Li, and Hung-Chun (James) Yu who worked in the same group with me on human chromosome 22 comparative analysis and zebrafish genes profiling studies; Dr Axin Hua, Steve Kenton, Jim White and Hongshing Lai who assisted me in the sequence analysis; Shweta Deshpande, Sara Downard, Mounir Elharam, Ying Fu, KayLynn Hale, Xiangfei Kong, Dr. Doris Kupfer, Cathy Lai, Jennifer Lewis, Shaoping Lin, Simone Macmil, Phoebe Loh-Marley, Rose Morales-Diaz, Goldameir Osisanya, MaryCatherine Pottorff, Sulan Qi, Baifang Qin, Chunmei Qu, Jiaxin (Carson) Qu, Jiaxi Quan, Iryna Sanders-Vasy, Majesta Seigfried, Ruihua Shi, Stephen Snow, Leo

Sukharnikov, Keqin Wang, Ping Wang, Doug White, Graham Wiley, Dixie Wishnuck, Junjie Wu, Yanbo Xing, Limei Yang, Ziyun Yao, Jing Yi, and Liping Zhou, in addition to members of Dr. Wang's laboratory Dr. Qingchun Zhou, Jason Kesinger, Eric Lee, Jason Yousif, Joe Ghatta, Fadalia Kim, and George Martin, who collectively made my learning experience a memorable one.

I dedicate this dissertation to my dear wife Dr. Charissa Lu-Ming Chin, and my family, whose love had made this experience all the more joyful and meaningful.

Table of Contents

ChapterIII Results and Discussions

A. Comparative sequence analysis

B. Gene expression profiling in zebrafish

List of Tables

List of Figures

Abbreviations

AIFL	Apoptosis-inducing factor like gene
ALL	Acute lymphoid leukemia
BAC	Bacterial artificial chromosome
BCR	Breakpoint cluster region
BLAST	Basic local alignment search tool
Bp	Base pair
cDNA	Complementary deoxyribonucleic acid
CES	Cat Eye syndrome
CML	Chronic Myeloid leukemia
DNA	Deoxyribonucleic acid
DGCR	DiGeorge syndrome critical region
EST	Expressed sequence tag
Hpf	Hours post fertilization
IGH	Immunoglobulin heavy chain
IGLK	Immunoglobulin light chain kappa locus
IGLL	Immunoglobulin light chain lambda locus
IGLV	Immunoglobulin lambda variable segments
IGLJ	Immunoglobulin lambda joining segments
IGLC	Immunoglobulin lambda contant segment
Indels	Insertions or deletions
Kb	Kilobase
LCR22	Low copy repeats of chromosome 22

LINE	Long interspersed repeat element
Mb	Megabase
MER	Medium element of repeat
MMP11	Matrix metalloproteinases member 11 gene
mRNA	Messenger ribonucleic acid
Mya	Million years ago
ncRNA	Non-coding ribonucleic acid
ORF	Open reading frame
PIP	Percent identity plot
PPIL2	Peptidylprolyl isomerase (cyclophilin)-like member 2 gene
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
tRNA	Transfer ribonucleic acid
SINE	Short interspersed repeat element
SLC7A4	Solute carrier family 7 member 4 gene
SLC2A11	Solute carrier family 2 member 11
snRNA	Small nuclear ribonucleic acid
snoRNA	Small nucleolar ribonucleic acid
STMY3	Stromelysin III gene
Thap7	Thanatos-associated protein member 7 gene
WMISH	Whole mount *in situ* hybridization

Abstract

Comparison of a 4.5 Mb region of human chromosome 22 between markers D22s1687 and D22s419, with the syntenic region in chimpanzee had revealed overall DNA sequence identity of approximately 97.6%, Ka/Ks ratio of known protein coding genes at approximately 0.25, with the majority of amino acid changes between hydrophilic amino acids, followed by changes between hydrophobic amino acids, and the least changes between hydrophobic to hydrophilic amino acids or vise versa. Thus, the first major conclusion of this study is that overall, this chromosomal region is highly conserved between human and chimpanzee, and the known protein coding genes are undergoing purifying selections, in which 75 % of nucleotide substitutions that led to amino acid changes were eliminated by adaptive evolution.

Major large scale insertions or deletions that resulted in gene number differences between human and chimpanzee were discovered in the IGLL and LCR22s within this region, with four human insertions from 6 Kb to 75 Kb and three chimpanzee insertions from 12 Kb to 74 Kb observed in the IGLL region, two human insertions of 59 Kb and 36 Kb in LCR22-6, and a 67 Kb chimpanzee insertion in LCR22-8. Small scale insertions and deletions, in addition to exon shuffling, elevated nucleotide divergence rate and positive selection were also observed in the putative genes, partially duplicated genes and pseudogenes in the IGLL and LCR22s. Thus, the second major conclusion of this study is the major differences between human and chimpanzee in this region lies in the highly repetitive regions of the IGLL and the LCR22s.

Through whole mount *in situ* hybridization studies, a total of 12 human orthologs in zebrafish, including 4 newly predicted putative genes with no previously

known expression profile and function, showed specific expression in the developing zebrafish embryonic central nervous system, optic system, the neural crest cells, ottic vesicle, liver, and notochord. Thus, the third major conclusion from this present study is that many predicted genes which currently lack expression data and functional information likely are time and tissue specific during developmental processes.

Chapter I: Introduction

1.1 Human Genome: Structural Organization and Content

1.1.1 Hereditary information

DNA

Hereditary information of almost all living organism, except for some retroviruses, are stored in deoxyribonucleic acid (DNA) molecules. DNA is a polymer of deoxyribonucleotides, each composed of a base, a pentose sugar, and a phosphate group (Avery et al. 1944; Watson and Crick 1953). The human genome contains approximately 3 billion base-pairs (bp) of deoxyribonucleotides (IHGSC 2001). There are 4 types of nitrogenous bases in DNA, the double-ring purines: adenine (A) and guanine (G); and the single-ring pyrimidines: thymine (T) and cytocine (C). The pentose sugar in DNA is deoxyribose. The four bases are attached to deoxyriboses via covalent bonding from C-1 of the deoxyriboses to N-9 of purines and N-1 of pyrimidines. The deoxyribose sugars are linked by phosphate groups in a phosphodiester bond, forming the sugar backbone of DNA. The two complementary DNA strands then form Watson-Crick base pairs through hydrogen bonding with a purine pairing with a pyrimidine, i.e., adenine pairs with thymine through 2 hydrogen bonds, and guanine pairs with cytosine through 3 hydrogen bonds. This results in the DNA forming a double helix (Watson and Crick 1953a) with the complementary strands acting as templates for each other during semi-conservative DNA replication (Watson and Crick 1953b).

RNA

Ribonucleic acid (RNA) differs from DNA by having ribose in place of deoxyribose in its sugar backbone, and by replacing thymine with uracil (U). Apart from storing the hereditary information for some viruses, RNA also acts as an information transmitter in the form of messenger RNA (mRNA), and stable functional RNAs such as ribosomal RNA, tRNAs and microRNAs that together regulate and fascilitate the transmission of information from DNA to producing functional proteins. Some functional RNA molecules such as the tRNA contain modification on the bases of the 4 standard nucleotides.

1.1.2 Functional sequences

Genes

A gene is defined as a segment of DNA that is transcribed. There are two types of genes, protein-coding genes and noncoding stable RNA (ncRNA) genes.

In the human genome, as in all eukaryotic genomes, protein-coding genes are a combination of regulatory regions and a mosaic of protein coding exons and intervening non-coding introns (Wenkink et al. 1974; Berget et al. 1977). Within a particular gene sequence, the 5' flanking region contains specific DNA sequences that regulate gene transcription. This 5' flanking promoter region of the gene that often is comprised of one or more copies of GC boxes (most common seen in house-keeping genes), consisting of the sequence GGGCGG that often is followed by a CAAT box and a TATA box, located approximately 19-27bp upstream of transcription start point. The GC, CAAT and TATA boxes are sites for transcription factor binding and the TATA

box determines the start point of RNA polymerase directed transcription.

The first and the last exons of a gene are flanked by 5' and 3' untranslated regions, respectively. The 5' untranslated region begins at the transcription start point and is downstream from the promoter region. There are 3 enzymes involved in transcribing DNA: RNA polymerase I synthesizes ribosomal RNAs (rRNA), RNA polymerase II synthesizes pre-mRNA, and RNA polymerase III synthesizes transfer-RNA (tRNA), small nucleolar RNA (snoRNA) and small nuclear RNA (snRNA). In the case of pre-mRNA, the introns are cleaved by an enzymatic complex called the spliceosome. The splicing sites or junctions of introns are determined by the presence of 5' and 3' end donor and acceptor sites. Most eukaryotic introns have the dimer GT as their 5'end and the dimer AG as their 3'end (GT-AG). Each intron also contains a specific TACTAAC box located approximately 30bp upstream of the 3' end of the intron thatparticipates in lariat formation . These unique features of introns are essential for the correct excision and splicing of introns.

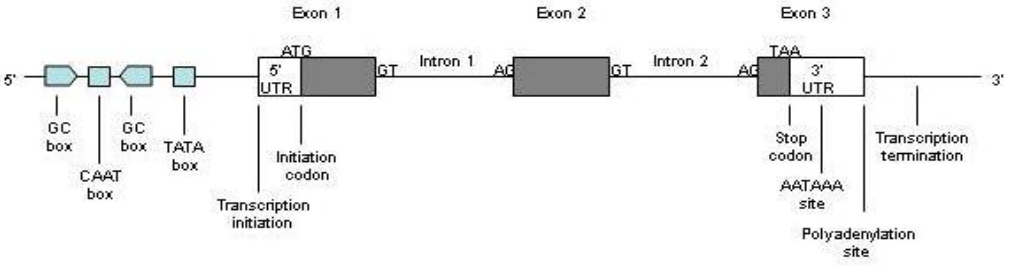

Figure 1.1 Schematic representation of a eukaryotic protein coding gene including the 5' and 3' flanking region.

Non-coding RNA genes

Non-coding RNA (ncRNA) genes are transcribed but not translated into proteins. They do not have ORFs, are usually small, and are not polyadenylated. The major classes of ncRNA in the human genome are transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs), telomerase RNA, 7SL RNA and Xist, a nontranslated RNA transcript that is involved in X chromosome inactivation in mammals (IHGSC 2001; Allaman et al. 2001).

Enhancer and Cis-regulatory elements

Enhancers or cis-regulatory elements are DNA sequences that are responsible for regulating the transcription of a gene, by specifying temporal and spatial pattern of expression of a transcript. They are distinct from the promoter region that is located directly 5' of transcription start site in a gene, and may be located 5' or 3' to a gene, or within exon and/or intron region of the gene (IHGSC 2001).

Genetic Code

The genetic code is a series of non-overlapping nucleotide triplets called codons, and each codon specifys one of the 20 amino acids that make up a protein (Crick et al. 1961). In a protein coding gene, codons are contiguous in the final mRNA with translation occurring in the 5' to 3' direction. With only a few exceptions, all eukaryotes and prokaryotes use the same set of universal genetic code.

With 4 different nucleotides A, C, G, and T, there are 64 possible arrangements for 3 nucleotide codons. 61 codons code for specific amino acids (sense codons), and 3 codons signal the termination of translation (stop codons). In human, as in most eukaryotes, the first amino acid in proteins typically is a methionine specified by the

initiation codon AUG. Since 61 sense codons are responsible to specify only 20 amino acids, most of the codons are redundant. 18 out of 20 amino acids are specified by more than one codon. The different codons specifying the same amino acid are called synonymous codons. Codons specifying different amino acids are termed non-synonymous codons.

As mentioned above, proten synthesis involves translating the genetic information from mRNA to amino acid sequences. This is accomplished through transfer RNA (tRNA), each aminoacylated with a specific amino acid that binds to the ribosome-mRNA complex and facillitates the growing polypeptide chain (Crick 1966).

1.1.3 Repeat Sequences

More than 50% of the human genome are repeat sequences (IHGSC 2001). The repeat sequences can be categorized into 5 distinct category, transposon derived repeats, segmental duplications, processed pseudogenes, simple sequence repeats, and tandem repeats.

Transposon-derived repeats

The majority of repeat sequences in the human genome are transposon derived, make up approximately 45% of the human genome (IHGSC 2001), and are grouped into 4 major classes, long terminal repeats (LTRs), long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and DNA transposons.

In the human genome, the most abundant class of interspersed repeat sequence are the LINEs that make up approximately 20% of the genome (IHGSC 2001). An autonomous LINE is 6.1 kb in length, has a polymerase II promoter, and contain genes

that encode its own transcription and integration proteins. In some instances, autonomous LINEs can be transcribed and after translation, both the transcript and its resulting protein will move into the nucleus where an AT rich region of the genome is nicked by the endonuclease activity of the LINE proteins, and the nicked single stranded DNA will prime the reverse transcription of the transcript. The reverse transcribed DNA product then is inserted at the nicked site. Reverse transcription usually does not proceed to the end of the transcript, resulting in many truncated LINEs. Most of the approximately 3500 full length LINEs and several hundred thousand truncated copies in the human genome are non-autonomous LINEs that have lost their transposition ability.

SINEs are the second largest interspersed repeat class in humans, representing approximately 13% of the human genome (IHGSC 2001). SINEs are approximately 100-400 bp long, and have a 7SL derived polymerase III promoter, but lack the transcription and integration machinery associated with LINEs. It therefore is belived that SINEs borrow the LINEs machinery for its retrotransposition events (Okada et al. 1997). Unlike LINEs, SINEs, that mostly are found in the GC rich region of the human genome, can be grouped into 3 major classes, e.g. Alu, MIR and MIR3 with the most common human SINE, the Alu class, exceeding one million copies in the genome.

Approximately 8% of the human genome consists of LTRs (IHGSC 2001). The LTRs found in the human genome are derived from endogenous retroviruses (ERV) that have integrated into the vertebrate genome (Malik et al. 2000). Although some LTRs still are active and might direct synthesis of exogenous viruses, most are inactive LTRs because of point mutation, insertions or deletions. There are three major classes of ERV

in human genome, namely ERV-classI, ERV(K)-classII, and ERV(L)-classIII.

DNA transposons are the smallest class of interspersed repeats making up of approximately 3% of the human genome (IHGSC 2001). DNA transposons do not rely on RNA intermediates in transposition since they have terminal inverted repeats and the autonomous DNA transposons encoding a transposase that allows them to be transposed by being cleaved from one location and integrated into a different location in the genome. DNA transposons have diverse families and origins. There are seven major classes of DNA transposons, namely MER1-Charlie, Zaphod, Mer2-Tigger, Tc2, Mariner, PiggyBac-like, and Unclassified (Smit, 1996).

Low copy repeats or segmental duplication

Low copy repeats or segmental duplications consist of the duplicated, transposed genomic DNA ranging in size from one to hundereds of kilobases (kb) that occur at multiple locations in the human genome (IHGSC 2001). These duplicons are highly identical, usually have >95% sequence identity, and they may contain introns and exons as well as repetitive elements such as Alus and L1 (Bailey et al. 2002). The two classes of low copy repeats or segmental duplications are interchromosomal duplications and intrachromosomal duplications. Interchromosomal duplications occur between nonhomologous chromosomes and intrachromosomal duplications occur within a chromosome or chromosome arm. The pericentromeric and subtelomeric regions of human chromosomes are 10 times more likely to have segmental duplicons compared to other regions in the human genome (IHGSC 2001, Bailey et al. 2001). This is consistent with previous studies that pointed to numerous segmental duplications at the pericentromeric and subtelomeric regions during the course of hominoid evolution

(Eichler et al. 1999; Horvath et al. 2000; Jackson et al. 1999; Monfouilloux et al. 1998; Trask et al. 1998). The highly identical structure of the duplicons results in deletions, duplications, translocations and inversions.

Chromosome 22 specific low copy repeats (LCR22s) have been found within 9 Mb of the centromere (Dunham et al. 1999; Bailey et al. 2002) as shown in Figure 1.2. The repeating units within these LCR22s include the BCR-like, DGCR-like, GGT-like, and PIK4CA-like genes that are partially duplicated copies of their functional counterpart. The highly similar sequence identity of these repeating units can lead to chromosomal rearrangements (Edelman et al. 1999) that have been implicated in cat-eye syndrome (CES; OMIM 115470) (Schinzel et al. 1981), der(22) syndrome (Zackaiand and Emanuel 1980), velo-cardio-facial syndrome (VCFS ; OMIM 192430) (Shprintzen et al. 1978), DiGeorge syndrome (DGS; 188400) (DiGeorge 1965), chronic myeloid leukemia (CML; OMIM 608232) (Nowell and Hungerford, 1960), t(8;22) associated Burkitt's lymphoma (BL; OMIM 113970) (Emanuel 1984; Davis 1984), Ewing sarcoma (ESWR1; OMIM 133450), malignant rhabdoid tumors and miningiomas (OMIM 601607).

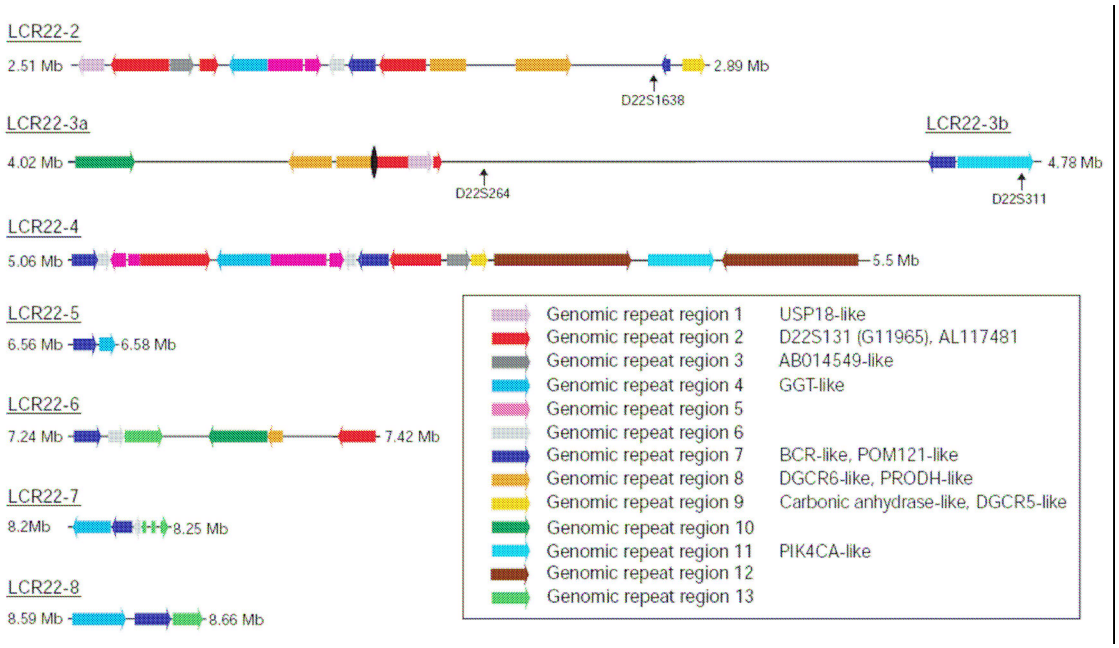

Figure 1.2 LCR22s characterized in in human chromosome 22 (Dunham et al. 1999), showing 7 LCR22s positions from the centromere and the repeating units within each LCR22.

Pseudogenes

Pseudogenes are segments of DNA that are similar to genes, but have nonsense and/or frameshift mutations, and often missing exons, introns, promoter regions or enhancers, that prevent them from being transcribed or translated (Vanin 1985; Mighell et al. 2000; Harrison et al. 2002). Non-processed and processed pseudogenes are the 2 types of pseudogenes. Non-processed pseudogenes are formed when whole or part of functional genes containing both introns and exons are duplicated. Since they are redundant, selective pressure does not prevent the accumulation of mutations in some of the copies that will ultimately turn them into pseudogenes. Such pseudogenes usually are characterized by shifted reading frame or truncated genes. Processed pseudogenes are formed when mature mRNA is reverse transcribed and integrated into the genome. Such pseudogenes are characterized by the absence of introns.

Simple sequence repeats

Simple sequence repeats (SSRs) are tandemly repeating units of one of several unique 1-500 bp sequences (IHGSC 2001). Short repeating units of 1-13 bp often are classified as microsatellites, while longer repeating units, typically in the range of 14-500 bp, are classified as minisatellites. Aside from poly-A tails that are reverse transcribed, most SSRs are believed to be formed from DNA replication slippage (Kruglyak et al. 1998; Toth et al. 2000). DNA replication slippage also results in SSRs length polymorphisms in the human population, making SSRs important in human disease mapping and forensic studies, where the lengths of specific repeats, such as $(CA)_n$ repeats, are determined and used extensively as disease or individual-specific tags (Dib et al. 1996; Broman et al. 1998).

Tandem repeats

Tandem repeats are similar to SSRs but contain a larger repeated sequence. They are hypothesized to have arisen through mechanism involving either replication slippage or by DNA recombination (IHGSC 2001) to create linked tandem repeats.

1.1.4 Organizational unit: chromosomes

The enormous amount of genetic material contained in the human genome is condensed and packaged into units called chromosomes. The human genome comprises 22 pairs of autosomes and 1 pair of sexual chromosomes, each consisting of two arms linked by centromeres. In the human genome, 18 chromosome pairs have arms of almost equal length and 5 have one arm significantly shorter than the other. The short or p-arms (p from the French word petite) in the acrocentric chromosomes encode

tandemly repeated ribosomal RNA genes as well as other tandem repeats (Dunham, 1999), and their long arms encode protein coding genes.

DNA in human chromosomes is packaged in the nucleus by coiling around histone proteins to form a packed, condensed structure called the nucleosome. The nucleosomes are coiled further to form chromatin (Kornberg 1974 and Finch et al. 1977), which in turn must be uncoiled from its condensed form to allow transcription (Kornberg & Lorch 1992).

The centromeres of chromosomes are essential for effective separation of sister chromosomes during meiosis and mitosis. The centromeres in human chromosomes consist of the alphoid satellite DNA, a 169 and 172 bp primate specific tandem repeat family (Warburton et al. 1993 & 1996; Lee et al. 1997). Telomeres at the distal end of each chromosome arm are the sites at which pairing of homologous chromosomes are initiated. Telomeres in humans and other primates consist of minisatellite repeats containing tandem hexanucleotide, TTAGGG extending up to 15 Kb (Allshire et al. 1989; Brown 1989; Luke and Verma, 1993).

1.2 Human Phylogeny and Genome Evolution

1.2.1 Human and other vertebrates

Humans, *Homo sapiens,* belong to the phylum vertebrata which is defined by the presence of the vertebral or spinal column which encloses the dorsal nerve cord, and the cranium which houses the brain. Living vertebrates can be divided into two main groups based on morphology (Field et al. 1988; Adoutte et al. 2000; Klein & Takahata 2002), the Agnatha (jawless vertebrates) and the Gnathostomata (jawed vertebrates). The jawless vertebrates include the hagfishes and lampreys while the jawed vertebrates are divided into six groups: the Chondrichthyes (cartilaginous fishes), Osteichthyes (bony fishes), Amphibia, Reptilia, Aves, and Mammalia.

Based on molecular timescale, the earliest vertebrates diverged approximately 564 Myr ago (Kumar & Hedges 1998). This is consistent with fossil record for the first appearance of vertebrates (Kumar & Hedges 1998; Benton 1997) at 514 Myr ago. It was hypothesized that the vertebrate genome had undergone two successive whole-genome duplications by polyploidization (the 2R hypothesis) (Ohno 1970, Sidow 1996). The discovery of genes and gene families such as the homeobox (Hox) clusters that have four copies in most vertebrates, including human on chromosomes 2, 7, 12, and 17, but only one copy in invertebrates such as fruit flies and round worms, had been cited as evidence supporting the hypothesis (Holland 1994, Sidow 1996, Spring 1997, Thornton 2001). However, there is strong skepticism to this hypothesis (Martin 2001, Friedman 2001, Hughes 2001) that is compounded by the initial analysis of the human genome where majority of the genes do not fall into the “4 copies in vertebrates and 1 copy in invertebrates” model. However the hypothesis cannot be completely ruled out

because other genome evolution events, such as segmental duplication or deletion, could have skewed the interpretation of the analysis (IHGSC 2001).

The Osteichthyes or bony fishes can be further divided into subclasses of Actinopterygii (ray-fin fish) and Saarcopterygii (lobe-fin fish). Based on comparative genomics studies that showed many genes and gene clusters have two copies in ray-fin fish compared to only one copy in other vertebrates, it was proposed that an additional round of whole genome duplication had occurred in ray-fin fish lineage (Wittbrodt et al. 1988; Amores et al. 1998; Postlewait et al. 2000; Aparicio et al. 2002; Taylor et al. 2003).

1.2.2 Human and other mammals

Mammals, 1 of the 4 terrestrial vertebrates that are distinguished from the other classes of vertebrates by their ability to lactate, are divided into 3 subclasses (Klein & Takahata 2002), Prototheria, Metatheria, and Eutheria. Prototheria (monotremes), include the platypus and echidnas that lay eggs and posses a chamber receiving discharge from the digestive, excretory and reproductive tracts termed the cloaca. Methatheria (marsupials) and Eutheria (placentals), have eggs that develop in the uterus of the female. The marsupials, which include the kangaroos, opossums, and wallabies, give birth to partially developed embryos that complete their development in a pouch outside of the female body called marsupium. The placentals, which include rodents and primates, have their embryos developed to an advanced stage in the female uterus, enclosed in an embryonic sac called the placenta.

It has been estimated that at least 5 major lineages of the placental mammals

appeared more than 100 Myr ago (Kumar & Hedges 1998). Recent independent molecular analyses have produced a concordant picture of earliest divergence events among the 18 modern orders of placental mammals (Murphy et al. 2001; Madsen et al. 2001). One of the most well studied models for human physiology, pathology and evolution had been the rodents, mainly mouse and rats. The completion of the draft genome sequences of mouse (MGSC 2002), and rat (RGSPC 2004) has yielded insights into the mammalian genome evolution. Despite an estimated divergent time of 75 Myr of the rodent lineage (MGSC 2002), large segments in the genome of the common ancestor have been passed on to human and rodent with minimal rearrangements in gene order within these segments (MGSC 2002; RGSPC 2004). Thus, over 90% of human and mouse genomes have conserved synteny (MGSC 2002) represented in approximately 280 human and mouse, and 278 human and rat orthologous segments with a minimal size of 1 Mb (RGSPC 2004). Since mouse and rat are estimated to have diverged from each other between 12-24 Myr ago (Adkins 2001; Springer 2003; RGSPC 2004), the conservation of synteny between these 2 rodent lineages is even greater with 105 large orthologous segments.

Genomic mapping projects for other placental mammals that have co-evolved with human including cow, pig, sheep and dog also have been reported (O'Brien et al. 1993; Edwards 1994; Eggen & Fries 1995; Womack & Kata 1995; Nadeau et al. 1995; Eppig 1996), and indicated that the genome organization between humans and cows is more conserved than between humans and mice (Band et al. 2000). A detailed comparison between humans and cows awaits the completion of the whole genome shotgun (WGS) sequencing effort for cow that presently is underway.

1.2.3 Human and other primates

Primates are one of the 18 living placental orders, and the human species belongs in this order. Primates are estimated to have diverged from other placental mammals 50-60 Myr ago (Martin 1993). The approximately 300 living species of primates are classified into 2 suborders: the Prosimii and the Anthropoidea. The suborder Prosimii includes lemurs, lorises, tarsiers, and the suborder Anthropoidea includes New World monkeys, Old World monkeys, and the great apes (Groves 1997; Fleagle 1999). Primates are distinguished from other placental mammals by a combination of morphological traits that include shorter snout length and skull, a flatter face, forward projecting eyes which lead to binocular vision, increased mobility of their digits and the development of a thumb, replacement of crawl by flat nails, and an increase in brain size.

Lemurs, that are found at the south east coast of Africa (Enard & Paabo 2004), include indris, avahi, sifakas, mouse lemurs, dwarf lemurs, and true lemurs. They are about the size of squirrels, and are distinguished by their long furry tails, protruding snouts, and fluffy fur. Lorises are nocturnal and arboreal primates that are found in Africa and southern Asia. They are tailless and have slow movements. Tarsiers also are nocturnal, and can be found in islands of southeastern Asia. They are the size of rats, and possess head that can rotate so they can look backwards over their shoulders. They possess large forward-looking eyes, big ears, long hind legs that are lengthened by the elongation of their ankle bones and hairless tails.

New world monkeys can be found in central and south America (Enard & Paabo 2004), and they include marmosets, tamarins, howler, capuchin, squirrel monkeys,

spider monkeys and wooly monkeys. They posses broad noses with large nostrils, long limbs and tails that enable them to hang from tree branches, their exclusive habitat that places them close to the leaves and fruits that comprise their diets.

Old World Monkeys can be found across Africa and Asia (Enard & Paabo 2004), and they include rhesus macaques, baboons, guenons, mangabeys, langurs, drills, mandrills and colobus monkeys. They are distinguished by their narrow nostrils that face downward and outward, their opposable thumbs, non-prehensile tails, and pad-like buttocks that facilitate sitting on the ground, and their locked shoulders that prevent them from hanging or swinging on the branches, although their habitats include tree branches as well as the ground.

Apes, found in Asia (Enard & Paabo 2004), include gibbon, siamang and orangutan, and apes found in Africa (Enard & Paabo 2004) include gorilla and chimpanzee. Apes are tailless and have flexible arms. Among the different groups of primates, apes are the closest morphologically to the human species.

Divergence time of the human lineage from the other primates had been estimated based on fossil calibration points and various statistical methods (Glazko& Nei 2003). Based on the estimations, the human lineage had diverged from the New World monkey lineage 32-36 Mya, from the Old World monkey lineage 21-25 Mya, from the orangutan lineage 12-15 Mya, from the gorilla lineage 6-8 Mya, and from the chimpanzee lineage 5-7 Mya (Goodman 1999; Glazko& Nei 2003) It was hypothesized that the rate of substitution among hominoids (human and apes) has slowed by 50% since their divergence from the old world monkeys (Goodman et al. 1971; Koop et al. 1986; Li and Tanimura 1987).

The overall genomic organization of primates is highly conserved. Other than differences in their chromosomal reorganization in baboons, gibbons, owl monkeys, or lemurs (O'Brien et al. 1999; Enard & Paabo 2004), overall primate karyotypes have remained stable (Muller & Wienberg 2001). Using the G-banding, it was demonstrated that the human and chimpanzee karyotypes only differ by 10 large scale genomic rearrangements (Yunis and Prakash 1982). Two chromosomes in the human lineage have fused at the telomeres, resulting in a hybrid human chromosome 2 and thus human have one fewer chromosome than the other great apes. In addition, human also have 9 pericentric inversions that are not present in the other great apes (Yunis and Prakash 1982).

1.3 Sequencing the human genome

1.3.1 A historical perspective

The Human Genome project is a research effort involving an international collaboration of 6 countries and 20 research groups aimed at making the sequence of the human genome freely available to the public (IHGSC 2001). An effort of such magnitude has never before been attempted in biomedical research and has its roots in several key events in the late 1970s and throughout the 1980s. In 1977, two separate groups had developed and published DNA sequencing techniques. Maxam and Gilbert at Harvard University developed the chemical cleavage DNA sequencing method (Maxam & Gilbert 1977), and Sanger, Nicklen, and Coulson at the Medical Research Council Laboratory of Molecular Biology, Cambridge, England, developed the dideoxynucleotide termination DNA sequencing method (Sanger et al. 1977). The innovations in DNA sequencing methods coupled with successful efforts in sequencing the genome of bacterial viruses ΦX174 (Sanger et al. 1977, 1978) and lambda (Sanger et al. 1982) had demonstrated the feasibility of assembling short individual DNA sequences into whole genomes and can result in the complete genomic sequence of an organism (IHGSC 2001). Subsequently, the shotgun sequencing strategy that was introduced in the early 1980s (Anderson 1981; Gardner et al. 1981; Deininger 1983), was automated in the late 1980s and early 1990's by Lee Hood and colleagues (Smith et al. 1986), as well as by others (Ansorge et al. 1987; Prober et al. 1987; Brumbaugh et al. 1988; Kambara and Takahahi 1993), enabled large-scale, accurate, cost effective and rapid DNA sequencing.

At about the same time, collective efforts to create a human genetic map of

disease related unknown genes (Botstein et al. 1980) as well as physical maps of yeast (Olson et al. 1986) and worm (Coulson et al. 1986) genomes had begun. In 1984, the US Department of Energy and others organized meetings to discuss the idea of a collective effort to sequence the entire human genome (Palca 1986; Sinsheimer 1989; IHGSC 2001). As a result, the report "Mapping and Sequencing the Human Genome" produced by the National Research Council in 1988 called for a Human Genome Project, that was not confined only to sequencing the human genome, but also include establishing the genetic, physical and sequence maps for the human genome, and for other key model organisms such as bacteria, yeast, worms, flies and mice, developing technologies to support the projects mentioned, and initiating studies involving the ethical, legal and social issues associated with the human genome research.

In 1990, the Human Genome Project was launched as a multi-national effort, with different national agencies spearheading the effort in 6 different countries. In the US, The Human Genome Project was undertaken by both the Department of Energy and the National Institute of Health; in UK, it was the UK Medical Research Council and the Wellcome Trust; in France, the Centre d'Etude du Polymorphisme Humain and the French Muscular Dystrophy Association; in Japan, multi government agencies including the Ministry of Education, Science, Sports, and Technology. Under these agencies, genome centers were established in the various countries. Subsequently 2 additional countries, Germany and China, joined the collaboration after the initial launching of the Human Genome Project.

By 1996, genetic and physical maps for both human and mouse were well established and the 1st International Strategy Meeting on Human Genome Sequencing

was convened in Bermuda. Among the agreements reached in this meeting were that all sequences generated would be made freely available to the public domain, and sequence assemblies would be released rapidly. In the following year, the 2nd International Strategy Meeting on Human Genome Sequencing, again convened in Bermuda, resulted in sequence quality, submission and annotation standards, as well as methods to establish sequence claims and etiquettes (HGPI). By 1999, the sequence of human chromosome 22, the first human chromosome completed, was published by our laboratory in collaboration with groups in the U.S.A, U.K and Japan (Dunham et al. 1999). Subsequently, 2 separate working draft sequences of the human genome were published in February 2001, by the publicly sponsored Human Genome Project (IHGSC 2001) and the private company Celera Genomics (Venter et al. 2001). Since then, an updated, highly accurate and nearly completed sequence was published in 2004 (IHGSC 2004).

1.3.2 Chromsome 22: The first human chromosome completed

Chromosome 22, the second smallest of human autosomes that makes up approximately 1.6-1.8% of the total human genome, is an acrocentric chromosome that have both a short (22p) and a long (22q) arm. To date, at least 37 human disorders have been linked to this chromosome (Sibbald et al. 2000) as shown in Table 1.1.

The published sequence of human chromosome 22 reveals that it is comprised of approximately 33.4 megabases of DNA in its euchromatic region. In the initial annotation (Dunham et al. 1999), 22q was estimated to contain at least 545 protein coding genes, including 247 known genes that are identical to known human gene or

protein sequences; 150 related genes that have homology to gene or protein sequences from human or other species, 148 predicted genes with homology to ESTs; and 134 pseudogenes, that are sequences homologous to known genes or protein sequences but contain disrupted open reading frames. The initial Fgenesh and Genscan computer prediction identified 887 and 817 genes, respectively. Among these predicted genes, 325 did not form part of the annotated genes categorized above.

Chromosome 22 associated syndromes and disease

Amyotrophic lateral schlerosis	Meningioma
Breast cancer	Mental retardation
Cat-eye syndrome	Metachromatic leukodystrophy
Cataract, cerulean, type 2	Myoneurogastrointestinal encephalomyopathy
Bernard-Soulier syndrome, type B	Neurofibromatosis, type 2
Breakpoint cluster region (CML)	Opitz G/BBB syndrome
Colon cancer	Ovarian cancer
Deafness	Pheochromocytoma
Dermatofibrosarcoma protuberans	Pulmonary alveolar proteinosis
DiGeorge syndrome	Schizophrenia
Ewing's sarcoma	Schwannomatosis
Glioma of brain	Sorsby's fundus dystrophy
Glucose-galactose malabsorption	Spinocerebellar ataxia
Glutathionuria	Succinylpurinemic autism
Heme-oxygenase-1 deficiency	Thrombofilia due to heparin cofactor-2 deficiency
Hirschsprung disease	Transcobalamin 2 deficiency
Hyperprolinemia type 1	22q13 deletion syndrome
Lysosomal Nacetylgalactosaminidase deficiency	velolcardiofacial syndrome
Malignant rhabdoid tumor	

Table1.1 List of human chromosome 22 associated diseases previously reported (Sibbald et al. 2000).

A revised annotation of human chromosome 22 then was published in 2003 (Collins et al. 2003). Based on comparison with increased genome sequences and EST databases since the initial annotation, new genes were identified, fragmented genes were fused together, and missed exons were included. In this second generation of human chromosome 22 annotation, the number and category of genes are as follow (Collins et al. 2003): 393 complete protein-coding genes that are identical to human cDNA or ESTs in its entire length, and has at least 300 bp of ORF; 153 partial genes that have sequence similarity to cDNAs or ESTs and are potential coding genes but do not have the entire sequence match or do not satisfy the criteria of the protein coding gene; 31 non-coding RNA genes that include 6 small RNA genes, 9 genes with no ORF, and 16 potential anitsense genes; 234 pseudogenes that are similar to known genes but have disrupted sequences and ORF; and 125 IGLV and J gene segments. From the 936 structures annotated, there are 209 known genes that are identical to human cDNAs or protein sequences, have an entry in LocusLink, and have a RefSeq accession in NCBI.

1.3.3 Targeted region on human chromosome 22

The targeted region for my Ph.D. research is a 4 Mb segment of chromosome 22 between markers D22s1687 and D22s419, and including 4 low copy repeats (LCR22s), the Immunoglobulin Lambda Light Chain region (IGLL), and the Breakpoint Cluster (BCR) region. A total of 126 gene structures incuding 29 known coding genes, 20 putative coding genes, 34 partially duplicated genes, 43 pseudogenes and 1 non-coding genes, in addition to 125 Immunoglobulin Lambda Light Chain segments were

annotated in this region (Collins et al. 2003). 31 zebrafish orthologs of human chromosome 22 genes in this targeted region have been identified during the course of my research by comparison to the recent Ensembl zebrafish genome assembly Zv5. The different classes of genes in this region and information regarding them are summarized in the gene table in the Appendix.

1.4 Understanding the human genome: Model organisms

1.4.1 Human Genome research

The completion of the human genome sequence is not an end to itself, but rather the beginning of human genome research in a holistic and systematic way. Most of the information that dictates how humans develop and function is encoded in the human genome sequence, therefore decoding and retrieving meaning from the sequence is one of the ultimate purposes of obtaining the sequence. Thus, there are 3 major aims underlying current human genome research. The first is structural annotation by identifying and characterizing all genomic elements, cataloging all protein coding genes, cis-regulatory elements and enhancer sequences, non-protein coding genes, repetitive elements, and large scale genomic architecture. The second is functional annotation bydecipher ing the role of all functional elements, including more than half of the approximately 25,000 presently predicted genes in the human genome that have no known function and awaiting validation (IHGSC 2001; IHGSC 2004). The third is phylogenetic annotation, by comparing and tracing specific evolutionary changes in the genome structures and contents that had ultimately led to unique developmental, morphological, and physiological features in human in contrast to other species. Currently, the main thrust towards achieving these aims lies in the genome sequencing and experimental design of animal models. Earlier genome sequencing projects such as those for budding yeast *Saccharomyces cerevisiae,* round worm *Caenorhabditis elegans* and fruit fly *Drosophila melanogaster* projects were launched by the HGP, other than their role as classical experimental model, for the purpose of testing large scale genome sequencing procedures and developing high-throughput methods for sequence analysis

(IHGSC 2001; Celniker & Rubin 2003). Since then, criteria for the selection of model organism for genomic sequencing are based on their phylogenetic relationship to human, relevance to human biology, and their potential to aid annotation of the human genome, although other considerations often included their significance for experimental designs, size of the genome, the cost for sequencing, and the model's economic value.

Much of the recent work in our laboratory has been focused on accomplishing the afore mentioned major aims related to human genome annotation. The focus of my Ph.D. research is described below.

1.4.2 Focus I: Multiple species comparative sequence analysis

Part of this focus of my research was to sequence and analyze the chimpanzee chromosome 22 region syntenic to human chromosome 22 between markers D22s1687 and D22s419, as part of a collective effort in our laboratory to complete the sequence of chimpanzee chromosome 22, previously known as chimpanzee chromosome 23 (chimpanzee chromosomes 2 and 3 was renamed chromosomes 2a and 2b corresponding to the human chromosomes).

Chimpanzee, our closest living relative (Caccone and Powell 1989; Ruvolo 1997), is estimated to have diverged from the human lineage approximately 5.5 million years ago (Mya) (Goodman 1999). There are two species of chimpanzees, the common chimpanzee *Pan troglodytes* and the pygmy chimpanzee or bonobo *Pan paniscus* (Olson et al. 2002). These two chimpanzee species are estimated to have diverged from each other approximately 2.5 Mya (Olson et al. 2002).

Human and chimpanzee share extensive similarities, but the most exciting and valuable information that may be obtained by comparing the two is the genome sequences that underlie the striking differences in anatomy, cognitive ability, physiology and pathology between human and chimpanzee. With the availability of the human genome sequence (IHGSC 2001, 2004; Venter et al. 2001), sequencing of the chimpanzee genome will enable sequence comparison between the two that will be instrumental in, for example, determining the genetic differences that underlie the uniquely human or chimpanzee characteristics in reproductive biology (Gagneux & Varki 2001), their unique vertebral column structure, highly developed human cognitive functions, bipedalism, and use of complex language. In addition, this sequence comparison may reveal the genotypic differences resulting in the high susceptibility of human to *falciparum* malaria (Ollomo et al. 1997), as well as the different rates of epithelial cancers (McClure 1973; Schmidt 1975), Alzeheimer's diseases (Gearing et al. 1994)), and HIV progression to AIDS (Novembre et al. 1997), in humans and chimpanzees.

Previous comparative studies have demonstrated that human and chimpanzee differ by 1.2% to 1.6% in nucleotide sequence (Koop et al. 1989; Chen and Li 2001; Fujiyama et al. 2002). These minor changes in nucleotide sequence could have great consequences as single nucleotide substitutions between the two species, especially the non-synonymous changes that occur in coding regions, and substitutions in cis-regulatory elements could be one of the major contributing factor to the qualitative and quantitative differential gene expression between human and chimpanzee. A classic example of this is the inactivating mutation in human CMP-N-acetylneuraminic acid

hydroxylase thatis functional in great apes and other mammals (Chou et al. 1998; Irie et al. 1998; Angata et al. 2001). This caused the human specific loss of a major sialic acid, N-glycolyl-neuraminic acid (Neu5GC), an integral part of pathogen and toxin recognition on mammalian cell surface, and has been postulated to increase human susceptibility to pathogens and epithelial neoplasms including carcinomas of the breast, ovary, stomach, lung, colon, pancreas and prostate (Muchmore et al. 1998; Varki 2000; Angata et al. 2001).

Comparative G-banding, fluorescence *in situ* hybridization (FISH), and long-range PCR studies have demonstrated that genomic rearrangements such as chromosomal fusion (Yunis and Prakash 1982), pericentric inversions (Nickerson and Nelson 1998), large scale segmental duplication (Bailey et al. 2002), insertions and deletions (Frazer et al. 2003; Watanabe et al. 2004) have occurred since the divergence of human and chimpanzee from their common ancestor. In some instances, these genome rearrangement events caused expansion or contraction of gene families. For example, in contrast to the chimpanzee, the human lineage underwent a duplication of V_{κ} immunoglobulin light-chain genes (Ermert et al. 1995) as well as the olfactory receptor gene family (Trask et al. 1998). Another example is that humans have 8 copies of the keratinocyte growth factor (KGF) gene while chimpanzees only have five (Zimonjic et al. 1997).

By sequencing the syntenic regions of human and chimpanzee chromosome 22 it may be possible to identify the emergence of new genes or expansion of gene families, to quantitate gene loss or contraction of gene families, to locate inactive genes or pseudogenes, to determine changes in regulatory sequences, and to detect genomic

rearrangements that are unique to each species.

The second part of this focus of my Ph.D. research was the comparative sequence analysis of the targeted regions in human chromosome 22 with multiple model species that are evolutionarily distant from human. Multiple species comparative sequence analysis is a powerful structural annotation method to identify functional elements in the human genome such as protein coding genes that have been conserved through evolution. This approach complements computational gene finding methods, often helps identify novel genes that were not identified by gene prediction programs (Roest et al. 2000; Venter et al. 2001; Jaillon 2004), and can locate conserved sequences outside of coding regions that could control gene expression, or be involved in gene imprinting, chromosome packaging and chromosome pairing (Hardison 2000; Pennacchio & Rubin 2001). For this purpose, selected BAC and PAC clones for baboon, cow, mouse and zebrafish were sequenced, and analyzed along with genomic sequences, ESTs and cDNA sequences that were available publicly in GenBank, but sequenced by others. The model species compared were:

Baboon

The olive baboon *Papio anubis,* an old world monkey, is estimated to have diverged from the human lineage approximately 25-40 Mya (Stewart and Disotell, 1998; Goodman 2000). The baboon serves as an excellent out group for human-chimpanzee comparison. Previous studies indicated that human and baboon are highly similar in genomic DNA sequence (Caccone and Powell 1989) and gene organization (Graves et al. 1995). They were also found to be very similar in physiological characteristics (Blanjero 1993; Van deBerg and Williams-Blangero 1996) particularly

in the neurophysiological functions (Kaplan et al.1995; Carey and Rice 1996). The major difference is human has 23 pairs but baboon has 21 pairs of chromosomes in the diploid genome. Human chromosome 2 is a fusion of baboon chromosomes 12 and 13 (Ijdo et al. 1991), baboon chromosome 3 a fusion of human chromosome 7 and 21 (Best et al. 1998), baboon chromosome 7 a fusion of human chromosome 14 and 15 (Rogers and Hixson 1997; Rogers and VandeBerg 2001), and baboon chromosome 10 a fusion of human chromosome 20 and 22 (Rogers et al. 2000). Baboon had been used successfully to study human conditions including cholesterol metabolism (Cox et al. 2002), cortical bone thicknes and peak bone density (Kammerer et al. 1995), osteoporosis (Jerome et al. 1986) and aging (Martin et al. 2002; Jayashankar et al. 2003).

Cow

The domestic cow, *Bos taurus*, estimated to have diverged from the human lineage approximately 85 Mya, has been subjected to selective pressures associated with domestication, e.g. meat and milk production, their ability to assist humans in chores such as pulling plows and carrying loads, their durability in different climates and their resistance to diseases (Gibbs et al. 2002). It also has long been a useful animal model in biological research, especially those studies pertaining to human health. Several important hormones and their effects that were first identified and demonstrated in cow, include parathyroid hormone (Collip 1925), warfarin (Stahmann et al. 1941), and luteinizing hormone (Wiltbank et al. 1961). The first amino acid sequence of insulin was that of bovine insulin used to treat human diabetes (Sanger et al. 1995; Sanger 1959). Bovine developmental and reproductive research also has contributed to the

development of reproductive techniques administered to human such as superovulation, oocyte culturing, *in vitro* fertilization, embryo maturation, transfer and freezing (Brackett et al. 1982; Robl et al. 1987; Iritani and Niwa 1977; Polge et al.1949; Phillips 1939; Johnson et al. 1987)

Data from genetic mapping project and small scale DNA sequence comparison had demonstrated that synteny is much more conserved between humans and cows than between human and mice or rats (Band et al. 2000; Gibbs et al. 2002; Green 2002; MGSC 2002; RGSPC 2004).

Mouse

The mouse, *Mus musculus*, a placental mammal, is one of the most well understood laboratory animal models. Although there are huge morphological and anatomical differences between humans and mice, detailed analysis revealed many similarities in organ systems, physiological homeostasis, reproduction, behavior and susceptibility to diseases between the two. As a result, the mouse has been used as a genetic model of human diseases for over a century (MGSC 2002), and is widely utilized as a research model for studies in embryonic development, behavior, metabolic disease, and cancer (Paigen 1995; Rossant & McKerlie 2001Bradley 2002). There also are enormous numbers of inbred mice strains available, hundreds of spontaneous mouse mutations were characterized, and various techniques for random mutagenesis, transgenic, knockin and knockout of genes have been developed (Hogan et al. 1994; Silver 1995; Joyner 1999; Copeland et al. 2001; Yu & Bradley 2001). Thus, sequencing the mouse genome was set as a high priority in the Human Genome Project to accompany the human genome sequencing and our laboratory was a major contributor

to the draft sequence of the mouse genome that was published in December 2002 (MGSC 2002).

Rat

The rat, *Rattus norvegicus*, was the first mammal domesticated for scientific research purposes, with the earliest record of its usage as a laboratory animal model as far back as 1821 (Hedrich 2000; RGSPC 2004). Ever since, the rat has been an ideal model system in various aspects of human medical research (RGSPC 2004), including surgery (Kuntz et al. 2002), transplantation (Kitagawa et al. 2002; Sauve et al. 2002; Wang et al. 2003), cancer (Alves et al. 2003; Liu et al. 2003) diabetes (Jin et al. 2003; Ravingerova et al. 2003), psychiatric disorders (Talor et al. 2002; Smyth et al. 2002; McBride et al. 1998), neural regeneration (Crisci et al. 2002; Ozkan et al. 2002), wound (Fray 2003; Petratos et al. 2003), bone healing (Hussar et al. 2001), space motion sickness (Yang et al. 2002), cardiovascular disease (Forte et al. 2003; Komamura et al. 2003; McBride et al. 2004), and drug development (Kastleleijn-Nolst et al. 2003; Malik et al. 2003; Kostrubsky et al. 2003; Lindon et al. 2003). In addition, several hundred inbred strains of *Rattus norvegicus* has been developed by selective inbreeding.

Tiger pufferfish

The tiger pufferfish, *Takifugu rubripes*, a marine fish that can grow up to 70 cm in length (Aparicio et al. 2002), is estimated to have diverged from a common ancestor with mammals approximately 450 Mya (Hedges 2002). Unlike other model organisms such as mice and rats that have long history as laboratory animal models, the tiger puffer fish has previously been known only to be a culinary delicacy in eastern Asia. However, with a genome of approximately 365 million base pairs, it is only about one-

nineth the size of human genome (Brenner et al. 1993; Aparicio et al. 2002; IHGSC 2001), and has extensive homology with the human genome (Baxendale et al. 1995; Trower et al. 1996; Venkatesh et al. 1998; Gellner et al. 1999; Aparicio et al. 2002). Its compact genome size and the remarkable homology between this teleost fish and humans were the major reasons why it was selected for comparative genomic sequencing (Aparicio et al. 2002). The tiger puffer fish genome sequence was published in 2002 as only the second vertebrate genome completed after the human genome (Aparicio et al. 2002).

Spotted green pufferfish

The spotted green pufferfish, *Tetraodon nigroviridis*, is a fresh water pufferfish that is estimated to have diverged from a common ancestor with the *Takifugu rubripes* approximately 18-30 Mya (Jaillon 2004), and from a common ancestor with the mammals approximately 450 Mya (Hedges 2002). The draft sequence of approximately 350 Mb *Tetraodon nigroviridis* genome was published in 2004 (Jaillon 2004). Its comparison with the human genome had helped identify approximately 900 previously unannotated human genes (Jaillon 2004), and revealed an ancient whole genome duplication (WGD) had occurred in the ray-finned fish lineage (Jaillon 2004).

Zebrafish

The zebrafish, *Danio rerio,* genome is estimated to be approximately 1.7×10^9 bp (Butler 2000). Conserved synteny, uninterrupted homologous segments containing 2 or more genes conserved between human and zebrafish, has been characterized through several gene and EST mapping projects (Postlethwait and Talbot 1997; Amores et al. 1998; Postlethwait et al. 1998; Gates et al. 1999; O'Brien et al.1999; Postlewait et al.

2000 Barbazuk et al. 2000). Presently the zebrafish genome is being sequenced by the Wellcome Trust Sanger Institute, as it also is an ideal model system in which embryonic gene expression studies can be performed.

1.4.4 Focus II: Gene expression profiling using Zebrafish

The second focus of this research is gene expression profiling using zebrafish whole mount *in situ* hybridization method. As vertebrates, human and zebrafish shares similarities in body plans and developmental constructs, thus making zebrafish an excellent model to study human orthologous genes that are expressed in developmental stages. Locating and timing human orthogous gene expression in zebrafish development is the initial step of a systematic experimental design that will allow subsequent studies into the function of the unknown but predicted orthologous genes. Initially, a portion of my research was devoted to the design of exon specific DNA probes and the development of a robust, 96-well format, high throughput protocol for large scale screening of zebrafish gene expression in different developmental stages. Subsequently, these techniques then were used to identify the embryonic expression provile of several zebrafish orthologs of human chromosome 22 genes.

Zebrafish as a model system

Zebrafish first was studied and developed as a model by George Streisinger at the University of Oregon, in Eugene, Oregon in the 1960s. This system was virtually unnoticed by the larger scientific community for almost 2 decades until its potential as a model for vertebrate embryogenesis and development was demonstrated by a series of elegant developmental studies in late 1980s (Kimmel 1989). Its unique features,

including its relatively small size, (1mm in diameter as a fertilized egg and up to 5cm in length as an adult), the availability of large quantities of embryos (each female fish lays up to 200 eggs every week), its transparent externally fertilized embryos that allow observations of its developing organ system, the viability of embryos outside the chorions before hatching, and the relatively short period of life cycle, make it an ideal model system for developmental studies. Following two large scale mutational screens that were performed in the 1990s, thousands of zebrafish mutations that result in specific developmental defects were discovered (Haffter et al. 1996; Driever et al. 1996).

Once the zebrafish was established as a vertebrate developmental model, efforts then were focused on investigating the genetic relationship between human and zebrafish. Detailed molecular analysis revealed correlations between the zebrafish mutations and human diseases or developmental defects (Zon 1999). For example, the phenotype of the zebrafish *sauternes* (*sau*) mutant was discovered to be equivalent to the human congenital sideroblastic anemia (Brownlie et al 1998), as in both humans and zebrafish, this phenotype is attributed to mutations in the erythroid synthase d-aminolevulinate synthase (ALAS-2) gene. This discovery has made zebrafish the first animal model for human congenital sideroblastic anemia. A very similar phenotype of anemia also was found later in the zebrafish *weissherbst* (*weh*) mutant. Molecular analysis of this mutation has revealed that *weh* encodes a novel iron transporter that is conserved among vertebrates (Donovon et al. 2000). The phenotype for the zebrafish *yquem* (*yqe*) mutant also was found to be identical to the human hepatoerythropoietic porphyria, and the phenotype in both human and zebrafish was found to be the result of

the mutation in the uroporphyrinogen decarboxylase (UROD) gene (Wang et al. 1998) and the *yqe* mutant therefore represents the first animal model for human hepatoerythropoietic porphyria. The phenotype of zebrafish mutant *gridlock* (*grl)* resembles the human malformation coarctation of the aorta (Weinstein et al. 1995). This discovery led to identifying a new regulator of cardiovascular development in humans and other vertebrate embryos (Zhong et al. 2000). Thus, it now is well established that there is a direct link between zebrafish mutations with many human diseases and developmental defects.

Chapter 2 Methods and materials

2.1 DNA sequencing

2.1.1 DNA libraries and sources

Chimpanzee BAC clone libraries from two individual chimpanzees (*Pan troglodytes),* the RPCI-43 library made from the DNA of a male chimpanzee name Donald constructed by Dr. Peter deJong at the Roswell Park Cancer Institute, and the PTB1 made from the DNA of another male chimpanzee name Gon constructed by Dr. Asao Fujiyama at the RIKEN Genomic Sciences Center and the National Institute of Genetics in Japan were used in the present studies. Since these BAC clones previously had been end-sequenced and mapped to the human genome (Fujiyama et al. 2002), those mapped to human chromosome 22 were selected for sequencing.

Zebrafish PAC clones from the BUSM1 PAC library produced by Dr. Chris Amemiya at Virginia Mason University were mapped to human chromosome 22 using a pooling PCR method in collaboration with Dr. Han Wang from the Department of Zoology at University of Oklahoma. Here, exon specific primers were picked from zebrafish ESTs that matched human chromosome 22 exons and synthesized on a MerMade oligonucleotide synthesizer (BioAutomation Corporation). These primers were used to amplify orthologous zebrafish exons from the pooled zebrafish PAC clones. Hierarchical pooling PCR that produced the desired product pointed to the 384 well microtiter plate, the row, and the column and ultimately the exact well of the PAC clone desired.

Baboon BAC clones from RPCI-41 Male Olive Baboon BAC library also

produced by Dr. Peter deJong were mapped to human chromosome 22 and obtained from either Dr. Tamim Shaikh at the Children Hospital of Philadelphia or from Dr. Evan Eichler at Case Western Reserve University. Cow BAC clones from the RPCl-42 Male Bovine BAC library also produced by Dr. Peter deJong were mapped to human chromosome 22 and obtained from Dr. Harris A. Lewin at the W.M. Keck Center for Comparative and Functional Genomics, Urbana, Illinois.

2.1.2 Random shot-gun sequencing strategy

Multiple species BACs that were syntenic to human chromosome 22q12 were sequenced via the random shotgun strategy (Roe 1997) followed by directed closure and finishing.

This random shotgun sequencing strategy entailed amplifying BAC clones in *E.coli* followed by isolation using the double acetate, alkaline lysis method. The purified DNAs then were randomly sheared in either a nebulizer or a Hydroshear. The single stranded ends of the nebulized DNA fragments were repaired by T4 DNA polymerase and the Klenow fragment of E. coli DNA polymerase, and the blunt-end DNA fragments were phosphorylated using T4 DNA polynucleotide kinase. The DNA fragments then were separated on a low melting agarose gel and those with a size range between 1.5-4kb were excised and recovered by phenol extraction. After ligating the DNA fragments into pUC/Sma1 vector using T4 DNA ligase, the ligation mix was transformed to *E.coli* competent cells to produce subclone containing colonies. After overnight growth, the white colonies were picked and grown in TB media in 96 wells microtiter plates. The subclone DNA then was isolated by acetate alkaline lysis,

followed by ethanol precipitation. The isolated double stranded DNA subclones were sequenced using the TaqFS DNA polymerase catalyzed reaction with fluorescent-labeled ET terminator (Amersham Biosciences). After incubation under cycle sequencing conditions and ethanol precipitation, the fluorescent-labeled DNA fragment set was loaded onto ABI 3700 capillary gel sequencers, for electrophoresis and signal detection. The data was base called by the computer program Phred (Ewing et al. 1998) and then assembled into contigs using the computer program Phrap (Green 1993). At least 7-fold sequence coverage usually was needed to ensure a high quality sequence (Green 1993; Ewing et al. 1998; Ewing et al.; Ewing and Green 1998) prior to gap closure and finishing.

Gaps that were not closed after 7-fold coverage were subjected to gap closing strategies. For the gaps that were covered by existing subclones in the subclone library, the primer walking method was utilized. Here, custom synthetic primers were used to extend the sequence on the template subclones. Several rounds of primer picking and walking typically were required. Gaps without subclone coverage could be amplified by PCR utilizing custom synthetic primers and the target BAC DNA as template. When the PCR product was shorter than 2kb it could be either directly sequenced using the synthetic PCR primers or cloned into pUC vector and sequenced with universal primers followed by primer walking. If the PCR product was over 2kb, it was nebulized and subcloned into pUC followed by sequencing and separate assembly of the subcontig. When there was no subclone covering the gap or the attempts to sequence off the target BAC and PCR products failed, the BAC DNA was renebulized using a lower pressure and lower temperature to generate larger DNA fragments (4-8kb) that then were

subcloned into pUC and end-sequenced. When a particular subclone was identified to be covering a gap of interest, this subclone was renebulized, subcloned into pUC and subclones from this new library then were sequenced. Finally, if the sequence contained larger repeats or are GC-rich regions that were difficult to close, 7-deaza-dGTP was used to amplify the region via PCR. Additives such as dimethyl sulfoxide (DMSO), betane, glycerol and formamide sometimes was added to the PCR mix to inhibit primer dimmer formation, reduce template secondary structure, stabilize the enzyme, or enhance primer template binding. Finally, the Applied Biosystems Big Dye, the dRhodomine and dGTP mixes were used to sequence if all other attempts failed.

2.2 Sequence analysis methods

2.2.1 Assembly programs

Automated sequencers obtain information from slab or capillary gels in the form of digitized signals. To make the information useful, the corresponding bases were identified, and then the individual sequences had to be assembled, and visualized. For these purposes, the programs Phred, Phrap, consed and exgap were utilized.

Phred and Phrap

Phred is a base caller which applies a four-phase procedure to identify the bases that are represented by traces from automated sequencers (Ewing et al. 1998; Ewing & Green 1998). Each base that is called is assigned a statistical quality value in the form of error probability. Phrap, an assembly program (Green 1993), then assembles the sequence output from phred to generate contiguous sequences that overlap each other based on several criteria, including repeat sequences length and quality threshold.

Consed

Consed is a graphical tool and a sequence editor intended to aid sequence finishing (Gordon et al. 1998). It allows one to view sequence assemblies, navigate the assemblies, view traces of sequences, tag or edit sequences that represents a misassembly and other problems.

Exgap

Exgap is a visualization tool that was developed by Dr. Axin Hua at the Advanced Center For Genome Tecnology. It enable visualization of contigs, and shows both their order and orientation.

2.2.2 Gene prediction and repeat masking Programs

One crucial step in genome analysis is to identify genes that are encoded in a genome. One way to achieve this is to utilize gene prediction programs such as Genscan and Fgenesh, and the other is to apply homology search and alignment between evolutionary related species. During this process it is crucial to identify repeat elements in genome sequences using Repeatmasker and to mask to reduce false assmblies.

Genscan

Genscan uses the Generalized Hidden Marlov Model approach for gene prediction (Burge & Karlin 1997). Genscan analyzes both strands of a double stranded genomic DNA sequence to identify distinct functional units of eukaryotic genes such as exon, intron, splice site, 5' and 3' untranslated region, and promoter. In each analysis it is able to identify multiple complete genes or partial genes. The program also is designed such that the intron/exon donor and acceptor sites are inter-dependent.

Fgenesh

Fgenesh is a program for predicting multiple genes in a genome sequence which is based on the Hidden Marlov Model similar to that of Genscan (Solovyev et al. 1994; Salamov & Solovyev 2000). However the main difference between Fgenesh and its analogs such as Genscan is in its scoring system, that has a much higher weight given to a defined sets of signals such as splice start sites of genes, as opposed to potential coding sequences highlighted by conserved sequences (Salamov & Solovyev 2000).

RepeatMasker

Repeatmasker is a computer program that identifies repeat elements in a genome sequence such as LINE, SINE, and LTRs (Smit & Green) by comparing a DNA

sequence to a list of known repeat elements. The program producesa copy of the original sequence replacing the repeat elements with Ns.

2.2.3 Alignment programs and visualization tool

A central activity in analyzing genome sequences is to compare them with sequence(s) from other species. This is instrumental in identifying important components in the genomic sequences because functional elements such as exons and cis-regulatory elements have a tendency to retain sequence similarity across related species, while genomic regions that are free from functional constraint are likely to diverge from each other.

The first step to compare genomic sequences entails aligning the sequences, or matching the bases in one sequence to the other in order to identify the similarities and differences. Alignments can be categorized into two broad categories: local alignments and global alignments. A local alignment uses a series of subsets from one sequence to search for similarities in the other sequence without regarding the position of this local region relative to other subsets. Segments with high similarities can be aligned without considering the entire sequence. On the other hand, a global alignment searches sequentially increasing similarities from the beginning of the sequences to the end and attempts to match these sequences even when parts of the alignment have low sequence similarity.

During the course of this research I used the Smith-Waterman (Smith and Waterman 1981), BLAST (Altschul et al. 1990, 1997), BLASTZ (Schwartz et al. 2003)

local alignment methods and the Align (Myers and Miller), Avid (Bray et al. 2003), ClustalW (Thompson 1994) global alignment methods.

BLAST

The **B**asic **L**ocal **A**lignment **S**earch **T**ool (BLAST), a search algorithm for finding ungapped, locally optimal sequence (Altschul et al. 1990, 1997), is used for sequence similarity search, gene structure and genetic feature identification. The BLAST family of programs consist of BLASTn, BLASTx, BLASTp, tBLASTn, and tBLASTx. BLASTn compares a nucleotide query sequence against a nucleotide sequence database. BLASTx compares the six-frame translations of a nucleotide query sequence against a protein sequence database. BLASTp compares an amino acid query sequence against a protein sequence database. tBLASTn compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames. tBLASTx compares the six reading frame translations of a nucleotide query sequence against the six reading frame translations of a nucleotide sequence database.

Once regions of the orthologous human 22q genes were determined in zebrafish, rat, mouse, cow, baboon and chimpanzee, they were subjected to BLAST homology alignment against the known and predicted genes of human 22q, to reveal the extent of homology between these evolutionarily distinct species. As we predict that exons will be highly conserved among the organisms while the intron region will vary, depending on evolutionary distance, the stringency of an alignment could be increased by using the tBLASTx program. tBLASTx translates both query and subject nucleotide sequence in all six reading frames into amino acid sequence allowing conserved regions to be more

easily revealed. Homology alignment of genomic sequences to cDNA sequences also is useful to reveal exon regions since only exon sequence will be present in cDNA.

Crossmatch

Crossmatch, a program for rapid DNA or amino acid sequence search and alignment tool that is included in the phredPhrap program package (Green copyright 1993-1996) utilizes the local alignment Smith and Waterman algorithm (Smith and Waterman 1981) to show conserved sequence regions.

SSAHA

The **S**equence **S**earch and **A**lignment by **H**ashing **A**lgorithm (SSAHA) is a search algorithm for very rapid matching and alignment of closely related DNA sequences (Ning et al. 2001). Although only effective for sequences that have more than 90% similarity, SSAHA completes an analysis quickly as it converts sequences into a `hash table' data structure, that then can be searched for matches.

Sim4

Sim4 is an alignment program specifically designed for aligning cDNA or mRNA to genomic sequences and was useful to specify the correct position of the boundary of exons and introns in the genomic sequence.

Spidey

Spidey, another alignment program for aligning cDNA or mRNA to genomic sequences, uses BLASTn and DotView, both local alignment tools, to generate alignments in a multi-step procedure in which a high-stringency BLAST would first be performed to identify the best genomic windows. Then, a new BLAST with lower stringency would be performed to align the cDNA or mRNA to the identified genomic

windows. Finally, the boundaries of the alignment would be adjusted to ensure exons are non-overlapping, and are adjacent to splice donors and splice acceptors. These features in Spidey helps to avoid exons of paralogs and pseudogenes, and to specify the correct boundaries of exons and introns in genomic sequence.

ClustalW

The ClustalW (Thompson 1994) program can perform a global multiple sequence alignment by obtaining the pair wise alignment of individual sequences, as well as calculating the overall multiple sequence alignment and generating the alignment scores needed to produce a phylogenic tree.

Mega 3.1

Mega 3.1 (Kumar, Tamura, Nei 2004) is a suit of programs for sequence editing and alignment, as well as phylogenetic and molecular evolutionary analyses. Tasks that requires multiple programs to accomplished such as retrieving sequences from Genbank, aligning the sequences, estimating the evolutionary distances of the sequences, building phylogenetic trees based on the alignment and testing the reliability of the tree can be accomplished on one platform in Mega 3.1.

PipMaker

PipMaker is a web-based (http://bio.cse.psu.edu/pipmaker) sequence comparative tool that uses BLASTz to generate sequence alignment and identity comparison between two sequences. The alignments are displayed as a percent identity plot (pip) in which a panel of dots that specify the degree of sequence identity between the compared sequences, aligned with the features of the reference sequence such as the exons and interspersed repeats labeled on top of the panel. PipMaker also can display a

2-dimensional dot plot showing the alignment with the two sequences and it also can produce a text-based alignment and a listing of the coordinates of the aligned segments.

Vista

The Vista website (http://gsd.lbl.gov/Vista/index.shtml) offers 5 distinctive categories of computer programs and databases for comparative genomics (Mayor et al. 2000). The mVista multispecies alignment tool, allows alignment of up to megabases of genome sequence, and present a visual representation of the alignment with annotation information. The rVista tool can detect regulatory sequences in a given genomic sequence using a transcription factor database search. The GenomeVista tools aligns a given genome sequence to whole genome assemblies and is useful to detect syntenic regions. The PhyloVista multi-species alignment also can calculate phylogenetic relationships. In addition, the VistaBrowser contains a graphic representation of pre-aligned whole genome assemblies for several species and can be focused on alignment information for a region of interest. Avid (Bray et al. 2003), a global recursive alignment program, is the alignment engine in all of the above Vista tools.

Alignment between human and chimpanzee sequence

Alignment of chimpanzee sequence to human chromosome 22 was done with a 'hash table' data structure approach similar to that used in SSAHA. When the minimum match value was set at 15, the chimpanzee sequences that were aligned to human sequenced could be chained and assembled, and subsequently visualized using either Pipmaker or Vista.

Multiple species alignment

The assemblies that were used in the multiple species alignment were the NCBI Mouse Build 33 (freeze May27 2004) using the data produced by the Mouse Genome Sequencing Consortium (MGSC), Zebrafish assembly version 4 (Zv4, freeze May 17, 2004) produced by Sanger Center; Rat genome assembly RGSC 3.1 produced by the Rat Genome Sequencing Consortium (RGSC), Fugu V. 3.0 (freeze August 26, 2002) produced by International Fugu Genome Consortium (IFGC) and the Tetraodon7 assembly produced in a collaboration between Genoscope and Broad Institute (MIT). The alignments of these sequences were done using both multiPipmaker and Clustal W.

2.3 Zebrafish whole mount *in situ* hybridization

2.3.1 Embryos collection and processing

Zebrafish were maintained in the light dark controlled (14 hours light, 10 hours dark) zebrafish room at the OU animal facility at a constant temperature of 28.5 ˚C. The fish that were at least 3 months old were segregated by sex into separate tanks with up to 15 males or females per tank. One male and one female (sometimes two females) zebrafish were placed in a breeding tank in the afternoon and left overnight in the dark. When the light comes on the next morning, it is a major stimulus for the zebrafish female to lay up to 200 eggs which are immediately fertilized by the sperm emitted by the male. These eggs sink to the bottom of the tank past a screen that is set up to prevent the adult fish from eating the newly spawned eggs. Although adult fish can be bred up to two times per week, the fish usually are separated for at least four days in between breeding.

After the adult fish were returned to their original tanks, the eggs were collected by pouring the entire content of the breeding tank through a strainer. Then the eggs were flushed into a Petri dish, and fish waste, unfertilized eggs and other debries were pipette out. Once eggs were rinsed several times with fresh Holtfreter's solution (Westerfield 2000), and before the embryos reach 24hpf, 1-phenyl-2-thiourea (PTU) in Holtfreter's solution was added to a final concentration of 0.006% to prevent pigmentation of the fish. The embryos then were left in the Petri dish to grow until the desired stage.

Since zebrafish embryos hatch at 72hpf and beyond, fixing embryos before 72hpf requires dechorion process. Here, the embryos in the Petri dish were transferred

into a beaker, and excess liquid removed. Pronase (Sigma Cat. No. P5147) is added to the embryos to a final concentration of 0.5 mg/ml and incubated at 28.5 ˚C (fish room temperature) for 2.5 min with occasional swirling. The pronase then was removed by rinsing the embryos repeatedly with water (~ 200ml total), and the embryos shed their chorion during the rinsing process.

Up to 100 embryos without chorion were pipetted into a 1.5 ml Eppendorf tube. After removing all excess liquid, the tubes were placed on ice to prevent active swimming of the embryos that interferes with the handling processes. Then 500 µl of 4% paraformaldehyde (4% PFA) was added into each tube to rinse the fish and after the first rinse, a fresh 1 ml 4% PFA was added to fix the fish. The tubes were kept at 4 ˚C overnight and the next day the 4% PFA in the tube was removed. The embryos then were dehydrated by serially washing with 1 ml of 1:3, 1:1, and 3:1 methanol : ddH_2O and agitated by shaking on the SpeciMix shaker (Barnstead International, model: M26125) for 5 minutes at room temperature. Finally, the embryos were stored in 100% methanol at -20 ˚C.

2.3.2 Zebrafish genomic DNA isolation

Seven days old zebrafish were collected in a 1.5 ml Eppendorf tubes containing between 50-60 embryos. After placing the tubes on ice to prevent zebrafish movements, excess water was removed and the fish were rinsed twice with fresh ddH_2O, then, 1 ml of DNA extraction buffer containing 10 mM Tris pH 8.2, 10 mM EDTA, 200 mM NaCl, 0.5 % SDS, and 200 µg/mL proteinase K, was added to each tube, to cover all the fish, and the tubes were placed in a 50 ˚C water bath over night.

By the following morning the fish had dissolved into the buffer and an equal volume of TE saturated phenol was added to the mixture, vortexed, and cenrtrifuged (Fisher Micro-Centrifuge Model 235A) for 5 minutes. The top aqueous layer then was transferred into a new tube, and the bottom phenol layer was discarded. An equal volume of TE saturated phenol plus chloroform was added to the aqueous fraction, vortexed and centrifuged for 5 minutes. After again transferring the aqueous layer to a new tube and discarding the bottom layer, chloroform was added to the new tube, vortexed and centrifuged for 5 minutes. The top layer again was transferred to a new tube, and an equal volume of ether was added. After vortexing again and centrifuging for 5 minutes, the top layer was discarded, and the bottom layer left open under the hood over night for the ether to evaporate.

On the third day, 100 µl 95 % ethanol with 0.12 M of NaOAc was added and the mixture was placed on ice for 30 minutes. After centrifug ation (Fisher Scientific Marathon 13 K/M) at 4 ˚C in the cold room, the supernatant was discarded, and the pellet washed with 100 µl 70 % ethanol. The pellet was dried and then resuspended in 100 µl 1:0.1 TE, adjusted to a final concentration of 50 ng per µl based on A_{260}.

2.3.3 Single Stranded oligonucleotide probe making

Two pairs of exon specific primers for each region of interest were picked using PRIMOU, where the first pair of primers was picked from the flanking region of the exons, and the second pair of primers was picked from within the exon region. The DNA fragments were amplified using Polymerase Chain Reaction (PCR) by incubation in a Perkin-Elmer Cetus DNA Thermal Cycler or the Perkin-Elmer Cetus Cycler 9600.

Here the first primer pair was used in first round PCR, with 50 ng Zebrafish genomic DNA as the template in a 50 µl PCR reaction. The PCR conditions were as follows: denaturing the template at temperature 95 ˚C for 3 minutes and 30 seconds; then thirty-five cycles of denaturing temperature at 94 ˚C for 1 minute, annealing temperature at 55˚C for 1 minute, and extension temperature at 72 ˚C for 1 minute. After the 35 cycles, the reaction was incubated at an extension temperature of 72 ˚C for another 1 minute and then the temperature was lowered to 4 ˚C indefinitely to stop the reaction.

When multiple bands or a smear of unspecific PCR products was produced, the touch-down PCR technique was employed. Here, 2 additional cycles with annealing temperature at 65˚C followed by 2 additional cycles with annealing temperature at 60˚C are added to the original PCR cycles. Thus, the reaction was started with 3 minutes and 30 seconds of denaturing temperature. This then was followed by 2 cycles of denaturing temperature at 94˚C for 1 minute, annealing temperature at 65˚C for 1 minute, and extension temperature at 72˚C for 1 minute, followed by 2 cycles of: denaturing temperature at 94˚C for 1 minute, annealing temperature at 60˚C for 1 minute, and extension temperature at 72˚C for one minutes, and followed by 35 regular PCR cycles of as described above.

The PCR products were analyzed by electrophoresis on a 2 % agarose gel or on the Caliper AMS90SE (Caliper Technologies Corp.), to determine their size. When the PCR product of the anticipated size was observed either on the agarose gel or the Caliper, it was treated with Shrimp Alkaline Phosphatase (SAP) at a concentration of 2 units for every 5ul PCR product, and Exonulease I (EXO I) at a concentration of 10 units for every 5ul PCR products, at 37˚C for 45 minutes to digest unused primers and

inactivate dNTPs. SAP and EXO 1 are inactivated by raising the temperature to 85˚C for 20 minutes. These cleaned PCR products then were used as the template for the second round of nested PCR. The conditions and cycles of the secondary PCR were the same as for the first round PCR, and after amplification the second round nested PCR products were analysed on a 2% agarose gel or the Caliper AMS90SE for validation of the product size. The PCR products then were cleaned up using SAP and EXO 1 as described above.

Both the first round and second round PCR products were sequenced using 5 µl of PCR product and 100 pmols of original primers that are used for the PCR with 2 µl of ET terminator dye in each sequencing reaction in a 96 well thermocycler plate, by incubating for 60 cycles of denaturing at 95˚C for 30 seconds, annealing at 50˚C for 20 seconds, and extension at 60˚C for 2 minutes. After the reactions were completed, the sequencing products were ethanol precipitated, dried and loaded on the ABI 3700 capillary sequencer or the ABI 377 slab gel sequencer.

Once the sequences were validated, the PCR products were used as template in unidirectional PCR to generate single stranded DNA probes for the *in situ* hybridization. All conditions and cycles were the same as above, but only a single primer and PCR DIG Labeling Mix (Roche Cat. No. 1 585 550) were used for the unidirectional labeling reaction. This Labeling Mix contained 2mM dATP, dCTP, dGTP each; 1.9 mM dTTP, and 0.1 mM digoxigenin-11-dUTP (DIG-11-dUTP). After the labeled products were analyzed through a 2% agarose gel electrophoresis or on the Caliper AMS90SE, they were ethanol precipitated and dissolved in 50% hybridization buffer, for use in the *in situ* hybridization.

2.3.4 *In Situ* hybridization

Whole mount *in situ* hybridization of zebrafish employed was a three days process. On the first day, the embryos were taken out from storage at –20°C, and the methanol was removed. The embryos then were rehydrated by washing with 3:1, 1:1 and 1:3 Methanol : Phosphate Buffered Saline with Tween-20 (PBST). For each wash, the tubes were placed on the SpeciMix shaker for 5 minutes. Then, the embryos were washed 4 times with 1 mL of 1X PBST, by adding in 1 mL PBST and shaking for 5 minutes.

The embryos then were treated with 10µg/mL proteinase K (Sigma) in PBST. The 24 hpf embryos were shaken for 1 minute and 30 seconds, while 48 hpf embryos were shaken for 4 minutes and 72 hpf embryos were shaken for 4 minutes and allowed to stay idle in the tube for 4 minutes. Immediately following this step, the embryos were treated with 2.5 mg/mL glycine in PBST and were shaken for 5 minutes. Then, the embryos were washed with 1X PBST byshak ing 3 times for 5 minutes. After that, the embryos were treated with 4% PFA and incubated at room temperature for 20 minutes, and then washed with 1X PBST and shaken for 5 minutes 6 times.

The embryos then were distributed into 96 wells microtiter plate (VWR Scientific, cat. #62402-933) with approximately 15 fish in each well. The PBST were removed from each well and 200 µL of 50% hybridization buffer (50% formamide, 5X SSC [standard saline citrate], 50 µg/mL heparin, 5 mg/mL Torula Yeast RNA, 0.2% Tween 20, and 10 mM citric acid) was added using the Hydra96 (Robbins Scientific). Suspension and removal of liquid involving the 96 well microtiter plate in all subsequent steps was done using Hydra unless stated otherwise. The microtiter plate

was then placed in a water bath at the designated hybridization temperature (50 °C, 55 °C, 60 °C, or 65 °C) and the fish were allowed to incubate in the 50% hybridization buffer for 2 hours. After 2 hours, the 50% hybridization buffer in the wells was removed and replaced with new 50% hybridization buffer containing the DNA probes. The microtiter plate then was sealed and placed in the water bath overnight.

On the following day, all 50% hybridization buffer with probes were removed using a 12 channel pippette. Then, 250 µL of fresh 50% hybridization buffer was added into the wells using Hydra96. and the microtiter plate was placed in the water bath for 5 minutes. Fresh 50% hybridization buffer was added and the plates were placed at the chosen incubation temperature for 5 minutes. Then, using the Hydra96 for removal and suspension of liquid, sodium based buffer saline sodium citrate (SSC) was added to the embryos bysubjec ting them to 2 washes each of 3:1, 1:1, and 1:3 ratio of 50% hybridization buffer to 2X SSC, and 2 washes with 100% 2X SSC, and l incubation in water bath for 5 minutes after adding each solution. The embryos then were washed 4 times with 0.2X SSC and incubated 15 minutes in the water bath after each new addition of fresh SSC. The phosphate based buffer PBST then slowly was reintroduced at room temperature by washing the embryos twice with 3:1, 1:1, and 1:3 ratio of 0.2X SSC : PBST, and twice with 100% PBST with shaking of the microtiter plate on the TiterPlate shaker shaker (Lab-Line Instrument Inc. Model: 4625) for 5 minutes at room temperature after each addition of fresh solution.

After the washes, all PBST were removed from the wells and replaced with 200 µl of blocking solution (2mg/ml of BSA Sigma A2153 and 0.02mL Normal Sheep Serum in 1 ml PBST). The microtiter plate then was placed on a TiterPlate shaker to

shake for 2 hours at room temperature. After the 2 hours shaking, the blocking solution was removed and 250 µL of blocking solution with Anti-Digoxigenin-AP (Roche Cat. No. 1 093 274), at 1 : 10,000 dilution or 75 mU/ml, was added to each well. The microtiter plate then was sealed and shaken on the TiterPlate shaker at 4 °C in the cold room over night.

On the third day, the blocking solution with Anti-Digoxigenin-AP was removed and the embryos were washed with 1X PBST 16 times by shaking on a TiterPlate shaker at room temperature for 5minutes with each addition. All 1X PBST then was removed from the wells and replaced with 200 µL of staining buffer NTMT (100mM NaCl, 50mM MgCl, 100mM Tris PH 9.5 and 0.1% Tween 20). After shaking on the TiterPlate shaker for 5 minutes at room temperature and repeating this step four times, all staining buffer was removed from the wells and fresh staining buffer with the dyes: 4-Nitro-blue-tetrazolium chloride (NBT, Roche 1383 -213) at a concentration of 4.5 µL per mL and 5-Bromo-4-chloro-3-indolyl-phosphate (BCIP, Roche 1383-221) at 3.5 µL per mL was added. Then after the microtiter plate was sealed with a silver sealer and wrapped with aluminum foil to prevent the dyes from exposure to light, the plate was placed on the TiterPlate shaker. After shaking for 1 hour, staining of the embryos was checked. If staining was observed, the staining step was stopped by removing the staining solution and rinsing the embryos twice with PBST.

Chapter III Results and Discussion

A. Comparative sequence analysis

3.1 Chimpanzee sequence and analysis

3.1.1 Overview

To understand underlying genetic differences between human and its closest living relative, the common chimpanzee *Pan troglodytes,* 32 overlapping BAC clones from 3 different male chimpanzee BAC libraries CHORI-251, PTB1, and RPCI-43 were sequenced, assembled, and compared to the syntenic region of human chromosome 22 at its upper q arm between markers D22s1687 and D22s419 that is approximately 4 Mb and includes 4 low copy repeats (LCR22s), the Immunoglobulin Lambda Light Chain region (IGLL), and the Breakpoint Cluster Region (BCR).

The corresponding chimpanzee sequence in this region is approximately 45,000 bp or 1.1% smaller than the human sequence. The G + C content of the region is 48.52 % for chimpanzee (Table 3.1), close to the G + C content of the orthologous human chromosome 22 region at 48.47%, and slightly higher than that of the entire human chromosome 22 which is 47.8% (Dunham et al. 1999). The interspersed repeats cover 44.24% in the chimpanzee region (Table 3.1) and 43.94% of the human orthologous region (Table 3.1) while noting that in the entire human chromosome 22 the interspersed repeats represent 41.9% of the DNA content (Dunham et al. 1999).

Human				**Chimpanzee**		
Total length:		4048001 bp		Total length:	4003489 bp	
GC level:		48.47 %		GC level:	48.52 %	
Repeat Type	Total Number	Coverage (Bp)	Coverage (%)	Total Number	Coverage (Bp)	Coverage (%)
SINEs:	3648	876778 bp	21.66 %	3692	889972 bp	22.23 %
ALUs	2882	776336 bp	19.18 %	2940	790777 bp	19.75 %
MIRs	766	100442 bp	2.48 %	752	99195 bp	2.48 %
LINEs:	1087	548532 bp	13.55 %	1109	543837 bp	13.58 %
LINE1	656	444751 bp	10.99 %	685	441155 bp	11.02 %
LINE2	392	95711 bp	2.36 %	386	95407 bp	2.38 %
L3/CR1	39	8070 bp	0.20 %	38	7275 bp	0.18 %
LTR elements:	521	256107 bp	6.33 %	498	244116 bp	6.10 %
MaLRs	228	80591 bp	1.99 %	217	77281 bp	1.93 %
ERVL	96	51096 bp	1.26 %	93	46324 bp	1.16 %
ERV_I	174	100170 bp	2.47 %	167	96368 bp	2.41 %
ERV_II	22	24152 bp	0.60 %	20	24045 bp	0.60 %
DNA elements:	386	82520 bp	2.04 %	374	80070 bp	2.00 %
MER1_type	257	49598 bp	1.23 %	249	48748 bp	1.22 %
MER2_type	58	23221 bp	0.57 %	58	22561 bp	0.56 %
Unclassified:	8	14856 bp	0.37 %	11	13098 bp	0.33 %
Interspersed repeats total :		1778793 bp	43.94%	1771093 bp		44.24 %
Small RNA:	20	3289 bp	0.08 %	20	3151 bp	0.08 %
Satellites:	45	18947 bp	0.47 %	46	18372 bp	0.46 %
Simple repeats:	536	45729 bp	1.13%	534	42570 bp	1.06 %
Low complexity:	304	16325 bp	0.40 %	312	16043 bp	0.40 %
Total Repeats		1862014 bp	46.00 %		1850546 bp	46.22 %

Table 3.1: Comparison of GC content and repeat elements between human and chimpanzee sequence.

The BAC-based chimpanzee sequences were aligned and compared to the current human sequence assembly (NCBI Human Chromosome 22 Build34) using a combination of BLASTN, Dot Plot and PIP analyses. The overall sequence and structure in the orthologous region between human and chimpanzee are highly similar,

although regions with insertions and deletions (indels), duplications and inversions were observed In coding regions, these changes sometimes resulted in altering amino acids in the translated proteins and discussed in detail below.

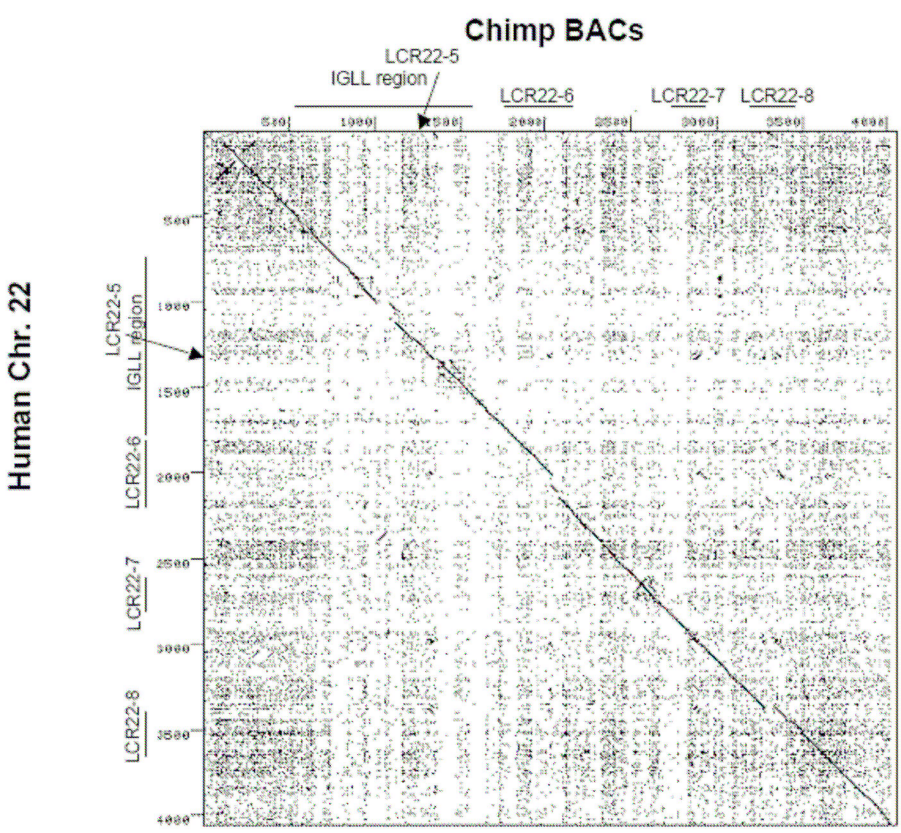

Figure 3.1 A dot plot alignment showing similarities between human and chimpanzee sequence. Each dot on the plot represents a match of at least 40 bases between the two sequences. The diagonal line from upper left corner to the lower right corner of the plot indicates the sequential match between the two sequences, lines perpendicular to the diagonal line indicate inversions, short lines away from the path of the diagonal line indicate duplication, and dots covering the plot indicate repetitive sequences. The dot plot was produced using the program Maxmatch with a stringency of minmatch 40.

3.1.2 Lineage-specific insertions and deletions

Alignment of the chimpanzee sequence to the human sequence, revealed a total of 102 lineage-specific insertions or deletions (indels) over the approximately 4 Mb region. As shown in Figure 3.2, the indels classified ranged in size from a few basepairs to > 50 Kb. For any given indel, an insertion observed in the sequence of one species could be a deletion in the other species or vise versa. For the purpose of this discussion, each indel was considered a lineage specific insertion. The largest indel observed was an approximately 75 kb human insertion previously reported (Robledo et al. 2004) in the IGLL locus when compared to chimpanzee sequence.

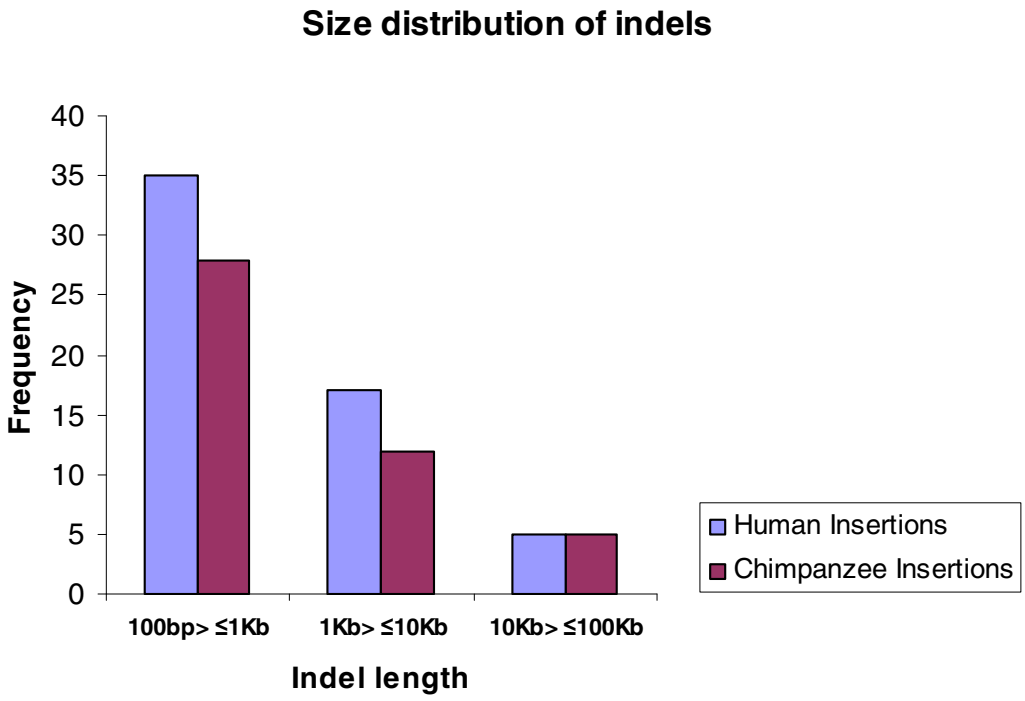

Figure 3.2: Distribution of human and chimpanzee indels by size. All indels are calculated as insertion either in human (blue) or chimpanzee (red). The x-axis shows the size range between 100 bp and 1 Kb, between 1 Kb and 10 Kb, and between 10 Kb and 100 Kb. The y-axis shows the frequency of indels.

The majority of the indels observed were less than 1 Kb, and mainly represented repetitive sequences. This is in agreement with the Repeatmasker human and chimpanzee comparison of SINEs, LINEs, LTRs, and other simple repeats shown in Table 3.1.

3.1.3 Identification of chimpanzee genes

The 4 Mb region of human chromosome 22 investigated in this study encodes a total of 126 genes including 29 known coding genes, 20 putative coding genes, 34 partially duplicated genes and 43 pseudogenes, 1 non-coding genes, in addition to 125 Immunoglobulin Lambda Light Chain segments (Dunham et al. 1999; Collins et al. 2003). Here, a gene was classified as coding when it possessed an undisrupted open reading frame (ORF) and had sequence identity greater than 99% to human cDNA or EST over its entire length. A partial gene typically either was partially identical to a DNA, EST or peptide sequence, or was identical to portion of a coding gene elsewhere in the human genome, and has an undisrupted ORF. Genes were classified as pseudogenes when they had sequence homology to cDNAs, ESTs or coding genes but contained disrupted ORF. Non-coding genes were those encoding small RNA genes, as they represented genes with no ORF and potential antisense sequences (Collins et al. 2003).

To identify chimpanzee genes in this region, human genes and pseudogenes were aligned to the chimpanzee sequence using the program Spidey. Every alignment was checked for exon coverage, intron spacing, splice sites, and percent identity. Every chimpanzee gene finally was confirmed through their syntenic relationship, sequential

order and intergenic spacing relative to other genes in the region. The Fgenesh and Genscan gene prediction programs were used to predict genes in regions of chimpanzee containing insertions, and any predicted genes then were searched against NCBI nr database using BLASTn. This resulted in identification of 29 known coding genes, 20 putative genes, 34 partial genes, 39 pseudogenes, and 1 non-coding gene in the chimpanzee sequence. Here, the syntenic chimpanzee region contained the same number of coding gene, one fewer partial genes, four fewer pseudogenes, and the same number of non-coding gene when compared to the 4 Mb human chromosome 22 region.

3.1.4 Gene Divergence

Every 1:1 ortholog between human and chimpanzee was identified and compared. To calculate the divergence rates between these orthologs, each pair was aligned using ClustalW, and the p-distance, which is the proportion (p) of nucleotide sites at which the two sequences being compared are different, was calculated for each ortholog using the program Mega3.1 (Kumar et al. 2004). Standard error of distance estimates were calculated using the Bootstrap method with 1000 replicates. In this analysis, both exon and intron regions are compared. The divergence of introns served as a control to local random or neutral mutation rate as mutation rate varies across the genome. Therefore it is crucial factor in mutational rate in the intron when divergence rate between genes are compared.

Through evolution, DNA sequences that have functional roles face functional constraints and are evolutionarily pressured to maintain sequence similarity, while sequences that are not functional will not have functional constraints and thus are

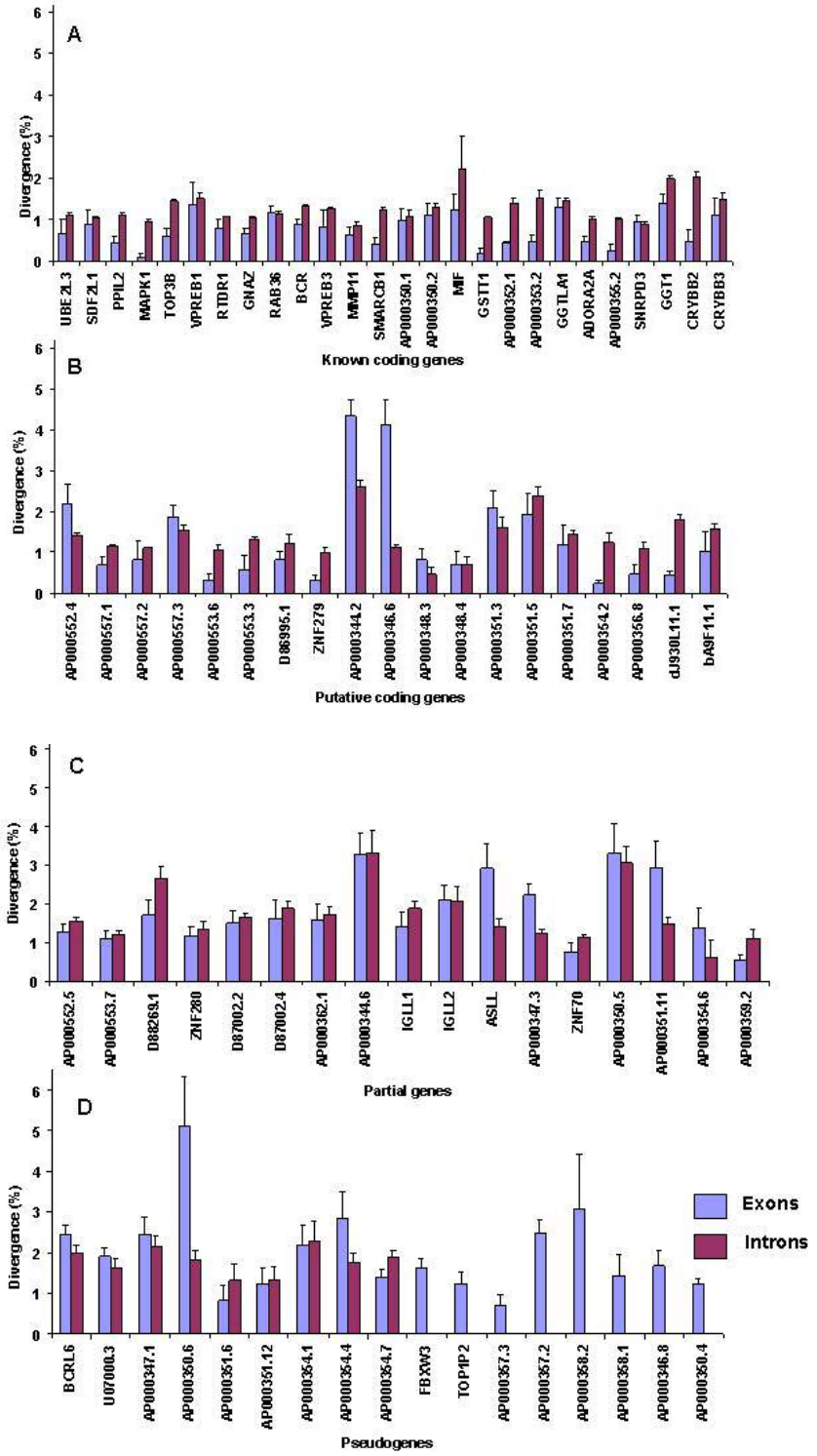

Figure 3.3 Graph showing percent divergence for (A) known coding genes (B) putative genes (C) partial genes and (D) pseudogenes in both the exons and the introns. Processed pseudogenes lacking introns have no intron divergence data. Percent divergence was calculated as p-distance using Mega3.1. Error bars depict standard errors for uncorrected percent divergence.

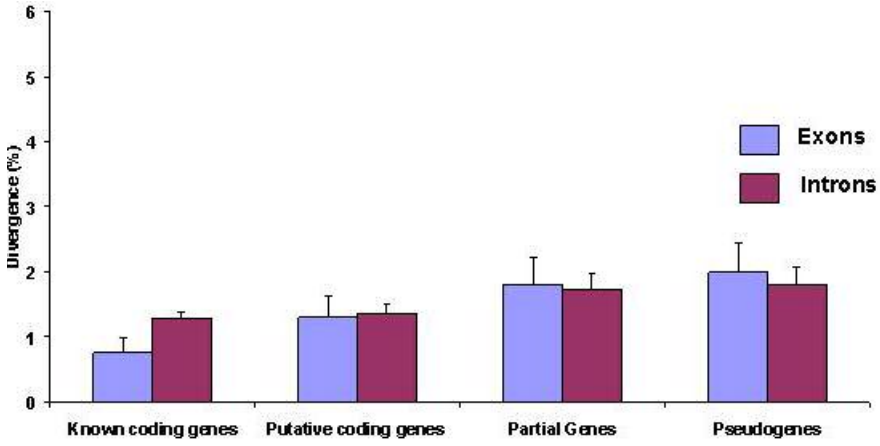

Figure 3.4 Graph showing average percent divergence for known coding genes, putative genes, partial genes, and pseudogenes. Percent divergence was calculated as p-distance using Mega3.1. Error bars depict standard errors for uncorrected percent divergence.

subjected to random mutation. By comparing the coding exons between human and chimpanzees, we were able to identify and characterize differences in the coding sequences between the two species, and in addition to that, differentiate among putative and partial genes between those facing functional constraints and those that are not. Since random mutation rates were different along the chromosomes in different regions, divergent rates of the introns were used as the control for random mutations for the coding sequence. As shown in Figure 3.3 all known coding genes except RAB36 and SNRPD3 have a lower divergence rate between human and chimpanzee in the exons compared to the introns, showing clear functional constraint applied on the exons. The average divergence for the exons of the known coding genes is 0.76% and for introns is 1.29%. The putative coding genes were shown to have higher divergent rate and were less consistant in their divergent pattern when compared to the known coding genes. Six of the genes in this class showed unusually high divergent rates in their exons, a sign of

rapid evolution in the coding regions. The average divergence rates for the exons and introns for putative genes are 1.31% and 1.36% respectively. Overall, the partial genes have a higher divergence rates both in the exons and the introns when compared to the known coding genes and some of the putative coding genes. The average divergence rates for the partial genes are 1.81% and 1.73% repectively in the exons and the introns and similarly, some partially duplicated genes showed unusually high divergence rate in the exons when compared to introns. The pseudogenes also have a higher divergence rate overall, and many having higher divergence rate in the exons than introns. The processed pseudogenes lack introns and therefore they only have divergence rates for their exons. Average divergence rate for pseudogenes were found to be 1.99% and 1.80% in the exons and introns respectively. Functional protein coding genes in this analysis shows clear functional constraints when compared to the random mutational rates represented by the divergence rate of the introns. Coding regions are highly conserved between human and chimpanzee, indicating their gene products likely have similar functions. For some of the putative genes, the partial genes and pseudogenes, signs of the loss of functional constraints when compared to the random mutational rates and the elevated divergence rate may point to lower evolutionary pressure on maintaining their similarity and an accelerated rate of amino-acid-changing base substitution in the coding region, leading to positive selection in the gene evolution.

3.1.5 Non-synonymous (Ka) versus synonymous (Ks) substitution

To assess how the divergence rates observed in coding sequences affect the evolution between humans and chimpanzees in their functional proteins, for each gene

studied, non-synonymous substitution rate (Ka) was calculated, where the Ka is defined as the number of nucleotide changes between the two species that led to amino acid changes as a fraction of all such possible sites. As random mutational rates varies across

chromosomes, demonstrated by an elevated Ka rates between human and chimpanzee by approximately 40% in the distal 10 Mb of chromosomes (TCSAC 2005), we also calculated the synonymous substitution rate (Ks) for each gene, where Ks is defined as the number of nucleotide changes between humans and chimpanzees in the coding region that did not change the amino acid sequences as a fraction of all such possible sites, to normalize Ka for comparison between genes.

The Ka/Ks ratio is an indication of the rate of amino-acid-changing base substitution compared to random mutationial drift, where Ka/Ks <1 indicates a substantial proportion of amino acid changes have been eliminated by negative or purifying selection, Ka/Ks =1 indicates the coding region is subjected to random mutational drift, and Ka/Ks >1 indicates the coding region is undergoing positive selection fixing advantageous amino acid changes (Zhang et al. 2003; TCSAC 2005).

As shown in Figure 3.5, both Ka and Ks values for the known coding genes are lower compared to other classes of genes, and the Ka values of the genes are predominantly lower than Ks values or are non-existent, resulting in Ka/Ks ratios <1 which indicates very low or zero nucleotide changes that led to amino acid changes in these genes. With only MMP11, with a Ka/Ks >1 at 1.34, showing positive selection, the average Ka/Ks ratio between human and chimpanzee for known coding-genes is 0.25, consistant with the values found in previous studies (TCSAC 2005). Under the

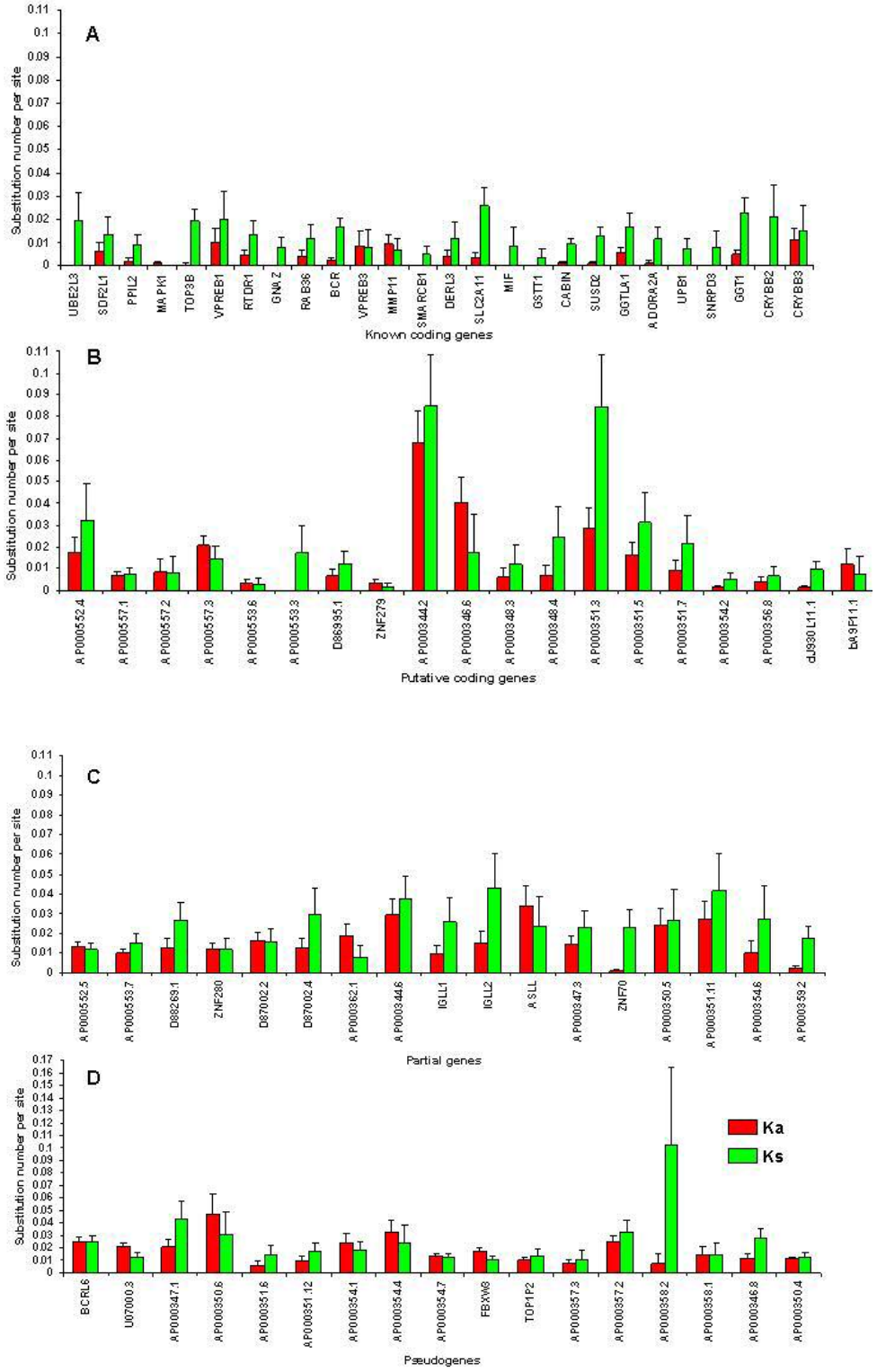

Figure 3.5 Graphs showing substitution number per site at non-synonymous sites and synonymous sites for (A) known coding gene, (B) putative coding gene, (C) partial genes and (D) pseudogenes.

assumption that the Ks values reflect random mutational drift the coding region is undergoing, this result implies that 75% of the amino acid changes between human and chimpanzee in these genes are deleterious mutations and thus are eliminated by natural selection.

Ka/Ks ratios for putative coding genes and partial genes are 0.82, and 0.75 respectively, partly contributed by an increase in genes undergoing positive selection.This is not surprising as many of the genes in these classes are duplicated genes or truncated genes that likely have no or relaxed functional constraints and allowed to mutate and fix advantageous changes at the amino acid level. The pseudogenes have an average Ka/Ks ration of 0.92, and very much like putative coding genes and partial genes, many of them are undergoing positive selection and can be found in the IGLL and LCR22s. This study had demonstrated that these duplicated segments harbors many genes that are evolving rapidly between human and chimpanzee, many of them by accelerated positive selection of amino acid changes.

3.1.6 Amino acid substitution

The known coding genes have the least number of amino acid substitutions between human and chimpanzee, 92 amino acid substitutions were observed among 26 genes. Here, 18 putative coding genes had 108 amino acid changes, while the 17 partial genes have 181 amino acid changes. When these substitutions were investigated in detail, the majority of the amino acid substitutions were observed between hydrophyllic amino acids. In the known coding genes, the next most prevalent changes was found between hydrophobic amino acids, and the least being from hydrophobic to

hydrophyllic or hydrophyllic to hydrophobic substitutions. The substitutions observed among the putative coding genes were found to be similar between hydrophobic to hydrophobic and hydrophobic to hydrophyllic and vice versa, with the least being hydrophobic to hydrophobic changes. Among the partial genes, most substitution was found among hydrophyllic to hydrophyllic, while hydrophobic to hydrophyllic and vice versa changes were greater than hydrophobic to hydrophobic changes.

3.1.7 Immunoglobulin Lambda Light Chain Locus (IGLL)

Immunoglobulins are tetrameric proteins composed of two identical heavy (H) chains measuring approximately 50-70 kDa and two identical light (L) chains measuring approximately 25 kDa linked by disulfide bonds. Each H and L chain contains a variable (V) as well as a constant (C) domain. The heavy chain locus (IGH) is located on human chromosome 14 at 14q32.33, while there are two separate light chain loci, the λ light chain (IGL), located approximately 6 Mb from the centromere on chromosome 22 at 22q11.2 (Dunham et al. 1999) and the κ light chain (IGK) is located on chromosome 2 at 2p11.2. All three immunoglobulin loci consist of multiple immunoglobulin gene components that are rearranged during B cell differentiation, with the IGH locus encoding the variable (V), diversity (D), joining (J) and constant (C) genes and both IGL and IGK encoding the V, J, and C genes.

Germline immunoglobulin entities, V-genes, D-genes, J-genes, and C-genes, are classified as functional, open-reading-frame (ORF), pseudogene, or vestigial (LeFranc 1998). A germline immunoglobulin entity is classified as Functional when it has an undisrupted open reading frame and undisrupted splicing sites, recombination

signals, and/or regulatory elements, while an immunoglobulin entity is classified as an ORF when an open reading frame is maintained but there either are defects in the splicing sites, recombination signals, and/or regulatory elements, or when changes occur in amino acid residues essential for correct folding such as cysteine at amino acid residues 23 (cys23) and 104(cys104), or tryptophan at amino acid residue 41 (trp41). Alternatively when an immunoglobulin entity is located outside of immunoglobulin locus, it is classified as an ORF even when there is no defect as described above because an immunoglobulin orphon is unable to recombine with other immunoglobulin entities to form a functional unit. An immunoglobulin entity is classified as a pseudogene when it contains a stop codon that disrupts the open reading frame that may or may not result from frame shifting mutations. Finally an immunoglobulin pseudogene is termed vestigial when it contains excessive insertions or deletions, stop codons and frame shift mutations, or only remnants of immunoglobulin motifs can be detected (LeFranc 1998).

The human IGLL locus on 22q11.2 spans approximately 1Mb, and encodes for 31 IGLV functional genes, 5 IGLV ORFs, 33 IGLV pseudogenes, 34 vestigial sequences, 7 IGLJ segments, and 7 IGLC genes, 2 non-immunoglobulin coding gene, 6 non-immunoglobulin partial genes, and 16 non-immunoglobulin pseudogenes (Frippiat et al. 1995, Kawasaki et al. 1997, LeFranc et al. 1999, Collins et al. 2003).

3.1.8 Identification of chimpanzee IGLL genes

To identify genes in the chimpanzee IGLL locus, human IGLV, IGLLJ/C genes and pseudogenes, as well as non-immunoglobulin genes and pseudogenes

(Collins et al. 2003) were aligned to the chimpanzee IGLL locus using Spidey. Fgenesh and Genscan gene prediction programs then were used to determine if the chimpanzee encoded any additional genes in inserted regions that then were searched against NCBI nr database using BLASTn. In total, orthologs of 23 functional IGLL genes, 5 IGLL ORFs, 47 IGLL pseudogenes, and 24 IGLL vestigial genes, in addition to 2 non-immunoglobulin coding gene, 7 non-immunoglobulin partial genes, and 13 non-immunoglobulin pseudogenes were identified in the chimpanzee IGLL locus. Each of this entity was named according to their orthologous human IGLL nomenclature (LeFranc 2001, Collins et al. 2003).

The translated amino acid and nucleotide sequences of the 23 chimpanzee functional IGLV genes were aligned to their respective human IGLV genes using ClustalX, as shown in Figure3.6. These IGLV genes were checked for open reading frame and essential IGLV elements such as cys23, cys104, and trp41. Among the 23 IGLV genes in chimpanzee, 3 were found to have either premature stop codons introduced by point mutations, essential elements missing caused by frame shift or point mutations. IGLV2-18 was found to contain all essential elements but a stop codon was introduced in the signal peptide domain by point mutation. Gene IGLV3-9 was found to have a 4 bp deletion prior to cys104 that caused a frame shift mutation while IGLV3-22 was found to be missing trp41 because a transversion occurred in the second base of codon 41 that changed the amino acid residue from trp to Ser. As a result of defects found above, the three genes that are classified as functional in human are classified as pseudogene in chimpanzee and none of these had been reported in the similar position in the human orhthologs. However, in chimpanzee, cys104 of IGLV3-32 was replaced

by tyr104, an observation also seen in the human IGLV3-32 haplotypes indicating a conserted genomic change and a conserved phenotype in both primates.

3.1.9 Phylogeny of IGLV genes

When the nucleotide sequences of exon 2 for 74 chimpanzee IGLV genes were aligned against each other using ClustalX by first aligning the 20 functional IGLV genes, and then adding the 5 ORFs and 49 pseudogenes to the alignment. The phylogenetic tree generated using the PAUP neighbour-joining distance method shown in Figure 3.7 indicates divergence of genes from ClanI through ClanV, giving similar phylogenetic relationships of the chimpanzee IGLV genes to that of the human IGLV genes (Kawasaki 2000).

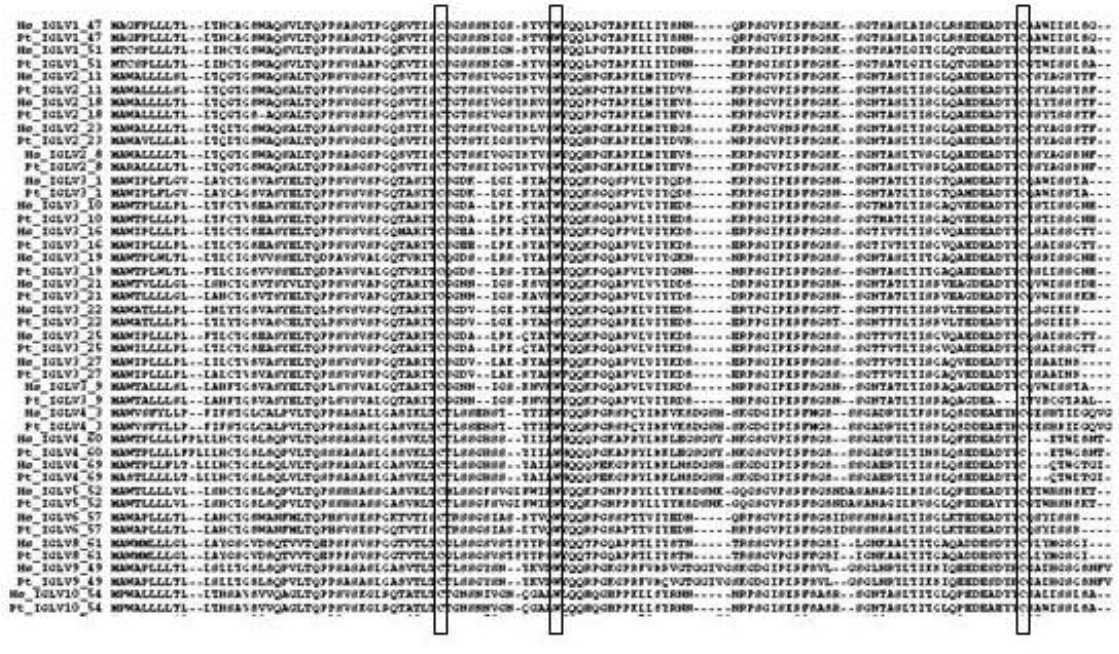

Figure 3.6 Multiple alignment of human and chimpanzee IGLV amino acid sequence. Essential amino acid residues cys23, cys104 and trp41 are highlighted.

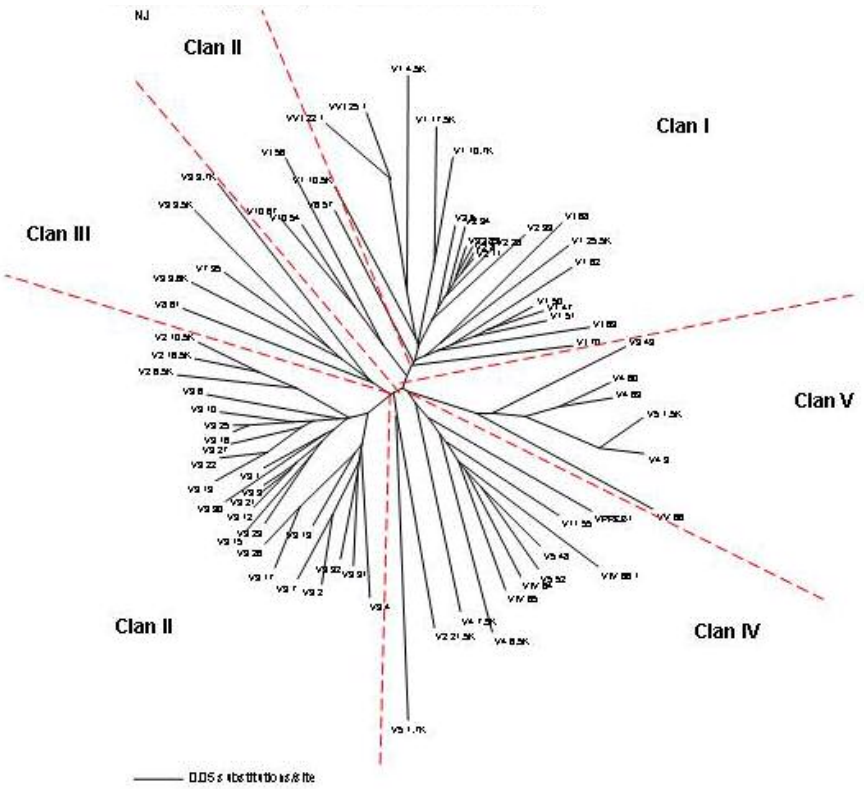

Figure 3.7 Phylogenetic tree for chimpanzee IGLV genes. The IGLV genes were grouped from Clan I through Clan V. Length of branches implicate distance as substitution per site. The scale shows 0.05 substitution per site.

3.1.10 IGLV gene divergence

The percent divergence between human and chimpanzee IGLV genes was calculated using Mega3.1 (Kumar et al. 2004). Here, each IGLV gene pair first was aligned using ClustalX before the p-distance and standard error of distance estimates were calculated using the Bootstrap method (Kumar et al. 2004) with 1000 replicates. These results are shown in Figure 3.8 (a –d).

As shown in Figure 3.8, the percent divergence for IGLV functional genes are slightly higher than both the ORFs and the pseudogenes, with an average percent divergence of 1.99% and 2.23% for exons and introns, respectively. This divergence is

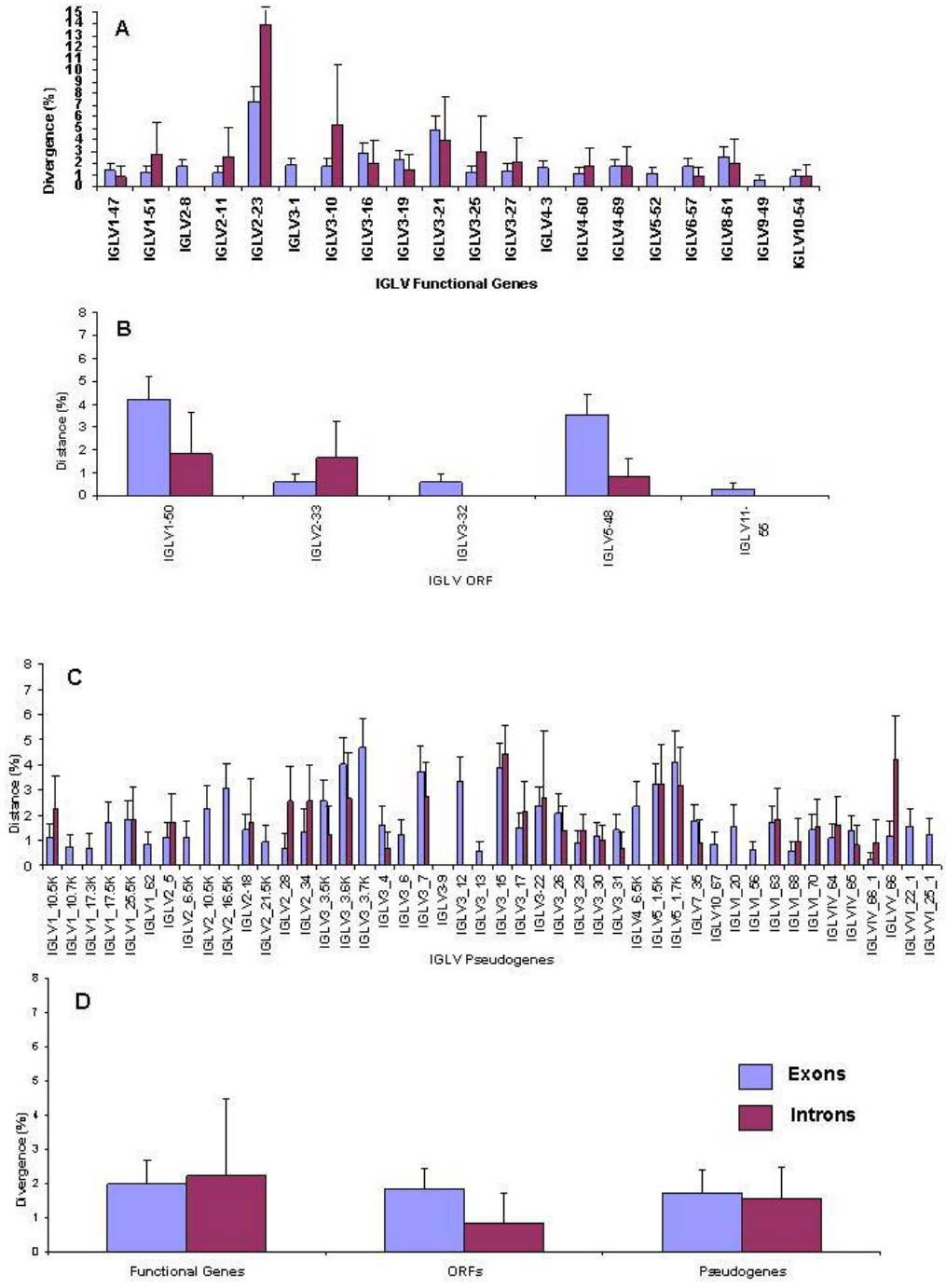

Figure 3.8 The percent divergence between human and chimpanzee IGLV genes. (A) Exon and intron divergence of human and chimpanzee IGLV functional genes (B) Exon and intron divergence of human and chimpanzee IGLV ORFs (C) Exon and intron divergence of human and chimpanzee IGLV pseudogenes. (D) Average divergence of human and chimpanzee IGLV functional, ORFs and pseudogenes. Error bars depict standard errors for uncorrected percent divergence calculated from 1000 bootstrap replicates using Mega3.1 software (Kumar et al. 2004).

unusually high because of the divergence seen in the IGLV2-23 gene. However, when IGLV2-23 is excluded from the calculation of the average, average divergence for the IGLV functional genes is 1.71% and 1.62% for their exons and introns, respectively, an observation similar to that observed for other ORF exons and the pseudogenes.

When compared to divergence rate of non-immunoglobulin known coding genes of 0.83% and 1.24% for exons and introns respectively, divergence rate more than doubled. Interestingly the divergence between the functional IGLV genes and pseudogenes were quite similar. Therefore, although these IGLV genes face selective pressure to maintain functional domains, they also face functional pressure to diversify, and do so at a divergence rate similar to the pseudogenes.

3.1.11 Large-scale differences between human and chimpanzee

The region of the human and chimpanzee genome sequenced were divided into four regions, I through IV, to simplify their comparative analysis.

3.1.11.1 Region I

Region I is immediately distal from LCR22-4 and proximal from the 1 Mb immunoglobulin lambda locus. Human sequence in Region I is approximately 638 Kb and the syntenic chimpanzee sequence is approximately 656 Kb. Overall, this region is highly conserved between human and chimpanzee but insertions in the chimpanzee lineage were observed. A 50 Kb inverted repeat present in both human and chimpanzee was flanked by two identical but inverted processed pseudogenes that are similar to human peripheral benzodiazepine receptor interacting protein. In chimpanzee, proximal to this repeat region an additional 10 Kb insertion that contain only repeated sequences.

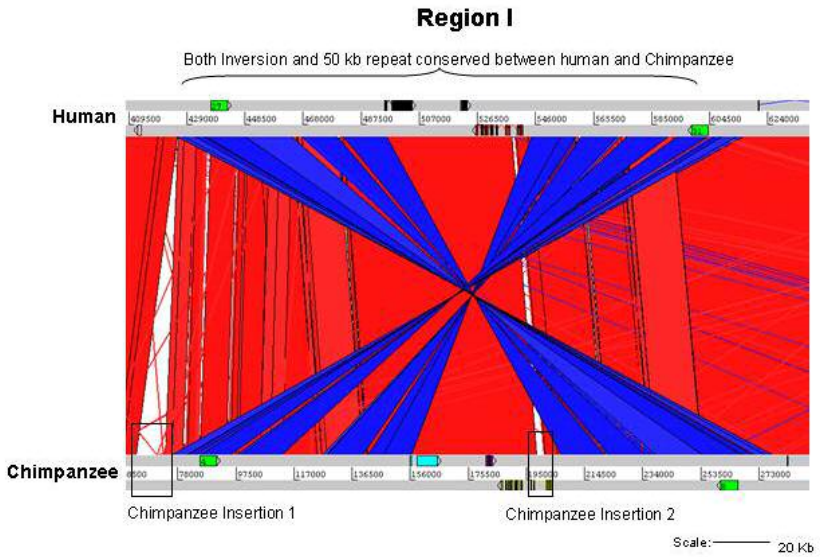

Figure 3.9 An ACT plot showing 2 chimpanzee insertions in Region I, and an approximately 50 Kb inverted repeat that was conserved between human and chimpanzee. Two chimpanzee insertions are shown.

3.1.11.2 Region II

Region II consists of the one megabase IGLL region and there are 5 IGLV gene clusters, IGLV region I – V. The approximately 35 Kb LCR22-5 was found to be embedded within this region between IGLV region I and II. A dot matrix and PIP comparison of the human and chimpanzee IGLL locus reveals 4 major human insertions and 3 major chimpanzee insertions in the IGLL locus as shown in Figure 3.9. There are four major human insertions in the IGLL locus when compared to the chimpanzee sequence. As a result of these human insertions, 6 functional IGLL genes, 12 IGLL pseudogenes, 4 non-immunoglobulin pseudogenes, and 2 non-immunoglobulin partially duplicated genes were specific to the human lineage.

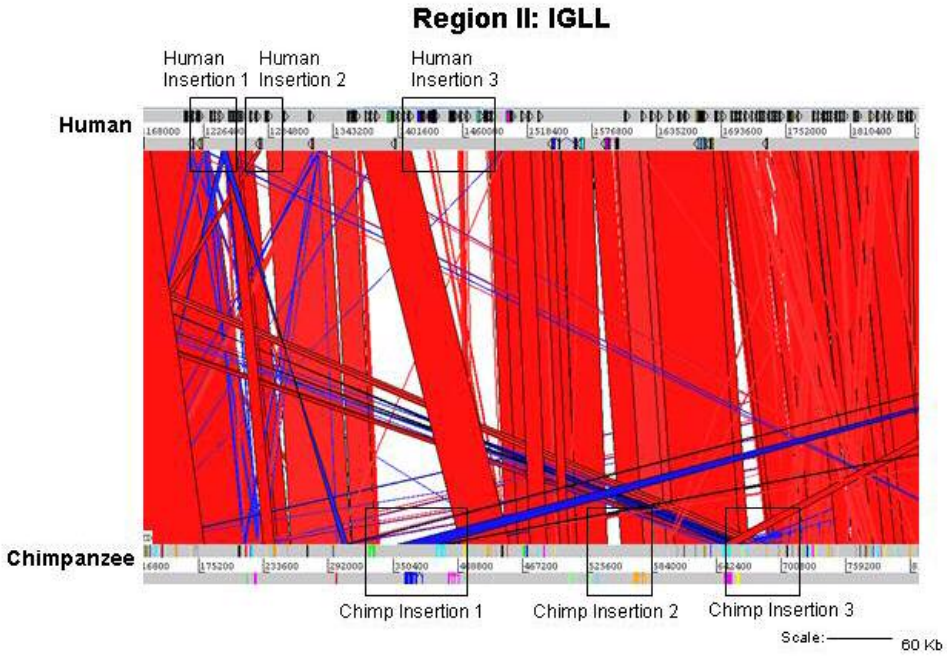

Figure 3.10 An ACT plot showing 3 human insertions and 3 chimpanzee insertions in the IGLL locus.

The first insertion in the human sequence is approximately 15 kb and located between approximately 231,470 bp and 246,383 bp in the human IGLL locus. Two immunoglobulin pseudogenes IGLVIV-59 and IGLVV-58, as well as a non-immunoglobulin pseudogene D87000.2, similar to the Tr:P220044 human bone morphogenetic protein 6 precursor are absent from the chimpanzee sequence.

The second insertion in the human IGLL locus is approximately 13 kb and located between 270,696 bp and 283,059 bp in the human IGLL locus. This additional DNA contains one immunoglobulin gene, IGLVIV-53, one non-immunoglobulin gene, partially duplicated topoisomerase III beta 2 gene (Top3B2) and one non-immunoglobulin pseudogene D88270.2 similar to human RPB5 mediating protein. Top3B2 appeared to be a partial duplication from the functional topoisomerase III beta (Top3B) gene approximately 250 Kb proximal from this human insertion.

The third insertion in the human IGLL locus, approximately 75 kb, and located between 411,000 bp and 486,000 bp in the IGLL locus previously was reported (Robledo et al. 2004). Two identical non-immunoglobulin processed pseudogenes ASH2L1 (absent small or homeotic Drosophila homolog) flank this region in the human sequence with another identical ASH2L1 embedded within the region. In the chimpanzee locus, only the distal ASH2L1 is present. A detailed study of the border of this human insertion revealed a 4-6 Kb highly identical Line1 and LTR repeat sequences. This 75 Kb human insertion encompasses 14 immunoglobulin genes and pseudogenes: IGLV7-46, IGLV1-16.5K, IGLV5-45, IGLV1-16.3K, IGLV1-44, IGLV7-43, IGLVI-42, IGLVVII-41-1, IGLV1-41, IGLV1-40, IGLVI-38, IGLV5-37, IGLV1-11.5K, and IGLV1-36 and was reported as one of the chimpanzee specific haplotypes (Robledo et al. 2004), based on the sequence of a BAC from the CHORI-251 and RPCI-43 chimpanzee BAC libraries determined in our laboratory. The RPCI-43 Chimpanzee BACs revealed the same haplotype.

The fourth insertion in the human IGLL locus relative to chimpanzee is approximately 6 kb, and located between 732,541 bp and 738,681 bp in the IGLL locus. This region contain one immunoglobulin gene IGLV3-24 and one predicted gene D86994.11 with no know function.

In contrast to the inserted sequence in the human IGLL locus, there are 3 larger than 10 Kb insertions in the chimpanzee IGLL locus. The first insertion in region II of chimpanzee is approximately 74 kb and located at 338, 357 bp to 411,424 bp within the chimpanzee IGLL locus, and encodes three predicted genes. The first predicted gene spans 2172 bp, coding for a predicted 723 aa protein that has homology

to the ral guanine nucleotide dissociation stimulators (Ral-GDS). The second predicted gene spans 612bp, and codes for a predicted 203 aa protein with homology to gamma-glutamyltransferase (GTT). The third predicted gene spans 480 bp coding for a 159 aa protein of unknown function with no significant matches to the NCBI nr database.

Further inspection of this region reveals an intra-chromosomal duplication that resulted in this chimpanzee insertion as shown in Figure 3.11. Approximately 1300 Kb distal from this human alignment gap, an identical 50 Kb region that contain both Ral-GDS and GTT genes were identified both in the human and chimpanzee sequence. This duplicated region is 700 Kb distal from the IGLL locus, and encodes two IGLL orphans proximal to the duplicated region that are non-functional immunoglobulin genes are as they are outside of the immunoglobulin locus and unable to recombine with the J and C functional domains.

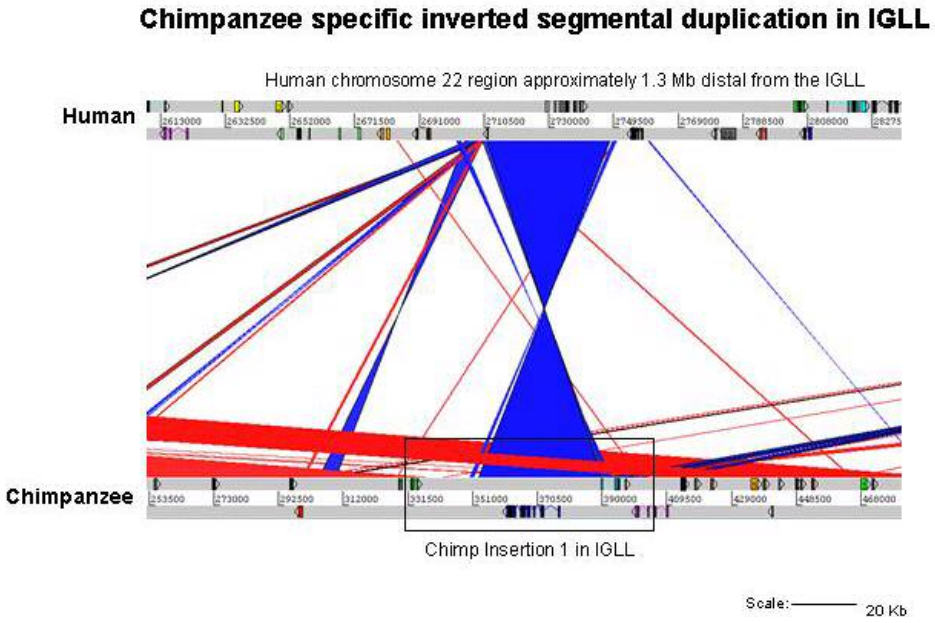

Figure 3.11 An ACT plot showing the inverted duplication from a 1.3 Mb distal region that resulted in chimpanzee insertion 1, a chimpanzee specific inverted segmental duplication in the IGLL locus.

The second chimpanzee insertion located at 544,000 bp to 556,000 bp, is approximately 12 Kb, lacks any detectable coding gene and consists mainly of LTR, Line1, Alu and other repetitive elements.

The third chimpanzee insertion in the IGLL locus approximately 17 Kb occurs between 671,000 bp and 688,000 bp in the chimpanzee IGLL locus and contains only one predicted gene, a gene with homology to the human KIAA0649 gene.

3.1.11.3 Region III

The one-megabase Region III immediately distal to the IGLL locus, corresponds to LCR22-6, one of the human chromosome 22 low copy repeat (LCR22) regions. LCR22-6 spans a region of approximately 180 Kb, and is made up of three repeat modules. The functional BCR gene is located in LCR22-6 and its unprocessed pseudogene copies map to other LCR22s (Collins et al. 2003; Babcock et al. 2003). When compared to the syntenic region in chimpanzee, 2 major human insertions were observed within LCR22-6. The first human insertion is approximately 59 kb, containing an eight-exon partially duplicated gene AP000344.1 that is similar to human gene carboxylesterase, Tr:Q16859. Immediately distal to this human insertion, a pseudogene similar to human ribosomal s10 protein, AP000343.2, occurred but was inverted in chimpanzee compared to human. Approximately 38kb distal from the first human insertion in this region, a 36 kb human insertion that encodes a three-exon, partially duplicated, unknown gene AP000344.7, and a processed pseudogene AP000344.3 that is similar to human gene TXBP181, Tr:Q13312 was observed.

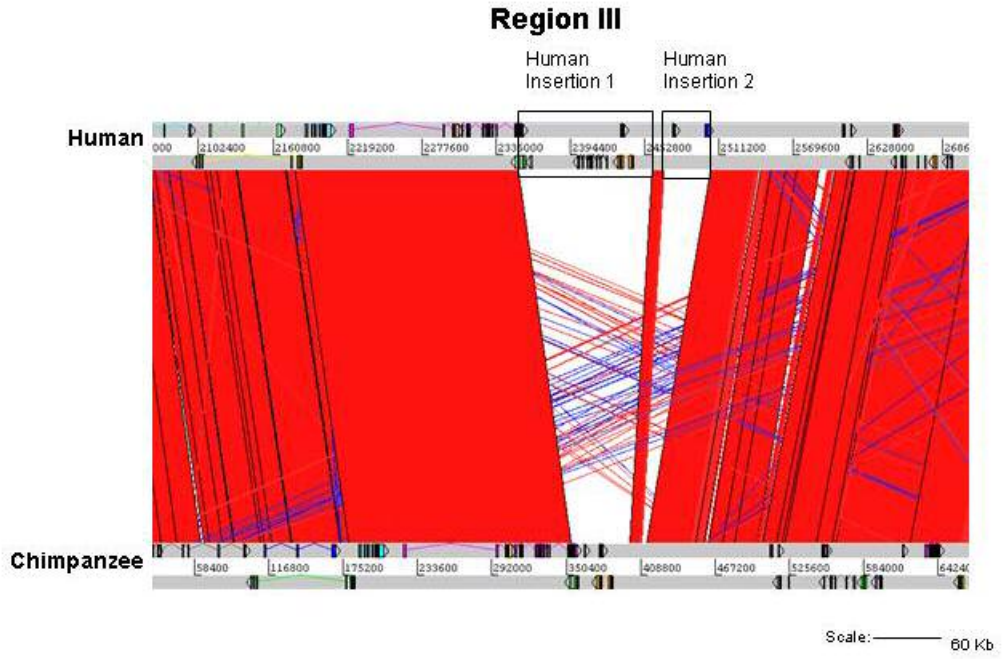

Figure: 3.12 An ACT plot showing human insertions 1 and 2 in Region III.

3.1.11.4 Region IV

The one-megabase Region IV contains LCR22-7 and LCR22-8 that are observed in both chimpanzee and human. LCR22-7 spans approximately 32 Kb and encodes a functional gamma-glutamyltransferase-like activity 1 (GGTLA1) gene, unlike the other LCR22s that often contain truncated, unprocessed, GGT-like pseudogenes (Collins et al. 2003; Babcock et al. 2003).

LCR22-8 spans approximately 90 Kb and among other LCR22 specific duplicated genes or pseudogenes, encodes a functional gamma-glutamyltransferase (GGT1) gene. As shown in Figure 3.13, proximal to LCR22-8 two human insertions relative to the chimpanzee sequence of approximately 6 Kb were observed. Both segments were found to be non-gene sequence filled with repetitive elements. In contrast, the orthologous region of chimpanzee contained a 67 Kb insertion within

LCR22-8, that is flanked by a GTT related pseudogene, AP000356.10, and a BCR-like pseudogene, BCRL6. Of the five genes predicted within the insertion, gene1 is similar to human cDNA FLJ46366, gene2 is similar to predicted hypothetical protein LOC391303, gene3 is similar to human predicted gene KIAA0649, gene4 is similar to membrane glycoprotein POM121 and gene5 is similar to gamma glutamyltransferase (GGT). A BLAST search of this segment against Human chromosome 22 as shown in Figure 3.14 demonstrated that this chimpanzee insertion is an inverted duplication of LCR22-4, a LCR22 that is located approximately 3.5 Mb distal on chromosome 22.

The chimpanzee syntenic region contains an approximately 12 Kb insertion devoid of any predicted genes immediately distal to LCR22-8, that is flanked by two pseudogenes AP000358.2 and AP000358.1.

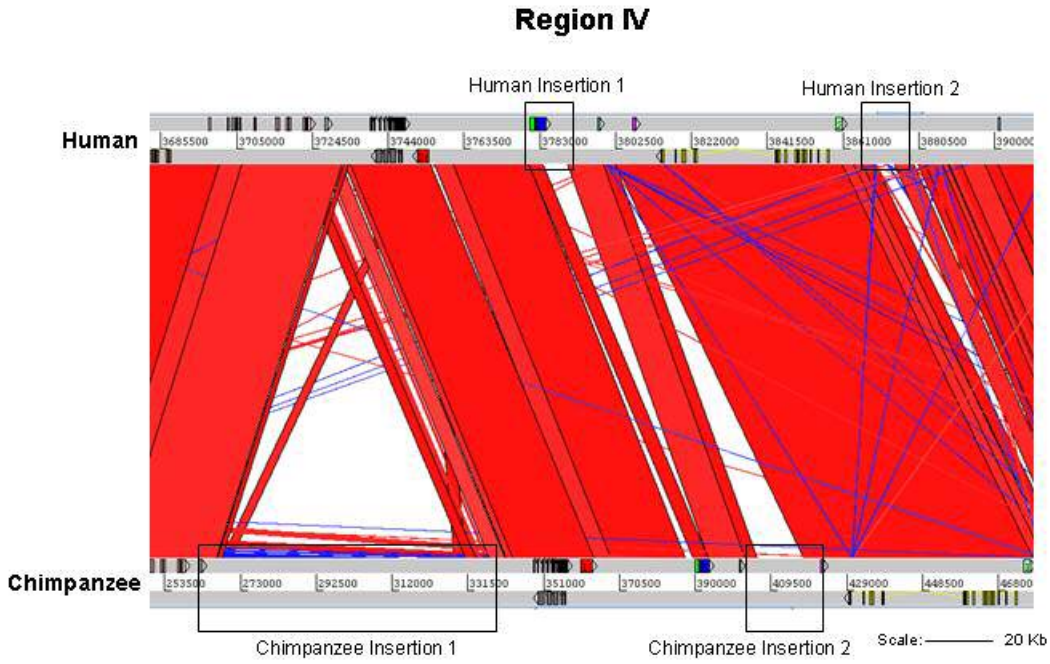

Figure 3.13 An ACT plot showing two major human insertions and two major chimpanzee insertions in Region IV.

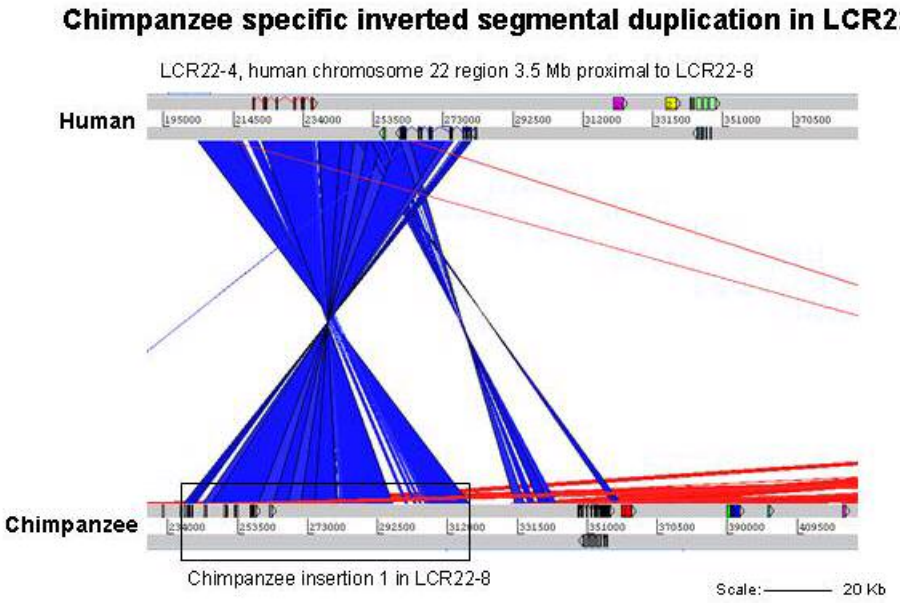

Figure 3.14 An ACT plot showing the inverted duplication that resulted in chimpanzee insertion 1 in LCR22-8.

3.1.11.5 Major differences in IGLL and LCR22s

IGLL is an approximately 1 Mb region consists mostly of tandemly duplicated regions containing IGLV genes. There are 5 major duplicated sub-regions, I, II, III, IV and V, which contain tandemly duplicated units of multiple IGLV genes or pseudogenes in the IGLL region. These sub-regions are interrupted by regions of non immunoglobulin genes and one chromosome 22 segmental duplication, LCR22-5 (Kawasaki et al. 2000). LCR22s are segmental duplications with >95% sequence identity that clusters within different chromosome 22 regions (Dunham et al. 1999; Bailey et al. 2002; Babcock et al. 2003). Four of the total eight LCR22s, LCR22-5, LCR22-6, LCR22-7, and LCR22-8, are located in this chromosome 22 region under study.

Comparison of human and chimpanzee sequence had revealed major

insertion and deletion events occurring in the the IGLL and LCR22s since human and chimpanzee shared a common ancestor. In the IGLL region, four major human insertions were discovered, and they are 15 Kb, 13 Kb, 75 Kb and 6 Kb respectively (Figure 3.11). Three major chimpanzee insertions were also observed in the IGLL region, and they are 74 Kb, 12 Kb and 17 Kb respectively (Figure 3.12). The 74 Kb chimpanzee insertion is an intrachromosomal inverted duplication from a distal region on chromosome 22. Major insertion and deletion events also occur in the LCR22s. Two human insertions with the size of 59 Kb and 36 Kb, respectively, was found in LCR22-6 (Figure 3.10) and a 67 Kb duplication from LCR22-4 was inserted in chimpanzee LCR22-8 (Figure3.13; Figure 3.14). As a result of these insertions and duplications, 6 functional IGLV genes, 12 IGLV pseudogenes, 4 partially duplicated genes and 6 pseudogenes specific to human without a 1:1 chimpanzee ortholog, in addition to 9 predicted chimpanzee genes without a 1:1 human ortholog were observed

Comparison of genes between human and chimpanzee in the IGLL region reveals that IGLV gene segments have a higher divergence rate when compared to protein coding genes. Comparison of the different gene classes reveals that the putative coding genes, partially duplicated genes and pseudogenes in the LCR22s have a much higher divergence rate and are evolving rapidly by changing exon numbers through small scale indels and exon shuffling, in addition to the rapid accumulation of amino-acid-changing base substitution through positive selection with Ka/Ks value >1. An increase of amino acid changes from hydrophobic to hydrophyllic was also observed.

Apart from their roles in diseases, duplications have long been considered as a major pathway for gene evolution (Ohno 1970), in which new gene functions are

believed to have emerged by adaptive evolution following gene duplication. To date, the evidence points to the creation and expansion of genes and gene families through segmental duplications during primate evolution, for example the Kruppel-associated zinc-finger genes (Eichler et al. 1998) on human chromosome 19 and the newly characterized morpheus gene family on human choromosome 16 (Johnson et al. 2001). It is clear from the present study that the major differences between human and chimpanzee lies in the highly repetitive regions of IGLL and LCR22s. While it is well established that immunoglobulin diversify by expanding the numner of IGLV genes through duplication, less is known about the role of LCR22s. The results from this study suggest that these duplicated regions may be unique evolutionary avenues for the creation of new genes. Since segmental duplications are primate specific, they might be the driving force for newly evolving genes underlying phenotypic differences between humans and other primates.

In addition, transcription of partially duplicated genes and peudogenes unique to LCR22s have been detected (Bailey et al. 2002). Though the roles and functions of these transcripts have not been determined, recent studies have shown examples of pseudogenes playing roles in transcriptional regulation, like the pseudo-NOS gene and the makorin1-p1 gene (Korneev et al. 1999; Hirotsune et al. 2003; Harrison et al. 2005). If this is true for the LCR22s transcripts, the loss and gain of the partially duplicated genes and pseudogenes specific to humans and chimpanzees also might play a significant role in the phenotypic differences between the species.

3.1.12 Chimpanzee gene polymorphism

To identify intra-species gene divergence in chimpanzees, and compare the rate to humans, chimpanzee orthologs from 3 different individual chimpanzees as represented in BAC libraries RPCI-43, PTB1 and CHORI-251 were compared. To accomplish this, chimpanzee genes and pseudogenes first were identified using the previously described gene prediction and homology approach on the corresponding sequence in the three different chimpanzee homologous sequences then were compared to the gene set from the CHORI-251 library, obtained from NCBI1.1 WGS assembly. Subsequently, both gene sets also were compared separately to human.

Both exon and intron regions of the orthologous genes were aligned using ClustalW and the percent divergence was calculated as p-distance, which is the proportion (p) of nucleotide sites at which the two sequences being compared are different, using the program Mega3.1 (Kumar et al. 2004). Standard error of distance estimates were calculated using the Bootstrap method with 1000 replicates.

This analysis shows that the divergence rate between two chimpanzee orthologs are overwhelmingly lower than divergence rate of either one compared to human. Comparison between chimpanzee and human orthologs also shows a trend of genes from the RPCI-43/PTB1 libraries having a lower divergence rate than CHORI-251 when compared to human. The only observed anomaly was in introns of IGLV functional genes where the RPCI-43/PTB1 genes have higher divergence rate compared to CHORI-251 genes.

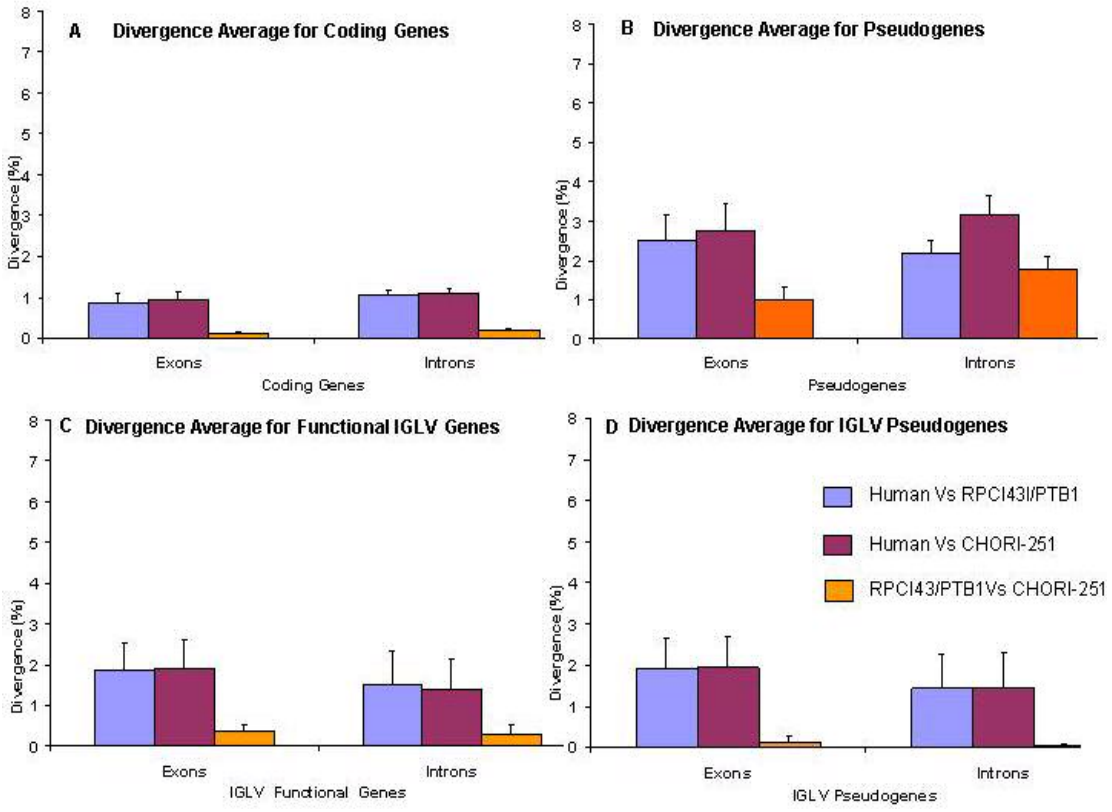

Figure 3.15 Graph showing percent divergence average for comparison between human and chimpanzee orthologous coding genes. Devergence average for (A) protein coding genes, (B) pseudogenes, (C) functional IGLV genes, and (D) IGLV pseudogenes were calculated between human and two chimpanzee orthologs as designated by the blue and purple color. Percent divergence was calculated as p-distance using Mega3.1. Error bars depict standard errors for uncorrected percent divergence.

For non-immunoglobulin genes, average percent divergence for the exons of coding genes for RPCI-43/PTB1 and CHORI-251 compared to human is 0.85% and 0.92%, respectively. However, the percent divergence between the the 2 chimpanzee's exons is only 0.11%. The divergence for introns of coding genes is higher at 1.04%, 1.09% and 0.20% respectively, and the relative divergence rate is similar to that of the exons. Pseudogenes have a higher divergence rate, similar to that observed in the gene coding regions. In comparison, the average percent divergence of exons between the two chimpanzees for coding genes and pseudogenes are 0.11% and 1.00%, respectively,

and the average percent divergence of introns are 0.20% and 1.78%, respectively.

For the IGLV genes, divergence for exons are generally higher than introns for both chimpanzee genes relative to human as within different chimpanzees. The average percent divergence of exons between the two chimpanzees for IGLV functional genes, ORFs and pseudogenes are 0.36%, 0.09% and 0.13%, respectively, and the average percent divergence of introns are 0.29%, 0%, and 0.03%, respectively.

3.1.13 Comparison of BAC and WGS Assembly

To compare the BAC by BAC sequencing approach to the whole genome short-gun (WGS) sequencing approach, chimpanzee sequence assembled from chimpanzee BACs was compared to corresponding region in chimpanzee NCBI1.1 WGS assembly. Both chimpanzee sequences were compared to each other and separately to corresponding human sequence using dot matrix anlysis.

One of the major differences observed between the BAC assembly and the WGS assembly involves a 50 Kb inverted repeat at the beginning of the sequence. The BAC assembly has both copies of the repeats, and comparison of the chimpanzee BAC assembly to human shows both the repeats, along with an approximately 50Kb unique region in between them conserved between human and chimpanzee. However, the WGS assembly only has one copy of the repeats.

Many sequence gaps in the WGS assembly also were detected when compared to both the human sequence and the BAC assembly, especially where there were known repeats, such as the IGLL region, LCR22-5, and LCR22-8. Thus, although useful in locating unique genomic features, it is clear that the WGS approach is unable

to produce a correct sequence assembly when dealing with the high number of repetitive sequences found in most genomes.

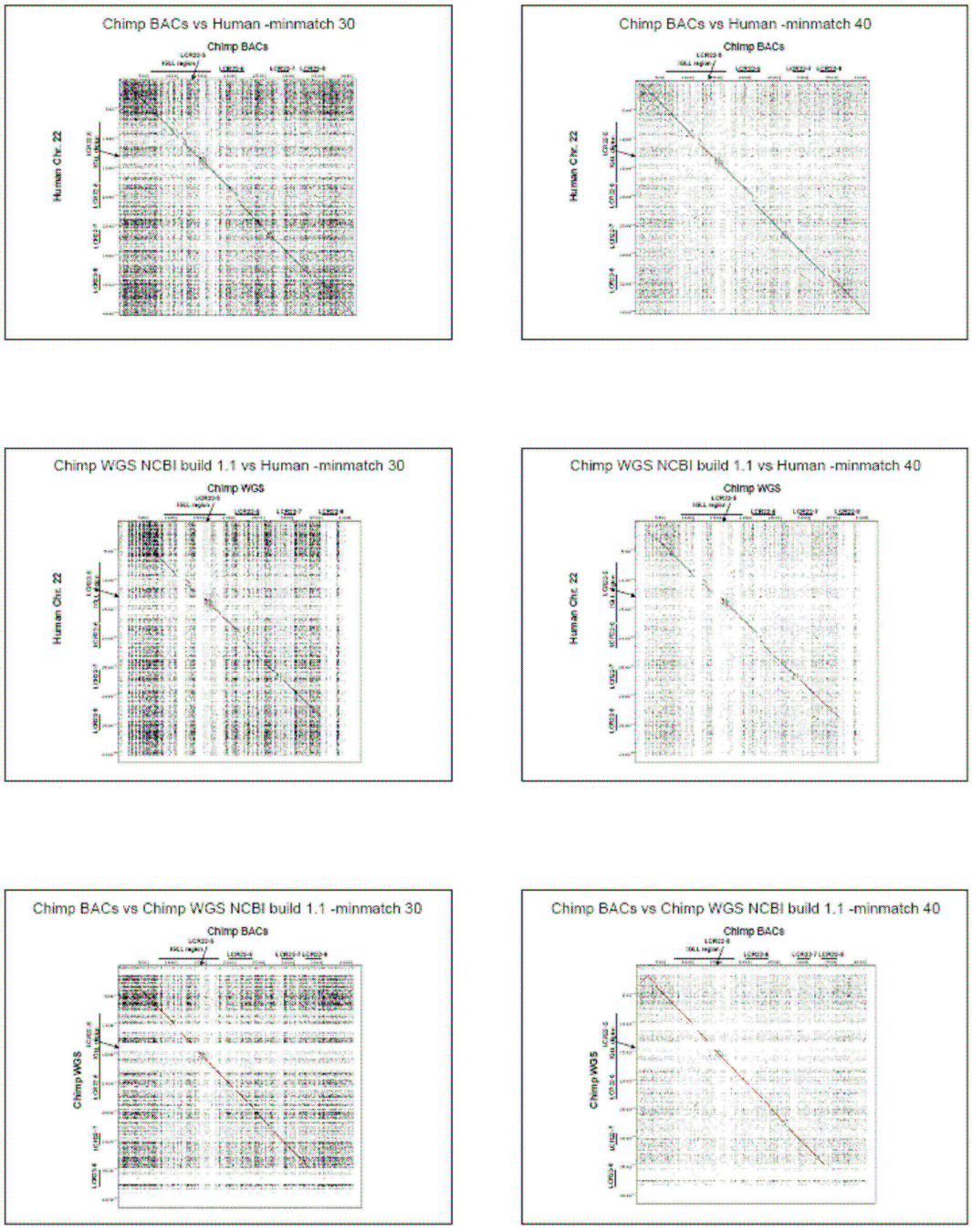

Figure 3.16 Dot matrix analysis using program Maxmatch. Chimpanzee BAC assembly was compared against human corresponding sequence. The chimpanzee NCBI WGS assembly was also compared to human. The two chimpanzee sequences were compared against one another. Minmatch value of both 30 and 40 were included in the figures above.

3.2 Multispecies comparison

Previous understanding of vertebrate genome organization and evolution was derived mainly from cytogenetic banding and painting studies, as well as gene order mapping comparison (Yunis and Prakash 1982; Nadeau and Sankoff 1998; O'brien et al. 1999; Murphy et al. 2001). These results led to the postulation that mere rearragements of genome structure could account for the differences in mammalian genomes and the estimation that the human and dog genome differ by 17 rearranged synthetic blocks (Wienber et al. 1997), while the human and mouse genome differ by 180 rearranged synthenic blocks (O'Brien et al. 1997). However, this view had been challenged as a result of the identification of highly homologous segemental duplications (IHGSC 2001) such as that of LCR22s on human chromosme 22 as well as comparative analysis of the sequences for fugu, mouse and rat genomes (Aparicio et al. 2002; MGSC 2002; RGSPC 2004), and recent assemblies of the dog, cow, chicken, and zebrafish genomes thatare now available

The targeted 4 Mb human chromosome 22 region was compared to syntenic regions in dog, cow, mouse, rat, chicken, frog zebrafish, fugu and tetraodon. The synteny of this region was conserved in dog chromosome 26, cow chromosome 17, and chicken chromosome 15, while mouse and rat shared two chromosomal breakpoints in this region, as regions of mouse chromosomes 16, 10 and 5, and regions of rat chromosomes 11, 20 and 12 are syntenic to the human chromosome 22 region. While the synteny decreased drastically when compared to the zebrafish genome, three syntenic blocks were observed, one on chromosome 5 and a duplicated syntenic block on both chromosomes 8 and 21 as shown in Figure 3.17. The observations are

consistant with previous assertion that a genome wide duplication occurred in the ray-fin fish lineage (Wittbrodt et al. 1988; Amores et al. 1998; Postlewait et al. 2000; Aparicio et al. 2002; Taylor et al. 2003).

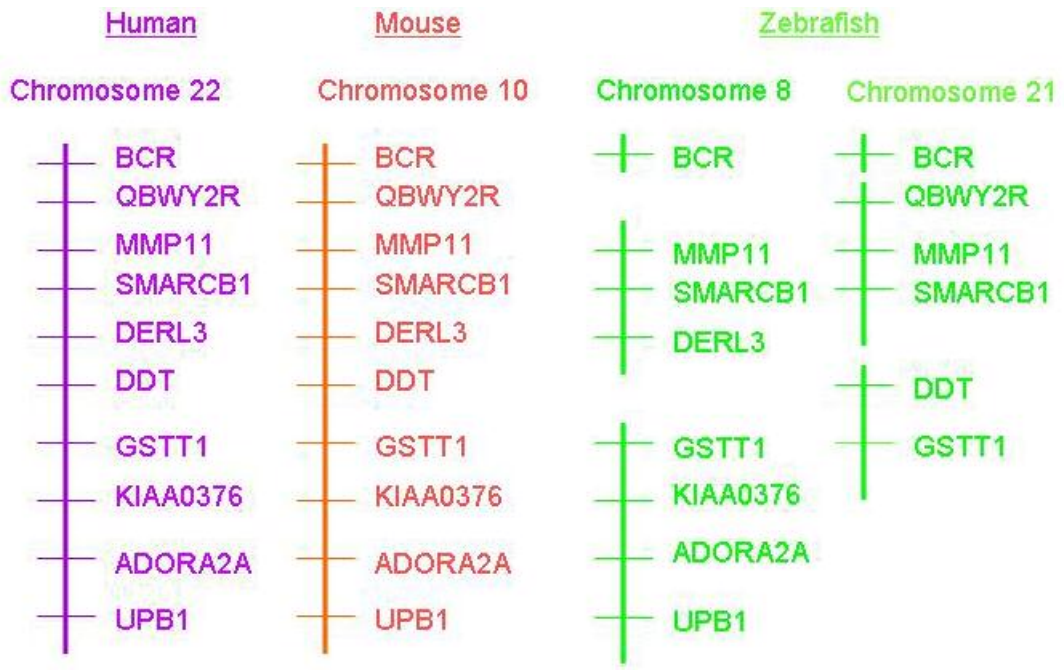

Figure 3.17 Diagram showing synteny in human, mouse and zebrafish. While human and mouse have one copy of each gene, the synteny blocks in zebrafish were duplicated and four mammalian orthologs were duplicated in zebrafish.

As a result, four genes in this human chromosome 22 orthologous region, BCR, MMP11, SMARCB1 and GSTT1, were observed as duplicated copies in zebrafish. Comparison of the diverse taxa also demonstrated that while many unique protein coding genes were conserved in all vertebrates, the LCR22s as well as several putative coding genes, partially duplicated genes and pseudogenes such as AP000354.1, AP000354.7, AP000354.4, AP000356.10, BCRL4, BCRL6, AP000356.9 AP000357.3, AP000357.2 AP000358.2 AP000358.1 were primate specific.

B Gene Expression Profiling in zebrafish

3.3 Development of strategy

3.3.1 Overview

Even before the human genome sequence was completed, elucidating the function of newly identified genes with unknown functions became one of the ensuing challenges for the genomics community. Since our laboratory sequenced a significant portion of human chromosome 22, our initial focus has been to investigate the expression of genes of unknown function predicted on this chromosome using the zebrafish as an animal model for whole mount *in situ* hybridization based (WMISH) gene expression profiling. Locating and timing human orthogous gene expression in zebrafish development is an initial step towards determining the function of these unknown genes. To this end, we have developed a robust, 96-well format, high throughput protocol for large scale screening of zebrafish gene expression during different embryonic developmental stages.

3.3.2 Pilot study with RNA Probes

Zebrafish orthologs of human chromosome 22 genes were obtained initially by a BLASTn search of a zebrafish cDNA library constructed by Dr. Han Wang at the Department of Zoology and sequenced by Dr. Yuhong Tang in our laboratory.

Three genes with the most significant EST matches initially were selected for a whole mount zebrafish *in situ* hybridization pilot study. These EST clones were b6n20zf, representing the mRNA for the zebrafish gene ENSDARG00000012849.2, an ortholog for human gene AP000557.1; a8h24zf, representing the mRNA for zebrafish

gene ENSDARG00000006719.2, an orthologo of human gene AP000354.2; and a4g17zf, representing a Josephin domain 1 containing gene similar to human predicted gene of unknown function and expression profile, dJ508I15.5. After these cDNA clones were isolated and digested with either Xho1 or EcoR1, the linearized plasmid was used as the template in an *in vitro* transcription reaction, where T3 and T7 RNA polymerase were used produce a digoxigenin labeled dUTP containing RNA probe. These probes then were used in the whole mount *in situ* hybridization study shown in Figure 3.18. Probe a4g17 was highly expressed in the head of developing zebrafish embryos, and based on this intinial study, probe a4g17 was used to further characterize the system.

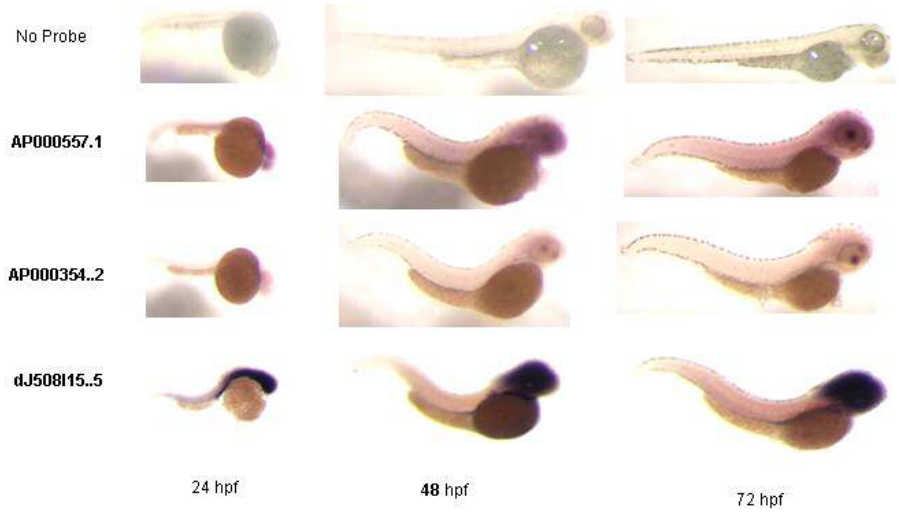

Figure 3.18 Expression pattern of the three initial RNA probes used. Probes were generated by *in vitro* transcription using cDNA clones as template.

3.3.3 Probes variable length study

To test the efficiency of probes in the zebrafish whole mount *in situ* hybridization vs probe length, PCR products of variable length containing the T7 promoter were generated by amplification of the clone a4g17. These PCR products then were used as templates for *in vitro* transcription to generate variable length RNA probes as shown in figure 3.19.

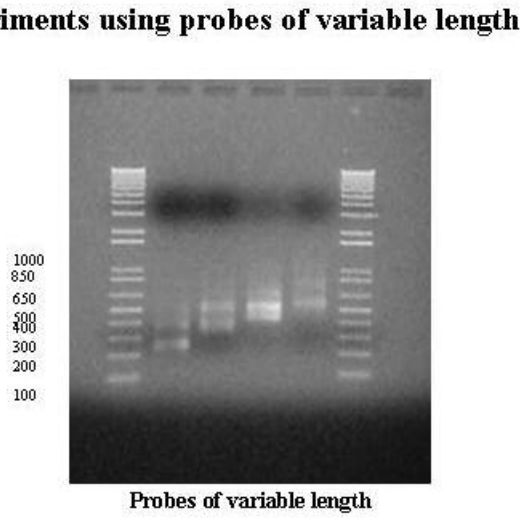

Figure 3.19 PCR products of variable length generate from cDNA clone a4g17 that contains the T7 promoter sequence was used as template for *in vitro* transcription. The RNA probes, shown here in the gel, are then used in whole mount *in situ* hybridization of zebrafish embryos.

These RNA probes of variable lengths then were used in zebrafish whole mount *in situ* hybridization experiments at 48 hpf and 72 hpf. As shown in Figure 3.20, the shortest probes at the length of 100 bp took at least 8 hours to show clear staining in both the stages. As the length of the probes increase, time needed for clear staining decreased. For probes at the length of 1000 bp, only 2 hours were needed to achieve the clear staining pattern. However, the longer probes have the potential to cause background staining, as shown in the trunk region of the embryos subjected to 1000 bp

probe hybridization. The results of this experiment provided us with an estimate of the staining time required for probes of variable lengths for future experiments.

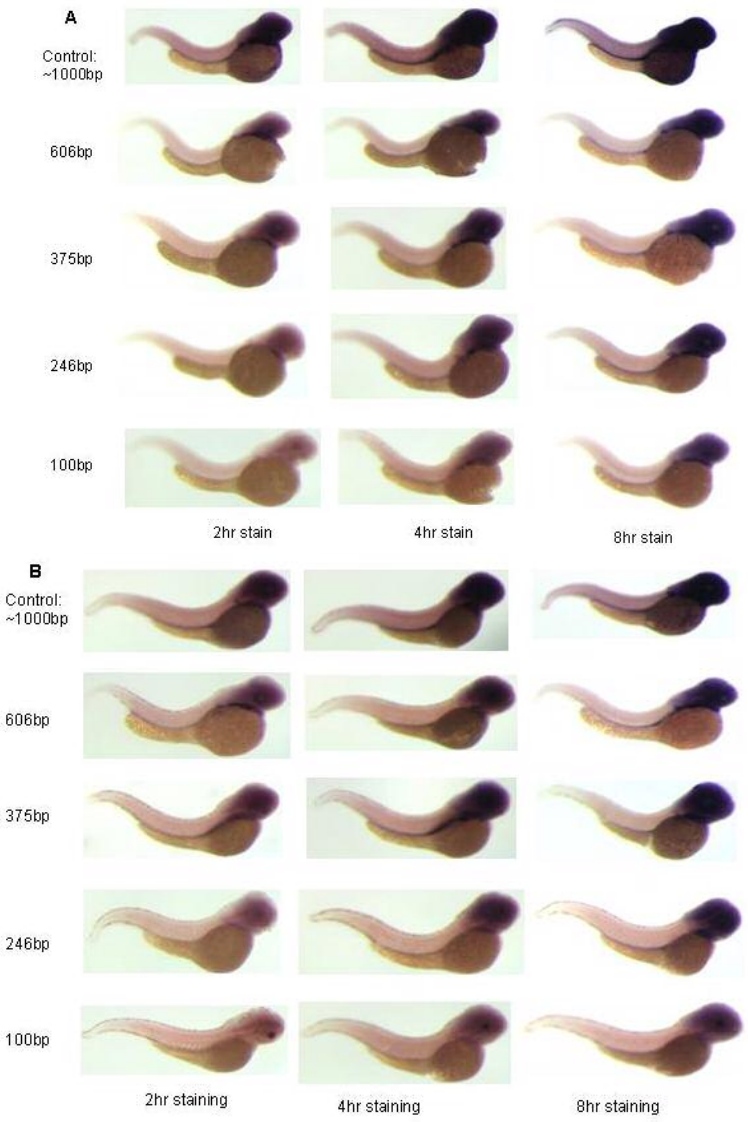

Figure 3.20 Experiments for probes of variable length. (A) 48 hpf and (B) 72 hpf embryos with their staining pattern using antisense probe for a4g17. Five different length of probes are used and staining time range from 2 hours to 8 hours. Intensity of staining pattern of probes is in correlation with their length. The shorter the probes, the longer to achieve staining intensity.

3.3.4 Scaling to 96 wells format

To improve its efficiency and throughput we adapted the whole mount *in situ* hybridization protocol zfin (zfin.org), optimized for 1 ml volume for hybridization and wash solutions in 1.5 ml Eppendorf tubes to a 96-wells microtiter plate format. This resulted in decreasing the hybridization and wash volumes to 200 µl to fit into the 96-wells microtiter plate and a concommitent increase in the number of wash steps to adjust for the decreased volume and gave the results shown in Figure 3.21.

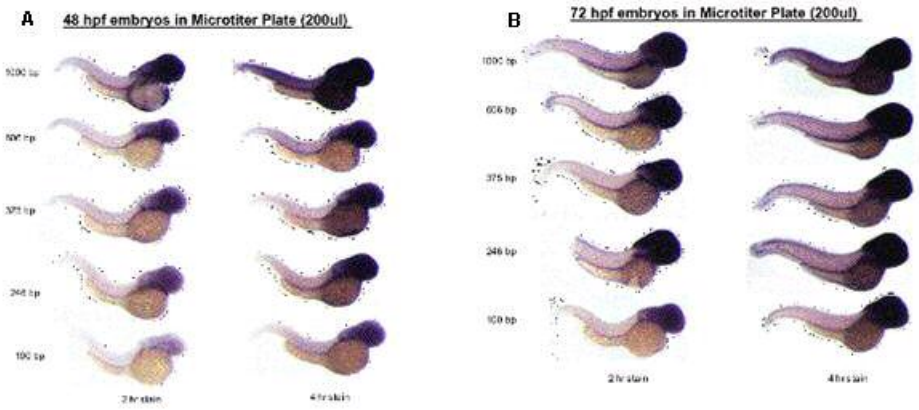

Figure 3.21 Hybridyzation in the 96-wells microtiter plate format. (A) 48 hpf and (B) 72 hpf embryos with their staining pattern using antisense probe for a4g17 using the 200 µl volume 96-wells microtiter plate format.

3.3.5 DNA Probes

In addition to scaling the *in situ* hybridization to a 96-wells microtiter plate format, a method to generate sufficient highly specific probes for all zebrafish orthologs of human chromosome 22 genes was required. Using the zebrafish genome sequences provided by the Sanger Center, we generated DNA probes by exon-specific PCR

amplification using zebrafish genomic DNA as template. These PCR products then were used as templates for single primer amplification to produce DNA probes that contained incorporated digoxigenin labeled dUTP.

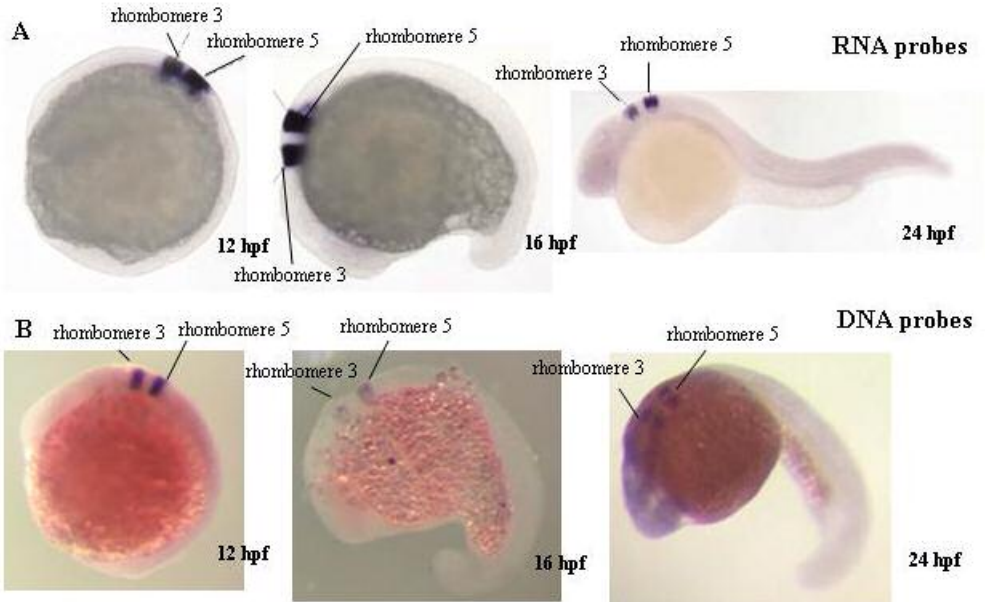

Figure 3.22 Comparison of (A) RNA probes and (B) DNA probes for Krox-20 that is expressed in rhombomere 3 and rhombomere 5 of 12 hpf, 16 hpf and 24 hpf zebrafish embryos.

As demonstrated in Figure 3.22, comparison of RNA vs. DNA probes was carried out using the Krox-20 gene that is expressed in rhombomere 3 and rhombomere 5 of 12 hpf, 16 hpf and 24 hpf zebrafish embryos. Under similar *in situ* hybridization conditions, RNA probes shows increased intensity of staining pattern. However, DNA probes demonstrated specific Krox-20 that is sufficient for differentiation from background staining. As a result of these comparative studies and the availability of zebrafish genomic DNA for efficient generation of DNA probes, the DNA probes became our method of choice for the subsequent expression profiling studies.

3.4 Expression of human orthologs in zebrafish

A total of 31 zebrafish orthologs were identified in the 4 Mbp region of human chromosome 22 immedicately downstream from the DGCR region and extending through the IGLL and BCR reigion by comparison to Zebrafish genome assembly Zv5 and checked for redundancy in the available zebrafish sequence. Only exon sequences that were unique were used in this present study.

Exon specific primers were picked using Primou and PCR products were generated using the zebrafish genomic DNA as template. After verifying the PCR product by sequencing, they then were used as templates for single primer amplification to generate the DIG-dGTP labeled exon specific probes, thatwere used in the 96-well format zebrafish whole mount *in situ* hybridization to determine the embryonic expression patterns.

As shown in Figure 3.23 and Table 3.2, a total of 12 zebrafish orthologs were expressed in specific tissues at specific stages in the zebrafish embryos. Eleven others had no specific expression pattern during developmental phases of zebrafish embryos. Probes of sufficient length were unavailable for 8 of the orthologs either because the exons predicted were short, or because Primou did not pick unique exon-specific primers.

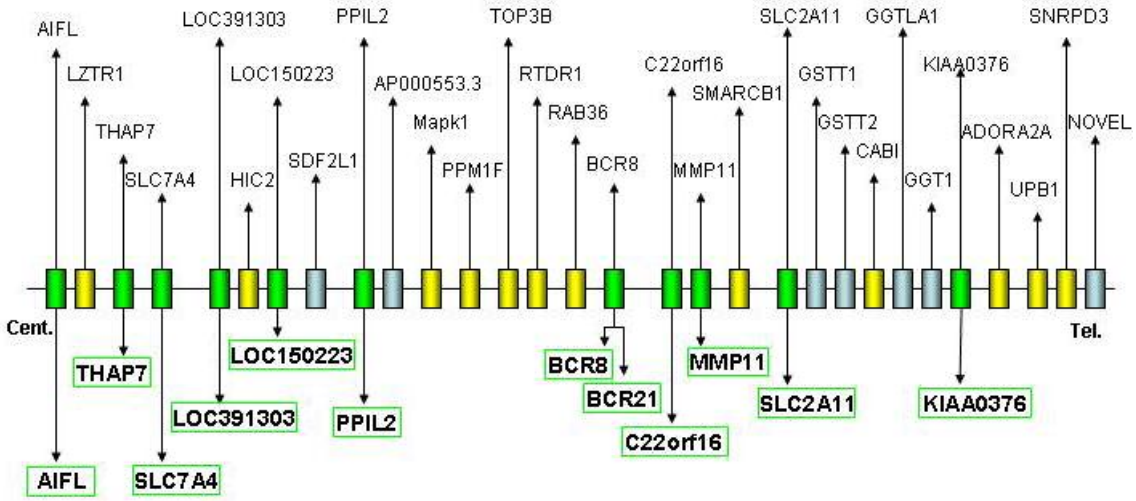

Figure 3.23 A schematic representation of the genes located in the human chromosome 22 of which zebrafish orthologs were identified in this study. Genes in the green boxes represents genes with observed specific expression pattern in developing zebrafish embryos. Yellow boxes represent genes in which expression studies were carried out but staining pattern was unspecific. Blue boxes indicate genes that probes were unavailable due to small intron sizes and/or unique primers could not be picked successfully for the amplification of probes.

Human Chr22 Ensembl Gene ID	Sanger Asscession #	Gene Name	Zebrafish Homolog Ensembl Gene ID	Zebrafish Expression
ENSG00000183773	AC002472.7	AIFL	ENSDARG00000002125	Specific – Fig 4.5
ENSG00000099949	AC002472.2	LZTR1	ENSDARG00000015905	Unspecific
ENSG00000184436	AC002472.8	THAP7	ENSDARG00000027585	Specific – Fig 4.6
ENSG00000099960	AC002472.5	SLC7A4	ENSDARG00000026245	Specific – Fig 4.7
ENSG00000169892	AP000552.4	LOC391303	ENSDARG00000027240	Specific – Fig 4.8
ENSG00000169635	AP000557.1	HIC2	ENSDARG00000038298	Unspecific
ENSG00000161179	AP000553.6	LOC150223	ENSDARG00000002884	Specific – Fig 4.9
ENSG00000128228	AP000553.4	SDF2L1	ENSDARG00000035631	Probe N/A
ENSG00000100023	AP000553.2	PPIL2	ENSDARG00000002016	Specific – Fig 4.10
ENSG00000100027	AP000553.3	YPEL1	ENSDARG00000035630	Probe N/A
ENSG00000100030	AP000555.1	Mapk1	ENSDARG00000027552	Unspecific
ENSG00000100034	D86995.1	PPM1F	ENSDARG00000005786	Unspecific
ENSG00000100038	D87012.1	TOP3B	ENSDARG00000027586	Unspecific
ENSG00000100218	AC000029.2	RTDR1	ENSDARG00000017983	Unspecific
ENSG00000100228	AC000102.1	RAB36	ENSDARG00000014058	Unspecific
ENSG00000186716	U07000.1	BCR	ENSDARG00000042474	Specific – Fig 4.12
ENSG00000186716	U07000.1	BCR	ENSDARG00000028844	Specific – Fig 4.13
ENSG00000138869	AP000348.4	C22orf16	ENSDARG00000010717	Specific – Fig 4.14
ENSG00000099953	AP000349.1	MMP11	ENSDARG00000026293	Specific – Fig 4.15
ENSG00000099956	AP000349.2	SMARCB1	ENSDARG00000011594	Probe N/A
ENSG00000133460	AP000350.2	SLC2A11	ENSDARG00000034501	Specific – Fig 4.16
ENSG00000133433	AP000350.7	GSTT2	ENSDARG00000017388	Probe N/A
ENSG00000184674	AP000351.10	GSTT1	ENSDARG00000042428	Probe N/A
ENSG00000099991	AP000352.1	CABI	ENSDARG00000039230	Unsepecific
ENSG00000099998	AP000354.3	GGTLA1	ENSDARG00000007929	Probe N/A
ENSG00000100031	AP000356.4	GGT1	ENSDARG00000023526	Probe N/A
ENSG00000100014	AP000354.2	KIAA0376	ENSDARG00000006719	Specific – Fig 4.17
ENSG00000128271	AP000355.1	ADORA2A	ENSDARG00000018790	Unspecific
ENSG00000100024	AP000355.2	UPB1	ENSDARG00000011521	Unspecific
ENSG00000100028	AP000356.7	SNRPD3	ENSDARG00000005825	Unspecific
ENSG00000167037	dJ930L11.1	Novel	ENSDARG00000028857	Probe N/A

Table 3.2 A list of chromosome 22 genes in the region studied, their Ensembl gene IDs, Sanger accession number, and gene name if previously characterized.

3.4.1 Apoptosis-inducing factor like (AIFL) gene

The gene for apoptosis-inducing factor like (AIFL), AC002472.7, encode a 598 amino acid putative protein that shares a 35% sequence identity with apoptosis-inducing factor (AIF). AIFL contains both an apoptosis-inducing characteristic Rieske domain and a pyridine nucleotide-disulfide oxidoreductase domain (Pyr_redox) (Xie et al. 2005), implying that this protein may be involved in apoptosis, or programmed cell death, a crucial cellular process required for metazoan embryonic development, essential in organogenesis, to successfully craft complex multicellular tissues, and normal tissue homeostasis (Danial & Korsmeyer 2004; Xie et al. 2005).

Expression of AIFL recently was detected in a wide array of adult human tissues (Xie et al. 2005), but no expression information was available for AIFL expression during human embryonic development. However, since the frog homolog of AIFL, Nfrl, was expressed exclusively in the embryonic ectodermal region that develops into the brain and spinal cord as well as into the nervous tissue of the peripheral nervous system, termed the neuroectoderm, during xenopus embryonic development (Hatada et al. 1997), AIFL expression was investigated in zebrafish embryos.

As shown in Figure 3.24, the expression of the AIFL zebrafish homolog, Ensembl gene ID ENSDARG00000002125, was observed at 24 hpf predominantly in the forbrain, tectum, midbrain and hindbrain. It also was expressed in the somites, the undifferentiated mesodermal component in the early trunk or tail segment that ultimately develop into myotome or sclerotome. By 48 hpf, expression clearly was restricted to the brain, concentrating at the telencephalon, tectum, cerebellum, and the hindbrain and at 72 hpf, expression of the genes was reduced further, concentrating only

at tectum and cerebellum.

Based on this evidence and the similar expression of the human and xenopus AIFL orthologs in human and xenopus brain respectively, it is clear that AIFL is a gene expressed in the central nervous system in both developing embryos and adults.

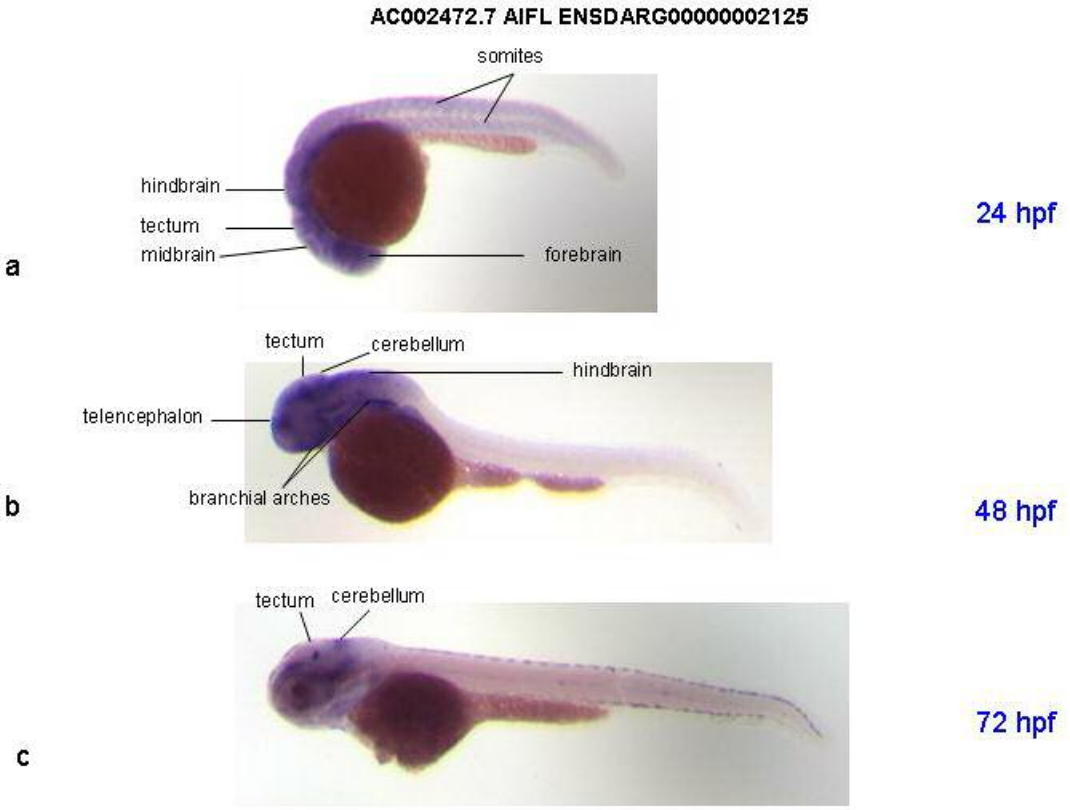

Figure 3.24 Expression pattern for AIFL Ensembl gene ID ENSDARG00000002125 at (a) 24 hpf observed in forebrain, midbrain, tectum, cerebellum, hindbrain and somites; at (b) 48 hpf observed in the telencephalon, tectum, cerebellum, hindbrain and branchial archres; and at (c) 72 hpf in the tectum and cerebellum.

3.4.2 Thanatos-associated protein member 7 (Thap7) gene

Thanatos-associated protein member 7 (Thap7), a 309 aa member of the recently identified transcription repressor thanatos-associated protein family, contains an N-terminal THAP domain followed by a proline-rich region and a C-terminal acidic domain. A total of 12 gene families (Thap0-Thap11) have been identified in the human genome and they all encode proteins that contain the approximately 90 aa Thap domain in their N-terminal (Macfarlan et al. 2005). To date, only 3 of these members have been characterized, they are (i) THAP0 (DAP4/p52rIPK), a protein that was isolated in a genetic screen for genes involved in interferon-gamma-induced apoptosis in HeLa cells (Deiss et al. 1995), and in a screen for regulators of the interferon-induced protein kinase R (PKR) functioning as an inhibitor of PKR (Gale et al. 1998); (ii) THAP1, a protein that cause both serum withdrawal and tumor necrosis factor alpha-induced apoptosis, and interacts with prostate-apoptosis-response-4 (Par-4), a well characterized proapoptotic factor, previously linked to prostate cancer and neurodegenerative diseases (Roussigne et al. 2003); and (iii) Thap7, the first Thap domain containing protein in human demonstrated to regulate transcription and was identified to be a transducer of the repressive signal of hypoacetylated histone H4 in higher eukaryotes, repressing transcription by associating with chromatin, and preferentially binds to H3 and H4 histones (Macfarlan et al. 2005). Genetic evidence in *Caenorhabditis elegans* and *Drosophila* also indicated that THAP domain proteins may have roles in chromatin-based processes, including transcription regulation (Boxem et al. 2002; Reddy et al. 2004).

Expression of Thap7, Ensemble gene ID ENSDARG00000027585, as shown in

Figure 3.25, was observed at 24 hpf in the forebrain, retina, cerebellum and the hindbrain, 48 hpf in the optic tectum and the tegmentum, and 72 hpf in the tegmentum.

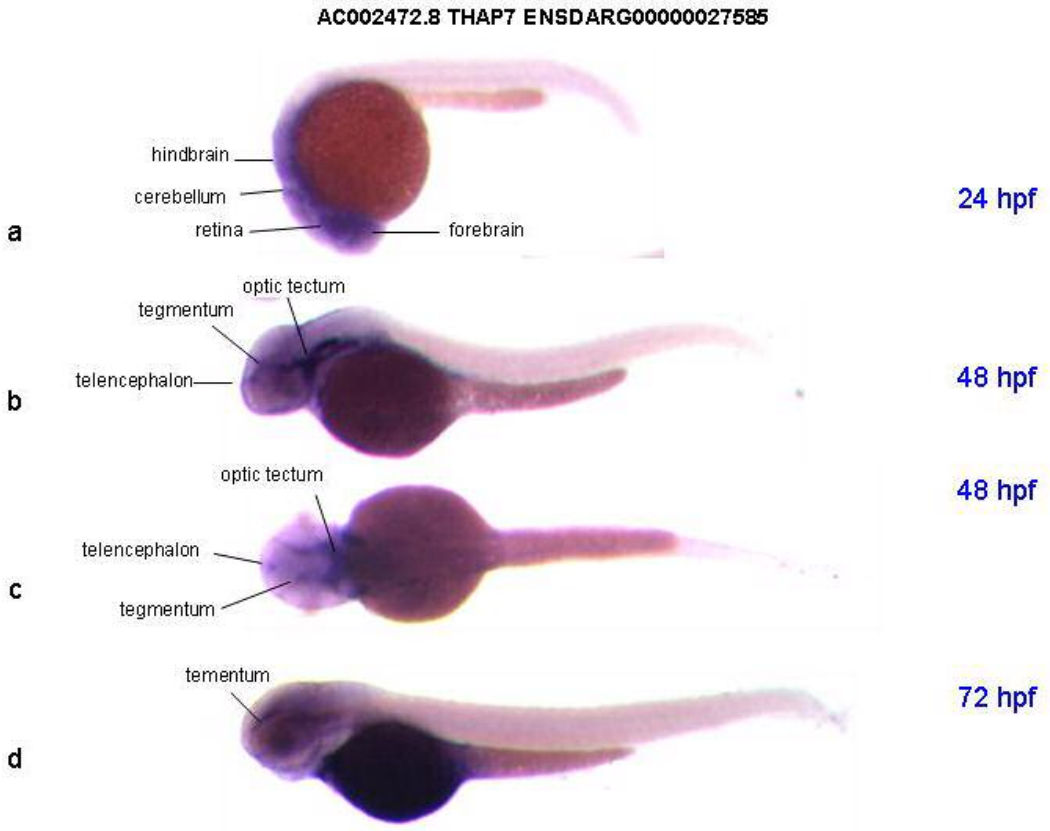

Figure 3.25 Whole mount *in situ* hybridization expression pattern for Thap7 Ensembl gene ID ENSDARG00000027585. Expression was observed at (a) 24 hpf in forebrain, retina, cerebellum, and hindbrain; at (b) 48 hpf in tegmentum, telencephalon, and optic tectum; and at (c) 72 hpf in the tegmentum.

The results from this study indicate that the expression of Thap7 in specific tissues of developing zebrafish embryos is consistant with its physiological role in developing vertebrates, and extended previous studies on its biochemical and cellular functions. Not surprisingly, previous studies had demonstrated that proteins involved in apoptosis (Danial & Korsmeyer 2004; Xie et al. 2005) and the repression of transcription (Hanna Rose & Hansen 1996; Seki et al. 2003) required for successful metazoan embryonic development and organogenesis.

3.4.3 Solute carrier family 7 member 4 (SLC7A4) gene

Solute carrier family 7 (cationic amino acid transporter, y+ system), member 4 (SLC7A4) shares 38.5% amino acid sequence homology with SLC7A1 and 37.8% homology with SLC7A2, two previously known SLC7 subfamily of human cationic amino acid transporters (Sperandeo et al. 1998). However, its function as an amino acid transporter recently has been disputed (Wolf et al. 2002), thus, the function of this gene needs further experiments for validation. Earlier Northern blot analysis detected expression of SLC7A4 in human brain, testis, and placenta (Sperandeo et al. 1998). Our present study, as shown in Figure 3.26 demonstrated the expression of SLC7A4, Ensembl gene ID ENSDARG00000041892, in the forebrain, midbrain, hindbrain and the retina of 24 hpf zebrafish embryos with no expression being observed there at later development stages but with expression observed in the ottic vesicle and pectoral fin at 48 hpf and the branchial areches at 72 hpf. These results implied that SLC7A4 is involved in early developmental processes in zebrafish.

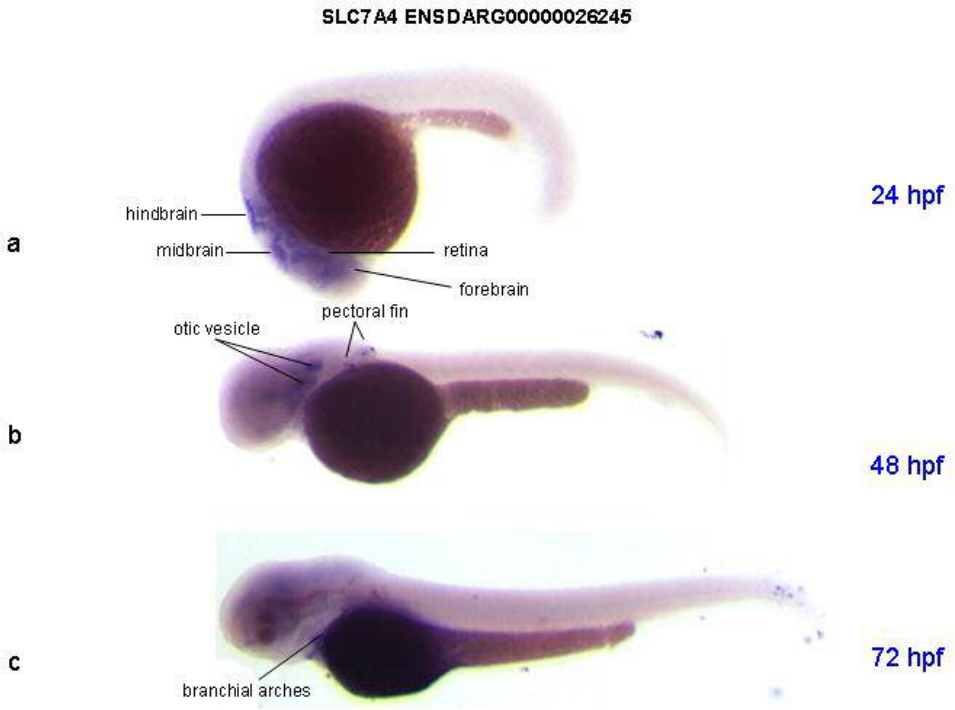

Figure 3.26 Whole mount *in situ* hybridization expression pattern for SLC7A4 Ensembl gene ID ENSDARG00000041892. Expression was seen at (a) 24 hpf in the forebrain, midbrain, and the hindbrain; at (b) 48 hpf some staining could be observed in the otic vesicle and the pectoral fin; and at (c) 72 hpf in the branchial arches.

3.4.4 AP000552.4 or LOC391303 novel gene

LOC391303 or AP000552.4 is a novel gene with no known function. Expression of this gene was detected in large scale cDNA sequencing project in the brain (Strausberg et al. 2002). This gene contains 4 exons and encompasses a locus of approximately 18 Kb with a transcript length of 1,530 bp would encoding a protein of 510 aa. It contains a pistil-specific extensi-like protein domain, that is a structural constituent of plant cell walls, and extensin gene expression was known to be organ-specific and temporally regulated during pistil development of plant (Goldman 1992).

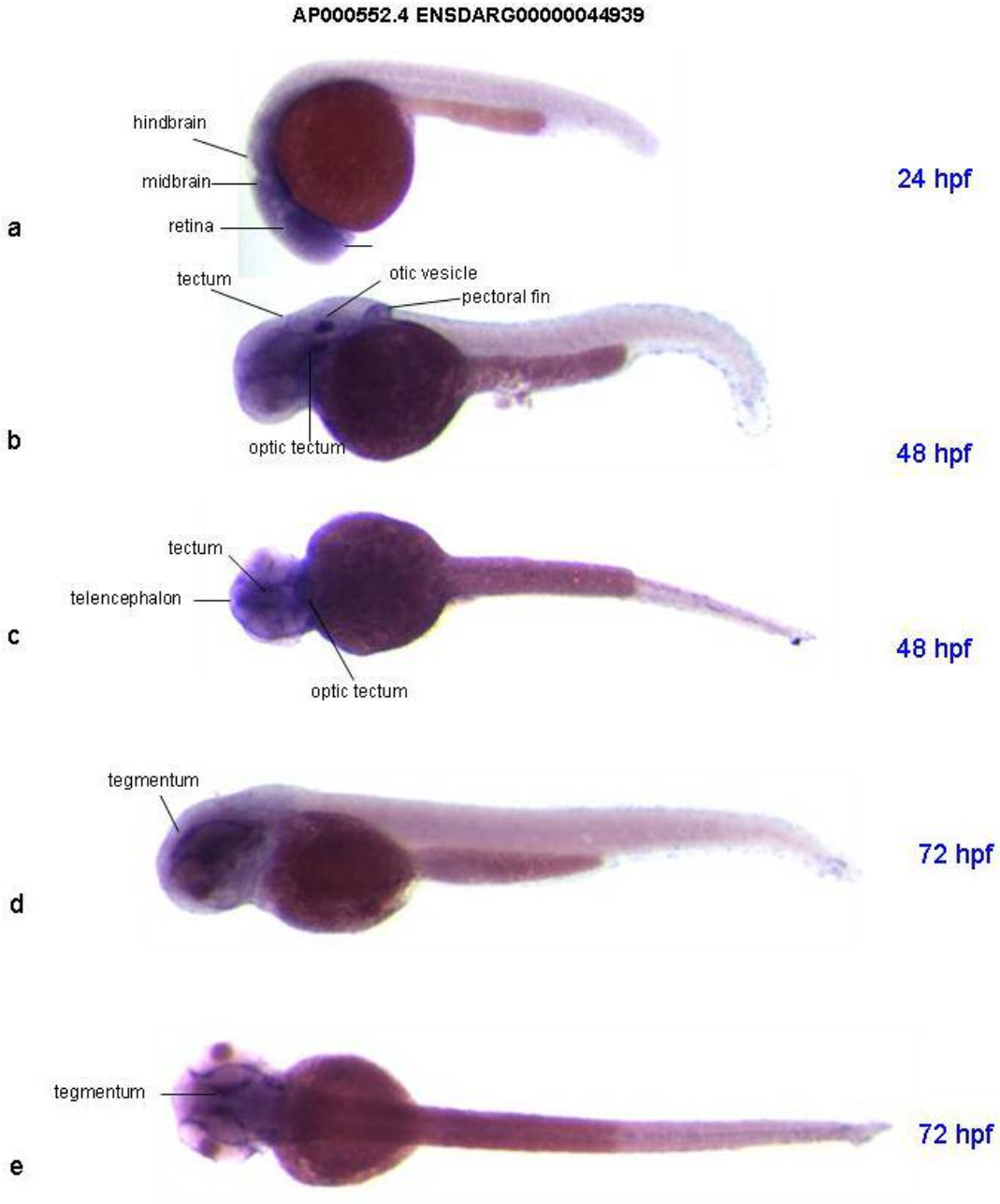

Figure 3.27 Whole mount *in situ* hybridization expression pattern for AP000552.4 Ensembl gene ID ENSDARG00000044939. Expression was observed at (a) 24 hpf in telencephalon, retina, midbrain and hindbrain; at (b) 48 hpf in telencephalon, tectum, optic tectum, otic vesicle and pectoral fin; and at (c)72 hpf in the tegmentum .

This present study confirmed earlier expression of AP000552.4 Enembl gene ID ENSDARG00000044939, in the brain and extended the information on expression specificity as shown in Figure 3.27, to the telencephalon, retina, forebrain, midbrain and the hindbrain at 24 hpf, the tectum, optic tectum and the telencephalon at 48 hpf, and in the tegmentum at 72 hpf.

3.4.5 AP000553.6 or LOC150223 novel gene

AP000553.6, is a noval 1,327 bp predicted gene with no known function encoding a 323 aa protein containing 5 exons covering a 1.95 kb locus. The predicted protein contains domains for both the YdjC-like protein possibly involved in the the cleavage of cellobiose-phosphate (Lai & Ingram 1993) and the flagellar hook-length control protein previously found to be involved in hook length control during flagellar morphogenesis in *Salmonella typhimurium* and *Escherichia coli* (Kawagishi et al. 1996).

As shown in Figure 3.28, expression of AP000553.6, Ensembl Gene ID ENSDARG00000002884, was detected in developing zebrafish embryos at 24hpf, in the notochord and at 48 hpf in both the notochord and otic vesicle, and at 72 hpf in the notochord, otic vesicle, and the liver. This experiment not only validated the prediction for this novel gene, it also demonstrated that this gene is involved in early zebrafish development.

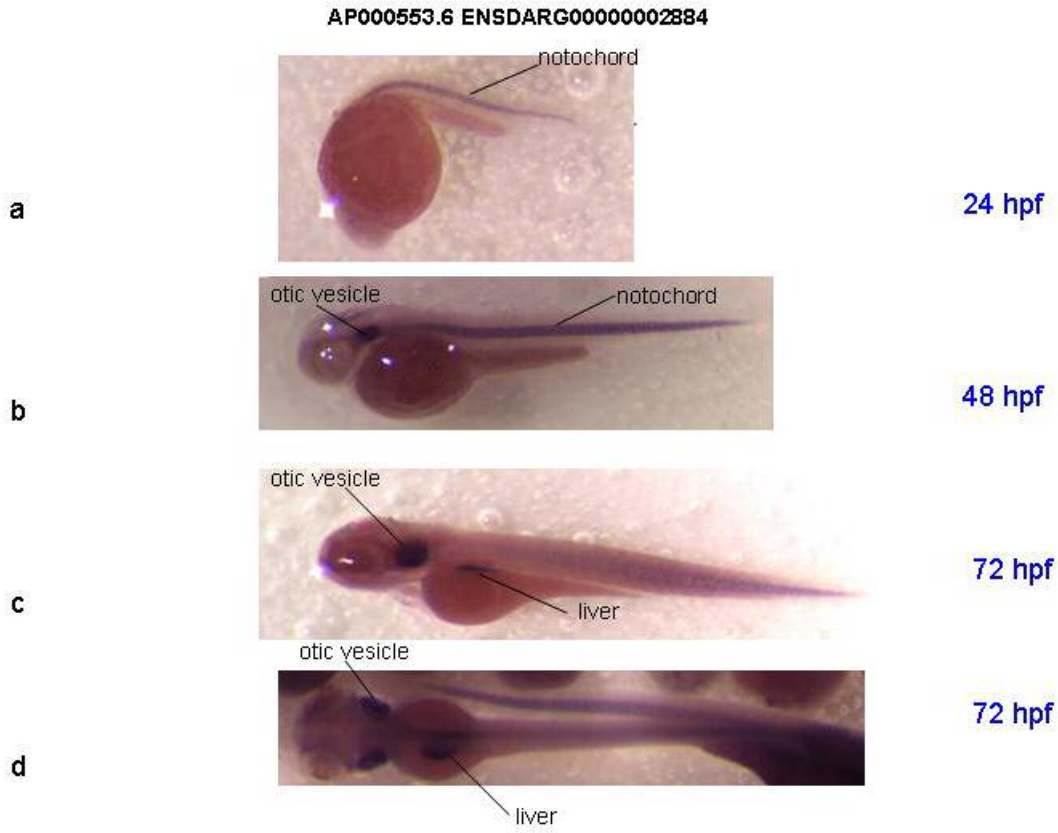

Figure 3.28 Whole mount *in situ* hybridization expression pattern for AP000553.6 Ensembl gene ID ENSDARG00000002884. Expression was observed at (a)24 hpf in the otic notochord; at (b) 48 hpf in the otic vesicle and the notochofd; at (c) 72 hpf, expression in the notochord decreased, but expression was observed in the otic vesicle and the liver.

3.4.6 Peptidylprolyl isomerase like member 2 (PPIL2) gene

Peptidylprolyl isomerase (cyclophilin)-like 2 (PPIL2) is a member of the peptidylprolyl cis-trans isomerases family also known as cyclophillins that play an important role in protein folding, protein trafficking in cells, intercellular communication, cell suface externalization of other proteins, infectious activity of HIV-1 virions (Pushkarsky et al. 2005). The cyclophilin family is highly conserved during evolution and have been observed in bacteria, fungi, plants and vertebrates

(Gothel and Marahiel 1998), and has been shown previously to be expressed in a wide variety of human tissues including thymus, pancreas, testis, small intestine, liver, colon and kidney, in addition to myelogenous leukemia cell line, and in most lymphomas and melanomas (Wang et al. 1996).

In this present study, expression of zebrafish PPIL2 Enembl gene ID ENSDARG00000002016, as shown in Figure 3.29, was observed in the forebrain, retina, hindbrain and the somites, at 24 hpf and in the brain, retina, pharyngyl arches, otic vesicle, thymus, pectoral fin, pancreas, kidney, and the caudal veins at 48 hpf. At 72 hpf, expression was observed in the brain, retina, and thymus.

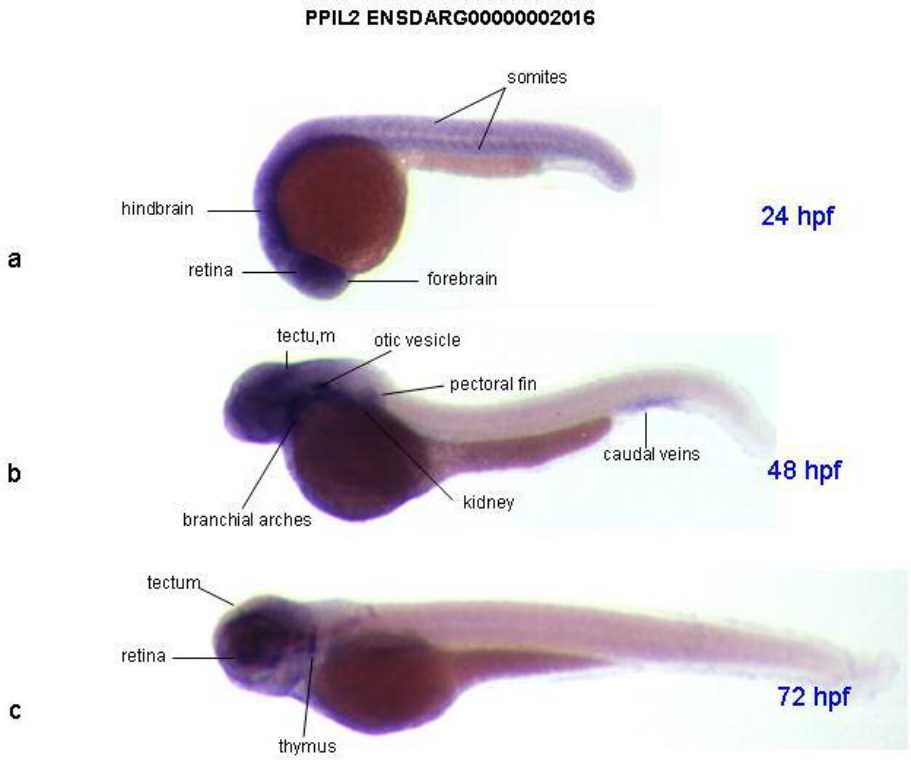

Figure 3.29 Whole mount *in situ* hybridization expression pattern for PPIL2 zebrafish homolog ENSDARG00000002016. Expression was observed at (a) 24 hpf in forebrain, retina, cerebellum, hindbrain, and somites; at (b) 48 hpf in forebrain, tectum, otic vesicle, pectoral fin, branchial arches, kidney and caudal veins; at (c) 72 hpf in the retina, tectum, and thymus.

It is interesting to see the expression of PPIL2 in the developing zebrafish embryo, indicating its roles in several of the organs that were previously demonstrated. The *Drosophila* cyclophilin homolog, NinaA, participates in trafficking of rhodopsin, and its mammalian counterparts are retina-expressed integral membrane proteins located in the lumen of the endoplasmic reticulum and intracellular transport vehicles (Pushkarsky et al. 2005). In addition, PPIL2 also was reported to be involved in the cell surface externalization of insulin receptor in the pancreas (Shiraishi et al. 2000). This study had extended and confirmed results of the previous studies mentioned, and demonstrated that PPIL2 is involved in early developmental stages of vertebrate in specific tissues.

3.4.7 Breakpoint cluster region gene (BCR) gene

The break point cluster (BCR) functional gene is located at position 21,847,625-21,984,774 on human chromosome 22 and was so named because clusters of breakpoints that cause reciprocal translocation between chromosomes 22 and 9 are located in the approximately 135 Kb BCR locus (Chissoe et al. 1995). The translocation produces the Philadelphia chromosome (Nowel & Hungerford 1960), which is often found in patients with chronic myelogenous leukemia. A consequence of the translocation is the production of a fusion protein which is encoded by sequence from both BCR and ABL, the gene at the chromosome 9 breakpoint. Although the BCR-ABL fusion protein has been extensively studied, the function of the normal BCR gene product is not clear, with the only information about it being a serine/threonine kinase and is a GTPase-activating protein for p21rac (Diekmann et al. 1991).

A phylogenetic analysis (Figure 3.30) was carried out on the BCR genes and revealed that both the zebrafish and tetraodon BCR genes are orthologous to tetrapods BCR genes and the duplication is an ancient event occured after the common ancestor of zebrafish and tetraodon diverged from the lineage of tetrapods. Alignment of the coding sequences also revealed that the human BCR gene has 75.4% and 80.6% amino acid identity with zfBCR8 and zfBCR21 respectively. The two zebrafish paralogs are highly divergent from each other having 75.6% amino acid identity with each other. Probes were generated from highly divergent regions in the exon 1 as well as from zfBCR21 specific exon 2 for *in situ* hybridization study.

From a PIP analysis as shown in Figure 3.31, the BCR gene is highly conserved in all vertebrates including studied including chimpanzee, cow, dog, mouse, rat, chicken, frog, pufferfish, tetraodon and zebrafish. While the chimpanzee show conservation in both exonic and intronic regions, other species only shows conservation in the exonic regions. However, both the tetraodon and the zebrafish possessed duplicated copies of the BCR genes. In zebrafish, one copy is located on zebrafish chromosome 8 (zfBCR8) and lacks exon 2, while the other is on zebrafish chromosome 21 (zfBCR21).

As shown in Figure 3.31A, another significant observation was an inverted repeat that occurred in the 5' promoter region and in the 3' coding and splice donor region of BCR exon 1. It was previously characterized and found to be essential for protein binding (Zhu et al. 1990, Chissoe et al. 1995), and is highly conserved in all vertebrates sequences compared, including zebrafish BCR gene on chromosome 8, except in the frog BCR gene and zebrafish BCR gene on chromosome 21.

Phylogenetic tree of BCR gene

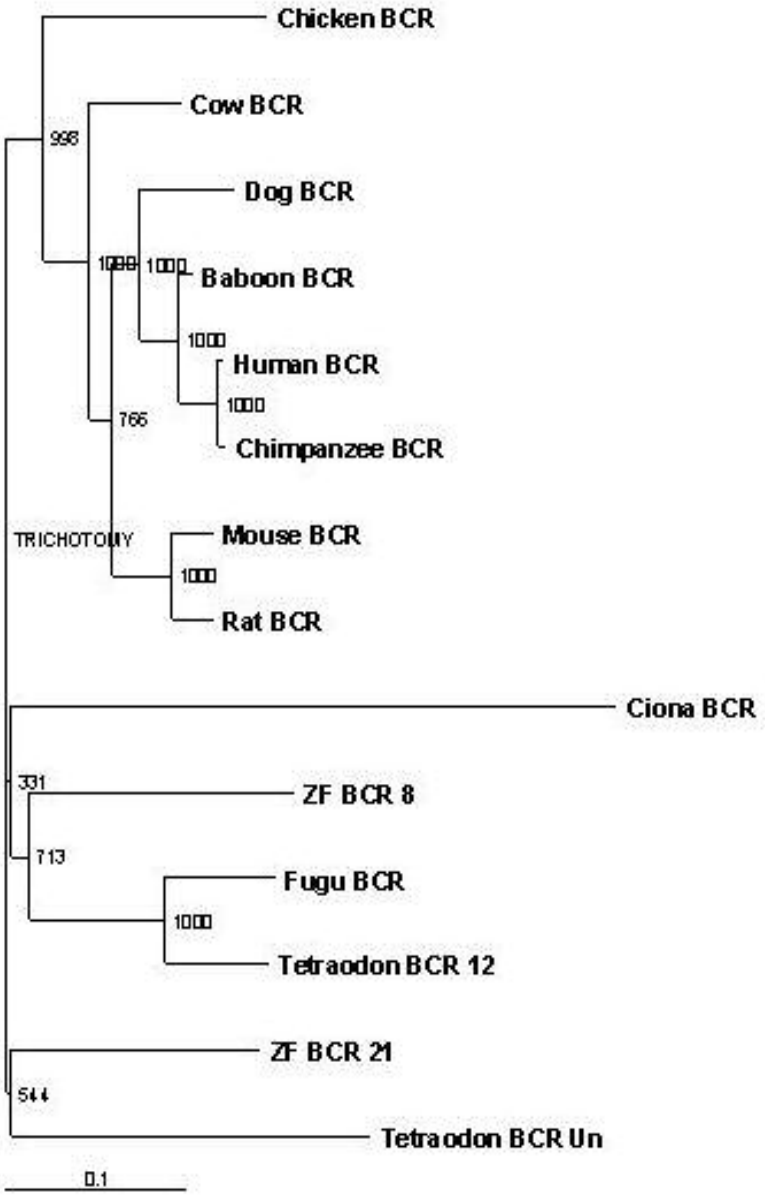

Figure 3.30 Phylogeny of BCR genes. Sequences were aligned by ClustalX and phylogenetic tree was constructed using the neighbour-joining method (Saitou and Nei 1987). Numbers on tree nodes are bootstrap values based on 1000 runs. Scale on the tree shows a 0.1 substitution per nucleotide.

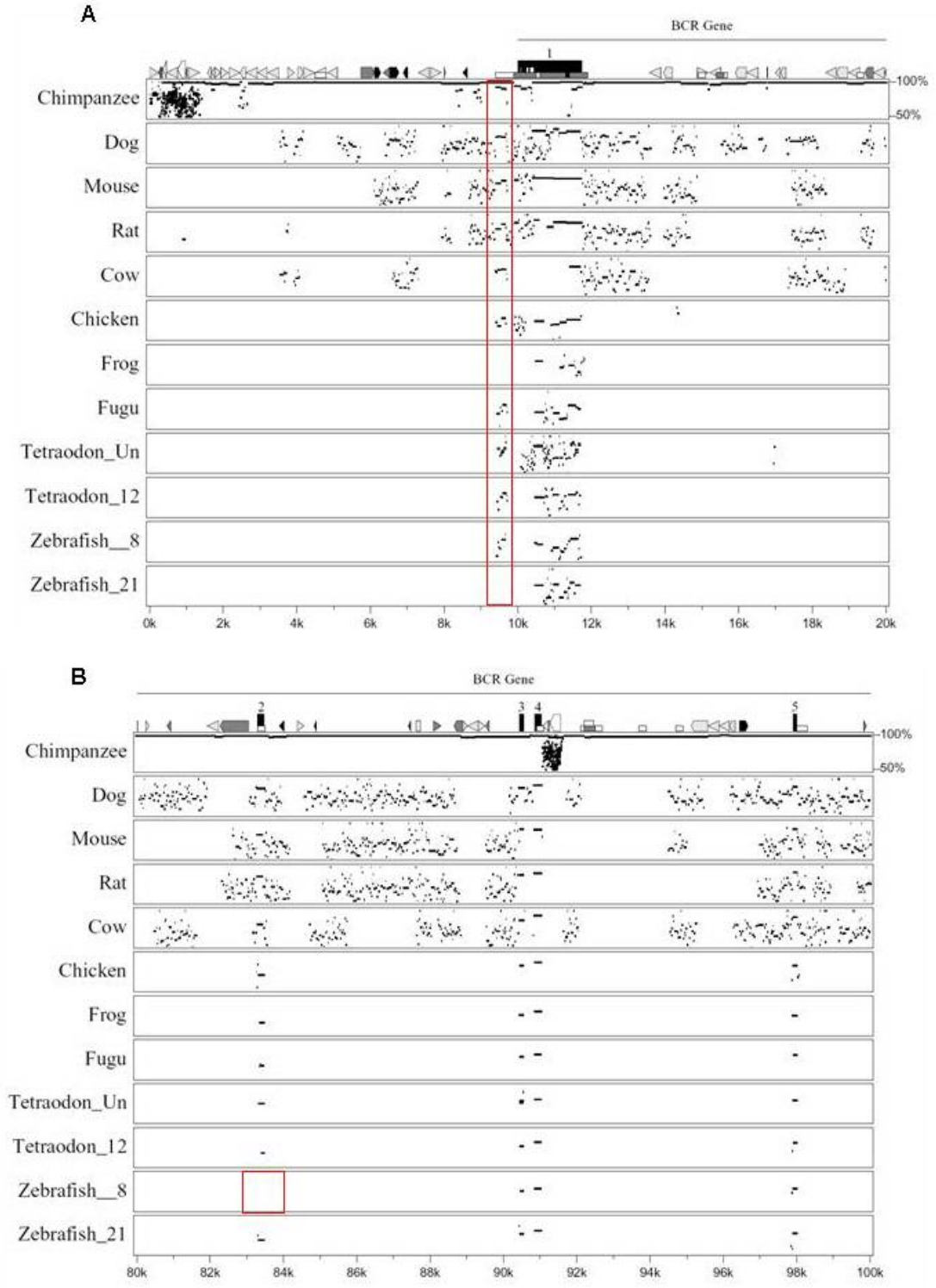

Figure 3.31 PIP plot showing (A) BCR exon 1, red box shows the inverted repeat in the promoter region, and (B) BCR exons 2 -5, red box shows zfBCR8 lacking the exon 2.

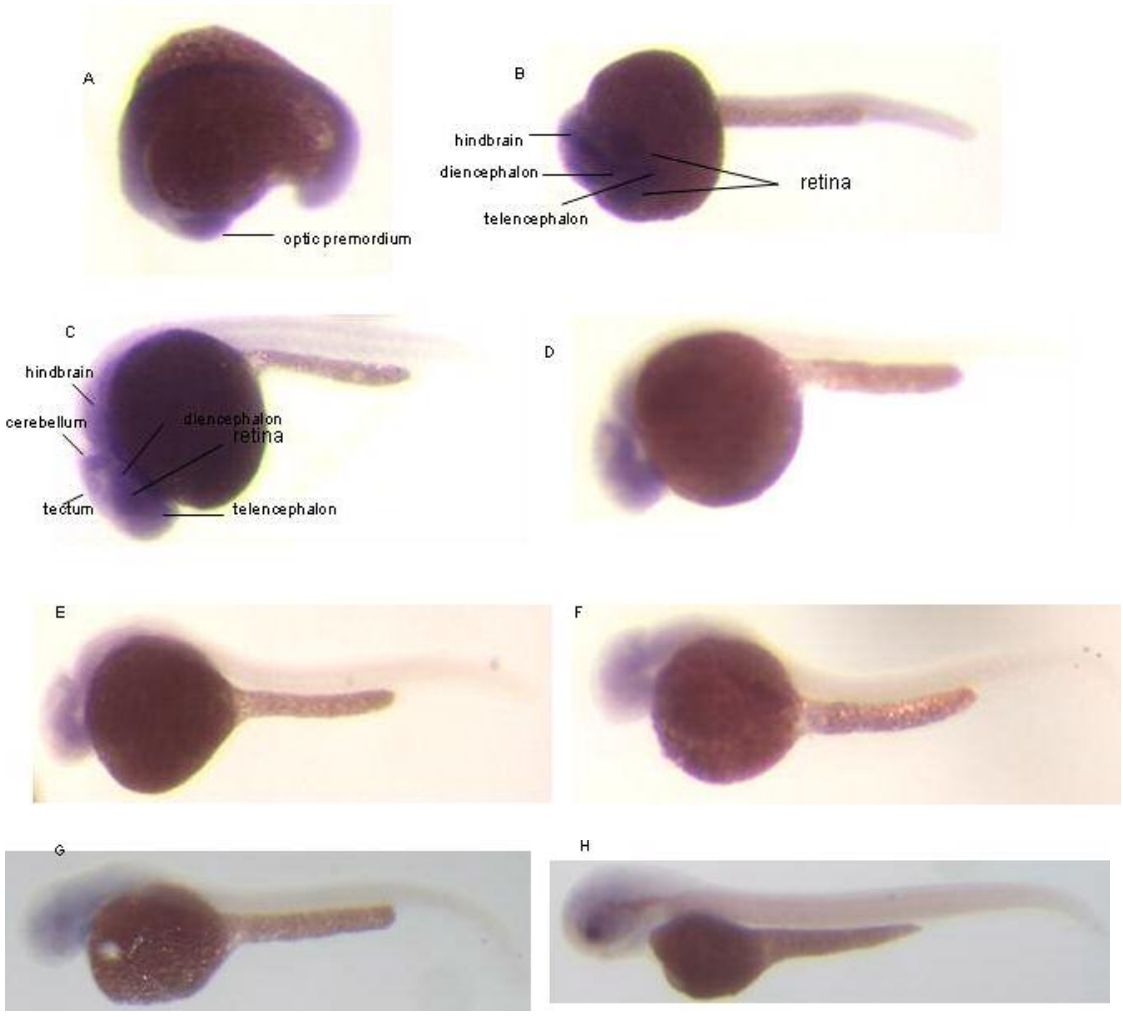

Figure 3.32 The whole mount in situ expression pattern of zfBCR8 shown here at (A) 16 hpf primarily in the optic primordium and undifferentiated brain rudiment that will later develop into various brain sections, and at 24 hpf (B),(C) in differentiated telencephalon, diencephalons, retina, tectum, cerebellum and hindbrain. Specific staining pattern was not observed in the trunk region. Expression pattern was observed to decrease from (D) 28 hpf, (E) 32 hpf and (F) 36 hpf where clear differenctiated expression pattern can be barely seen. Expression pattern pattern for this gene was not observed in the (G) 48 hpf and (H) 72 hpf embryos.

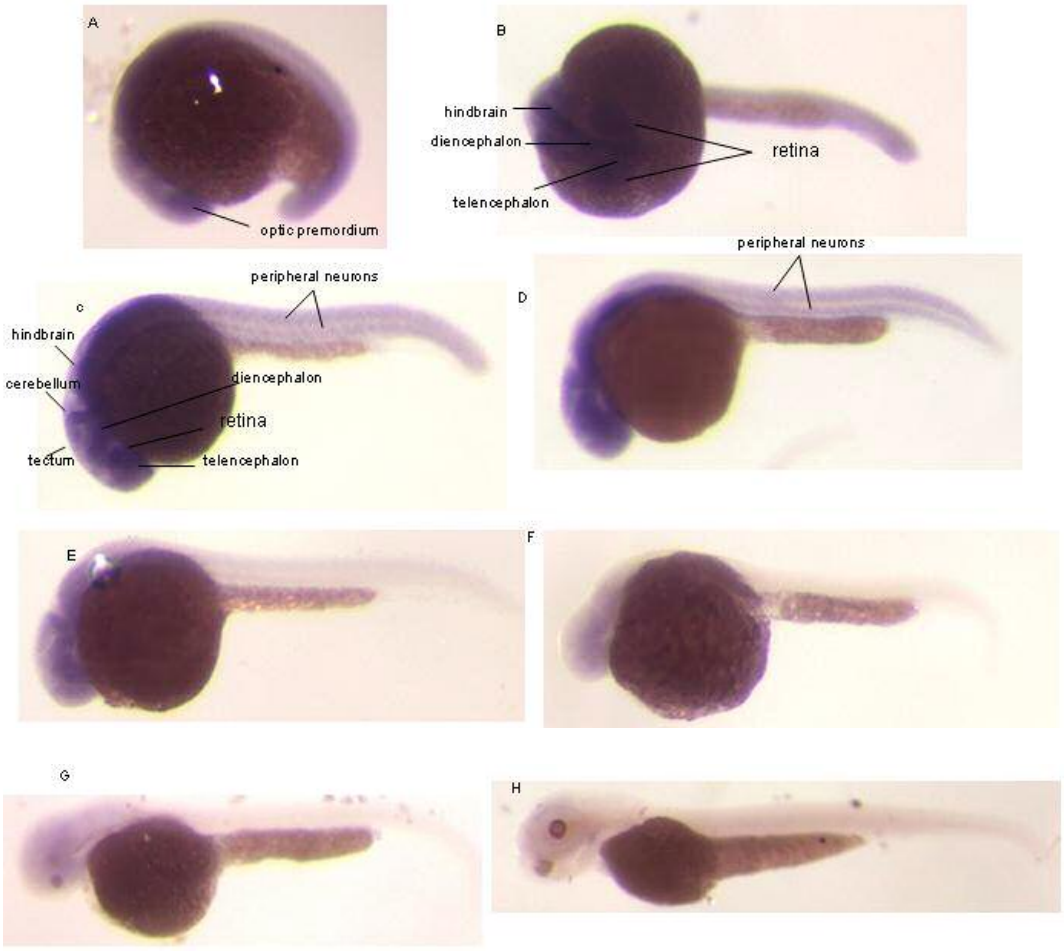

Figure 3.33 The whole mount in situ expression pattern of zfBCR21 shown here at (A) 16 hpf primarily in the optic primordium and undifferentiated brain rudiment that will later develop into various brain sections, and at 24 hpf (B),(C) in differentiated telencephalon, diencephalons, retina, tectum, cerebellum and hindbrain. Staining was also observed in the trunk believed to be the peripheral neurons. Expression pattern was observed to decrease in (D) 28 hpf, but expression in the trunk region could still be observed. At (E) 32 hpf, expression in the trunk was not seen and at (F) 36 hpf clear differenctiated expression pattern could hardly be seen. Expression pattern pattern for this gene was not observed in the (G) 48 hpf and (H) 72 hpf embryos.

The expression of zebrafish BCR homologs was observed in early developmental stages from 16 hpf to 36 hpf as shown in Figure 3.32 and 3.33 and shows for the first time that this gene is involved in early development of the central nervous system, particularly in the brain and the optic system, which at early embryonic stages starts as the optic primordium that is an outgrowth from the brain. Also, of the two zebrafish BCR copies, zfBCR8 showed specific expression only in the brain and retina and no expression in the trunk region, while zfBCR21 showed expression in the brain and retina, higher expression in the hindbrain and additional expression in the trunk region that likely corresponds to the peripheral neurons.

The location of the zebrafish BCR homologs on the duplicated syntenic blocks and phlyogenetic analysis demonstrates that both zfBCR8 and zfBCR21 are orthologs of the mammalian BCR gene, instead of one being the results of a tandem duplication events occurring within the zebrafish genome. This is consistant with the previously hypothesized genome duplication in ray-finned fish (Amores et al. 1998; Postlethwait et al. 1998). It is believed that the complexities and differences among organisms arise from gene duplication and the subsequent evolution of gene functions among the duplicates. Two models have been proposed for the preservation of duplicated genes. The classical model (Haldane 1933; Fisher 1935) hypothesizes that in most cases one member of the duplicated genes degenerates through time while the other retains the original function. Only in rare occasions, one duplicate may acquire new adaptive function, resulting in the preservation of both duplicates, one with a new function and the other retaining the old. A recently proposed duplication-degeneration-complementation (DDC) model (Force et al. 1999; Lynch and Force 2000)

hypothesized that degenerative mutations in regulatory elements increase the probability of the preservation of duplicated genes, and the partitioning of ancestral functions is usually how duplicated genes are preserved, rather than preservation by acquiring new functions.

Recent data is consistant with the DDC model. These studies include the Hox clusters genes (McGinnis et al. 1992; Godsave et al. 1994; Amores et al. 1998; McClintock et al. 2002), the engrailed genes (Amores et al. 1998) the SOX9 genes (Yan et al. 2002), and the Fox1 genes (Solomon et al. 2003). Sequence analysis and our expression pattern analysis also suggests that the zebrafish BCR duplicates fit the DDC model as well. Comparison of the BCR duplicates in zebrafish demonstrated that their major difference lies in the inverted repeat in the promoter region that previously was shown to be a protein-binding sequence through DNase1 footprinting studies (Zhu et al. 1990). Its role as a functionally important cis-regulatory element is further reinforced by its highly conserved sequence among all vertebrates compared. Expression pattern profiling shows both gene copies are expressed in the same organs at the same developemtal stages. The distinction between the two is zfBCR8, that retained this regulating region, is seen mostly in the anterior brain region, but zfBCR21, without this regulating region, increased expression was observed in the hindbrain and the peripheral nervous system. These observations supportthe DDC model of gene evolution and the present information gained for the BCR gene will facilitate future experimental designs to elucidate its cellular and physiological function.

3.4.8 AP000348.4 or Chromosome 22 ORF 16 (C22orf16) novel gene

AP000348.4 is a novel 681 bp, 142 aa residue protein coding gene, with 4 exons occupying a 2.1 Kb locus, that also is named C22orf16 based on it being chromosome 22 open reading frame number 16 that has no known function. C22orf16 encodes a protein with a coiled coil-helix coiled coil-helix domain (CHCH) that has a backbone consisting of a 10-amino-acid coiled coil region, followed by two 15-amino-acid alpha-helices connected by a coiled coil region of 5 to 10 amino acids (Westerman et al. 2004).

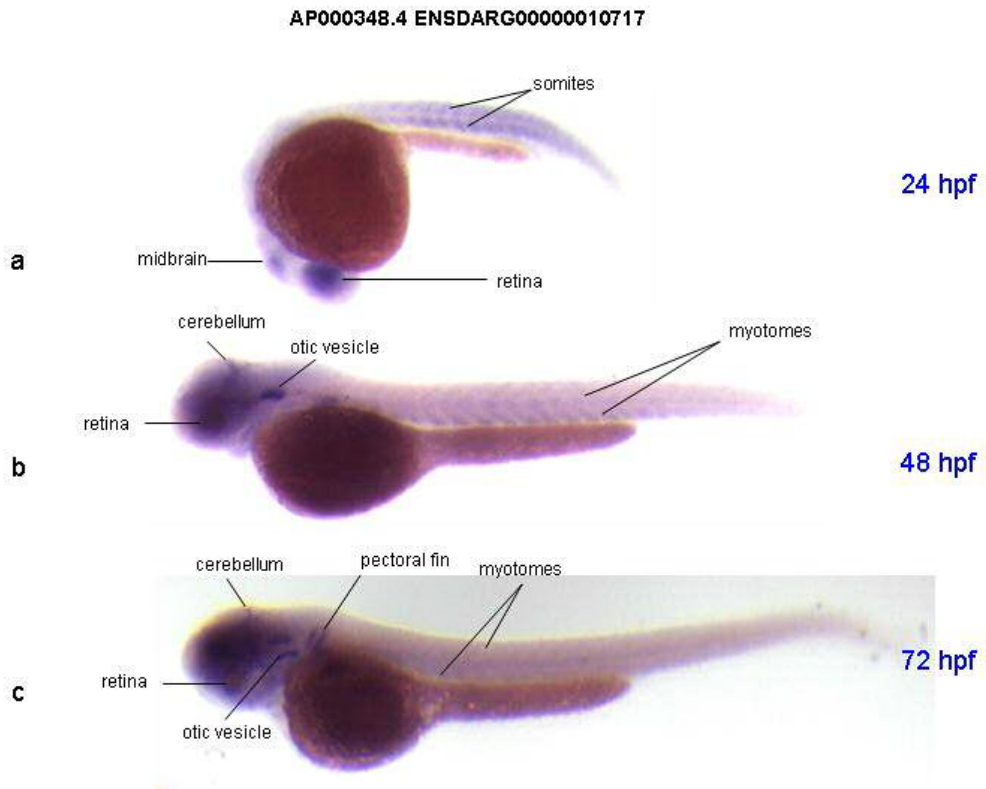

Figure 3.34 Whole mount *in situ* hybridization expression pattern for AP000348.4 Ensembl gene ID ENSDARG00000010717. The expression was observed at (a) 24 hpf in the retina, midbrain and the somites; at (b) 48 hpf in the retina, tectum, otic vesicle and the myotomes; and at (c) 72 hpf in the retina, tectum and myotomes.

The two alpha helices likely form a hairpin–loop that is stabilized bytwo disulfide bonds formed by two cysteine pairs separated by 9 amino acids. This novel protein family has a conserved motif that was found in plant, yeast, fruitfly, worm, mouse and human proteins, and members of this family were recently isolated from a unique first trimester placental cDNA library (Westerman et al. 2004).

Expression of this gene was observed in developing zebrafish embryos, mostly the brain, retina, and developing somites or myotomes developing zebrafish embryos as shown in Figure 3.34, with expression at 24 hpf detected in the retina, midbrain and the somites, at 48 hpf in the retina, tectum, otic vesicle and the myotomes and at 72 hpf, in the retina, tectum, the otic vesicle and the myotomes, supporting the previous study that implicated this gene family in the early developmental stages of vertebrates (Westerman et al. 2004).

3.4.9 Matrix metalloproteinases 11 or Stromelysin III gene

Matrix metalloproteinase 11 (MMP11), or Stromelysin III (STMY3), is a member of the matrix metalloproteinases family (Nagase et al. 1992) that is involved in the breakdown of extracellular matrix in normal physiological processes, including embryonic development, reproduction, and tissue remodeling, as well as in arthritis and metastasis (Matrisian 1990). Most MMP's are secreted as inactive proproteins that are activated when cleaved by extracellular proteinases. However, MMP11 is activated intracellularly by furin within the constitutive secretory pathway, and in contrast to other MMP's, this enzyme cleaves alpha 1-proteinase inhibitor but weakly degrades structural proteins of the extracellular matrix (Matrisian 1990).

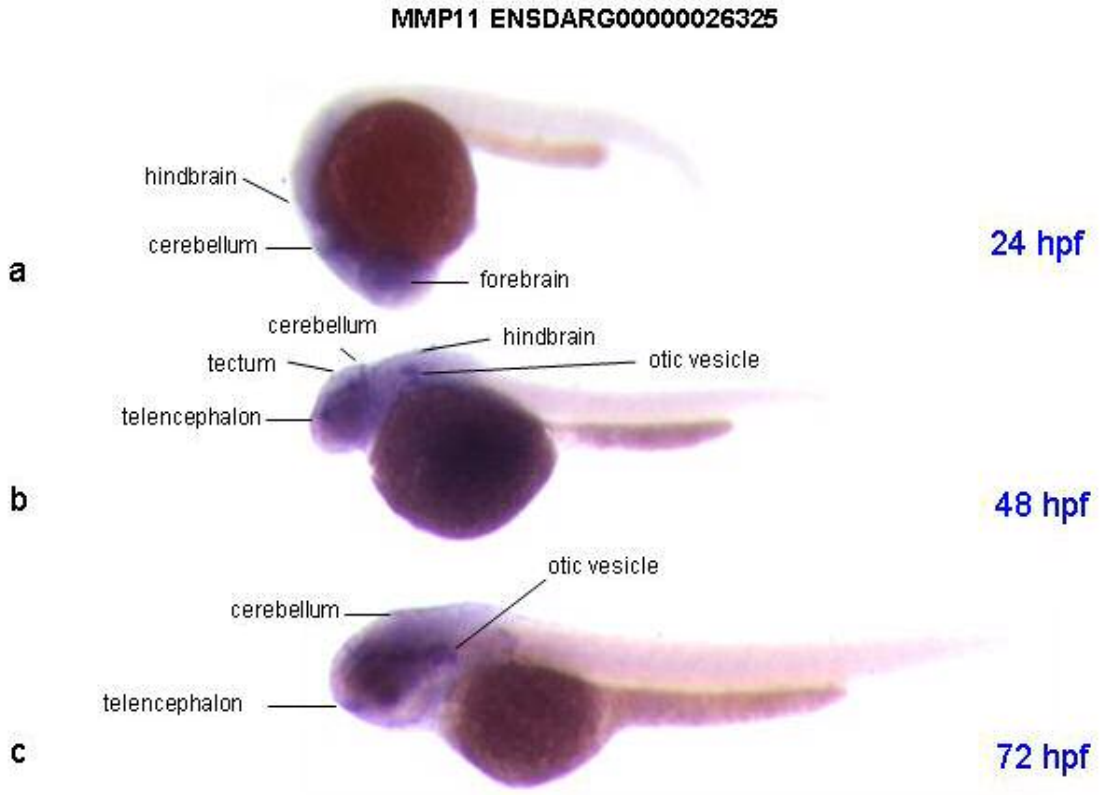

Figure 3.35 Whole mount *in situ* hybridization expression pattern for MMP11 Ensembl gene ID ENSDARG00000026325. Expression was observed at (a) 24 hpf in the forebrain, cerebellum, and the hindbrain; at (b) 48 hpf in the telencephalon, tectum, cerebellum, hindbrain, and otic vesicle; and at (c) 72 hpf in the telencephalon and the otic vesicle.

The MMP11 zebrafish homolog, Ensembl gene ID ENSDARG00000026325 was found expressed in developing zebrafish embryos, as shown in Figure 3.35, in the forebrain, cerebellum, and hindbrain at 24 hpf, in the telencephalon, tectum, cerebellum, and hindbrain at 48 hpf, and in the telencephalon at 72 hpf, extending previous knowledge of its role in embryonic development and specifying its involvement in the development of the brain

3.4.10 Solute carrier family 2 member 11 (SLC2A11) gene

The solute carrier family 2 (SLC2A) is a family of glucose transporters thatare integral membrane glycoproteins playing significant role in transporting glucose and maintaining glucose homeostasis in most cells. Three members of the solute carrier family, SLC2A1, SLC2A3, and SLC2A9 were found to be expressed in human chondrocytes (Mobasheri et al. 2002), the only cells found in cartilage which produce and maintain the cartilagenous matrix. Since glucose functions as the metabolite and structural precursor for articular cartilage, these SLC2A family members are critical to transport the glucose needed for cartilage development and function in the chondrocytes (Mobasheri et al. 2002).

The expression patterns of a zebrafish SLC2A11 homolog, Ensembl gene ID ENSDARG00000034501, as shown in Figure 3.36, initially occured at 24 hpf in the premigratory cranial and trunk neural crest cells and then at 48 hpf in actively migrating crest cells as the zebrafish embryos developed. Later in development, its expression was observed in structures believed to be the embryos' head cartilages at 72 hpf.

The neural crest cells are pluripotent cells that ultimately develop into diverse cell types. They are derived from the dorsolateral central nervous system primodium during the segmentation period, and undergo extensive migrations, resulting in stereotypic distribution of cell types within the zebrafish embryos. Neural crest cells have the potential to become a wide variety of differentiated cell types, including melanocyte, neuron, glial cell, cardiac smooth muscle, and head cartilage (Westerfield 1993).

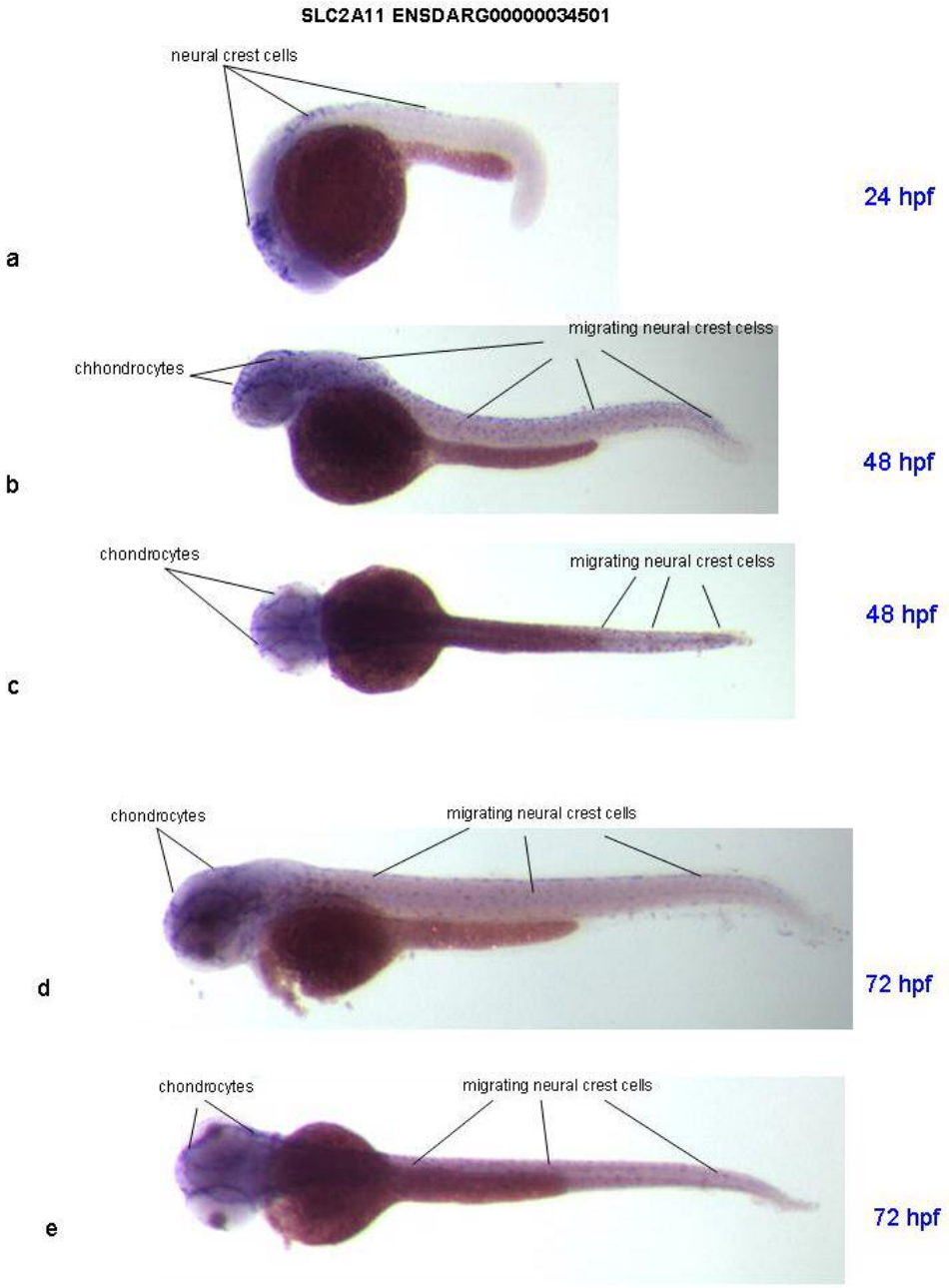

Figure 3.36 Whole mount *in situ* hybridization expression pattern for SLC2A11 Ensembl gene ID ENSDARG00000034501 at (a) 24 hpf in the cranial and trunk neural crest cells; at (b), (c) 48 hpf in the migrating neural crest cells and chondrocytes of the head cartilages; at (d),(e) 72 hpf in migrating neural crest cells as well as in head cartilages.

The expression pattern of the SLC2A11 gene ENSDARG00000034501 shows that the gene is expressed in developing neural crest cells as they first arise, migrate,

and differentiate into their specific lineage with gene expresson following the progression of the neural crest cells that led to the formation of head cartilage. This study has extended previous findings on members of the SLC2A gene family, and shows that SLC2A11 is involved in vertebrate development in a stage and tissue specific manner.

3.4.11 AP000354.2 or KIAA0376 novel gene

AP000354.2 is a novel predicted gene that has no known function but has homology with human cDNA sequence KIAA0376 from a human brain cDNA library (Ohara et al. 1997). The gene is encoded by 6,202 bp in exonic sequence, for a 1,117 aa residues protein containing 17 exons that encompasses an approximately 146 Kb region. The gene encodes a calponin homology (CH) domain that had been found in both cytoskeletal proteins and signal transduction proteins (Stradal et al. 1998).

Expression of AP000354.2, Ensembl gene ID ENSDARG00000006719 as shown in Figure 3.37, was observed in a wide range of tissue in the developing zebrafish embryos. At 24 hpf, expression was observed in the forbrain, posterior midbrain, hindbrain, notochord, pectoral fin, myotomes, tail bud and the proctodeum, at 48hpf, in the telencephalon, midbrain, hindbrain, otic vesicle, pectoral fin, lateral line ganglia and the proctodeum are stained and at 72 hpf in the heart, the otic vesicle, branchial arches, lateral line ganglia and the proctodeum. The expression profiling of AP000354.2 indicates that AP000354.2 likely is playing a ubiquitous role in a variety of developing tissues.

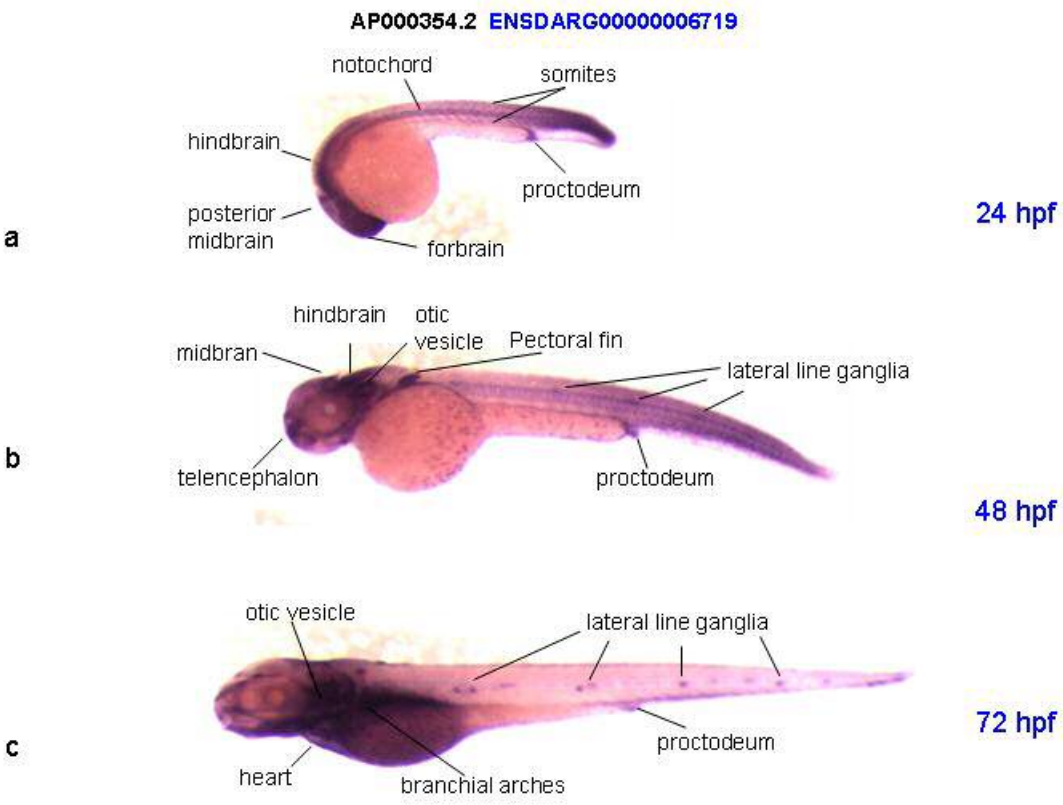

Figure 3.37 Whole mount *in situ* hybridization expression pattern for AP000354.2 Ensembl gene ID ENSDARG00000006719. Expression was observed at (a)24 hpf in the forebrain, cerebellum and the hindbrain; at (b) 48 hpf in the telencephalon, midbrain, hindbrain, otic vesicle, pectoral fin, the lateral line ganglia and the proteodeum; at (c) 72 hpf in the otic vesicle, heart, branchial arches, lateral line ganglia and the proctodeum.

Chapter IV Conclusion

4.1 Comparative sequence analysis

The region of human chromosome 22 between markers D22s1687 and D22s419, contains LCR22s with predominantly duplicated partial genes or peudogenes, the IGLL region with its the repetitive immunoglobulin gene segments, and the BCR region encoding the chromosomal breakpoints implicated in ALL and CML. Thus, a comparative genome study of this unique genome region provides evolutionary insights into these unique features. The comparison has revealed specific evolutionarily conserved and altered regions, that imply the mechanisms for their evolution, and to a certain extent how each of these affects or alters functions in either the normal or pathological phenotype.

Genome level comparisons between humans and our closest living relative, the chimpanzee, reveals highly conserved major chromosomal structures and gene organization as the overall DNA sequences are approximately 97.6% identical. Protein coding genes also are highly conserved, having an average of approximately 99.2% sequence identity in the exons and 98.8% sequence identity in the introns. IGLV genes have a slightly lower sequence identity with approximately 98.0% identity in the exons and approximately 97.8% identity in their introns resulting in minute changes at the amino acid level. The Ka/Ks ratio between known human and chimpanzee known protein coding genes is 0.25, indicating that 75% of the nucleotide substitution that led to amino acids changes are eliminated by natural selection, and major amino acid changes observed are between the same class of hydrophilic amino acids postulated to be on the protein surface, followed by changes between the same class of hydrophobic

amino acids in the protein interior, followed by the fewest amino acid changes between hydrophobic to hydrophilic amino acids.

Comparison between human and other species in the IGLL region revealed that the human IGLL region has undergone duplication followed by deletion of the duplicated sub-regions and/or tandemly duplicated units within the sub-regions. Five major IGLL sub-regions I, II, III, IV and V, are interrupted by regions containing no genes or non immunoglobulin genes and pseudogenes, including the LCR22-5 that separated IGLL sub-region I and II. These sub-regions contain duplicated units larger than 5 kb with multiple IGLV genes or pseudogenes in humans (Kawasaki et al. 2000). The light-chain immunoglobulin genes that are involved in vertebrate adaptive immune system arosesince jawed vertebrates shared a common ancestor mor e than 500 mya (Litman et al. 2005). Comparison between the human and chimpanzee IGLL region shows major insertion and deletion events that occurred since humans and chimpanzees diverged. Four major human insertions were discovered in this region, ranging in size from 6 Kb to 75 Kb, while three major chimpanzee insertions were observed in the IGLL region, ranging in size from 12 Kb to 74 Kb, while the 74 Kb chimpanzee insertion likely is an intrachromosomal inverted duplication from a distal region on chromosome 22. IGLV gene segments are more highly diverged between humans and chimpanzees when compared to known protein coding genes. Comparison of the LCR22s revealed that two human insertions of 59 Kb and 36 Kb were observed in LCR22-6 and a 67 Kb duplication from LCR22-4 was duplicated and inserted in chimpanzee LCR22-8.

As a result of these insertions and deletions, 6 IGLV functional, 12 IGLV

pseudogenes, 4 partially duplicated genes and 6 pseudogenes are human specific while 9 predicted genes are chimpanzee specific.

The IGLL and LCR22s, both consist of duplicated and repeated segments on human chromosome 22 that likely have evolved rapidly by changes in exon numbers through small scale insertions, deletions and exon shuffling. In addition, rapid accumulation of amino-acid-changing base substitution also occured through positive selection with Ka/Ks value >1, and increased hydrophobic to hydrophyllic amino acid substitutions.

The LCR22s only occured in human, chimpanzee and baboon, but are absent in other vertebrates. Thus, it can be concluded that the LCR22s is a unique feature that arise by segmental duplication prior to the time that humans shared a common ancestor with other primates, a finding consistent with previous studies that concluded thsese recent segmental duplications arose 35 Mya during early primate evolution (Bailey et al. 2002). While segmental duplications such as LCR22s were previously only known to be involved in pathological conditions such as Cat Eye Syndrome and DiGeorge Syndrome as a result of their highly identical sequences within each species, cross species comparison between human and chimpanzee LCR22s now brings a new perspective in their role for primate speciation. As shown in previous studies, for example the Kruppel-associated zinc-finger genes (Eichler et al. 1998) on human chromosome 19 and the newly characterized morpheus gene family on human choromosome 16 (Johnson et al. 2001), gene families in primates evolved by duplication of genomic segments followed by exon shuffling or rapid sequence divergence and positive selection in the coding region that fixes amino-acid-changing

base substitution (Johnson et al. 2001; Bailey et al. 2002). Thus, it is likely that the highly duplicated and rapidly evolving IGLL and the LCR22s regions of human and chimpanzee offer unique evolutionary avenues for the creation of new genes or gene families, and may in part account for underlying phenotypic differences in primates.

Transcription of partially duplicated genes and peudogenes unique to LCR22s previously has been previously reported (Bailey et al. 2002), and it is known that although truncated genes or pseudogenes might lack the ability to produce functional proteins, they are transcribed and greatly affect paralogous functional copies at the transcriptional level as in the case of pseudo-NOS and makorin1-p1 (Korneev et al. 1999; Hirotsune et al. 2003; Harrison et al. 2005). Thus, the loss and gain of the partially duplicated genes and pseudogenes specific to humans and chimpanzees might also be significant factors affecting the differences in transcription of their functional paralogs.

Comparison of the approximately 135 kb BCR region reveals that the functional BCR gene is highly conserved in all vertebrates and is duplicated in zebrafish. This entire region was highly conserved in humans, chimpanzees, and baboons in the exonic, intronic and intergenic regions. The protein coding BCR gene exonic region also was conserved in all vertebrates compared but when compared to human, the level of repetitive elements was decreased drastically in the other mammalian genome e.g. dog, cow, mouse and rat, and these elements were not present at all in the chicken, frog, pufferfish, and zebrafish genome. These results are consistant with the genome wide generalization of decreasing repetitive elements observed in other primates, mammals and vertebrates when compared to human. Since all vertebrates except teleost fish had

only one copy of the functional, 23 exons, BCR it is very likely that the duplicated copies in zebrafish on chromosome 8 and chromosome 21 resulted from the postulated whole genome duplication rather than a tandem duplication of this specific region. Two independent data sets, gene phylogenies based on sequence information, and location of the duplicated genes on the syntenic blocks confirms both zebrafish duplicated BCR genes shared common ancestors with that of mammalian BCR gene and arose as a result of chromosomal duplication. This finding is consistent with previous data that supports the theory of a genome wide duplication in the ray-finned fish (Amores et al. 1998; Postlethwait et al. 1998).

An inverted repeat in the 5' promoter region and in the 3' coding and splice donor region of BCR exon 1 was characterized previously and postulated to be a protein binding regulatory element (Zhu et al. 1990, Chissoe et al. 1995). This inverted repeat was conserved in all vertebrate sequence compared except the frog and the duplicate copy of the zebrafish BCR gene on zebrafish chromosome 21. The other differences between the two zebrafish BCR copies were that human BCR exon 2 was not present in the chromosome 8 duplicate. These results are consistant with previous assertion of adaptive evolution following gene duplication (Ohno 1970), which likely is the major underlying evolutionary pathway found among primates and among distant vertebrates such as the fishes.

4.2 Gene expression profiling in zebrafish embryos

Information on the expression of genes bridges the gap between DNA sequence and function in an organism. As an extension to my comparative sequence analysis, a high-throughput 96-well format whole mount *in situ* hybridization protocol for zebrafish expression profiling was developed, which provided reproducible analysis of the expression pattern of several zebrafish homologs of human chromosome 22.

Through these gene expression studies I observed for the first time that the zebrafish BCR genes were expressed in early developmental stages, with high expression detected clearly in the optic primordium, brain rudiment before morphological subdivision, diaencephalon, dorsal midbrain, cerebellum, and hindbrain during the segmentation period at 24 hpf, and slowly decreasing before the first half of the phyaryngula state before 36 hpf with no expression in 48 hpf and 72 hpf embryos. The interesting discovery was that expression of the chromosome 8 duplicate was found to be confined to the optic primordium and brain region but the chromosome 21 duplicate was found also expressed highly in the somites and myotomes in the trunk region. From the expression data, it can be concluded that the BCR gene is involved in early development of central nervous system, and that the inverted repeat in the promoter region conserved between the human BCR gene and zebrafish BCR chromosome 8 duplicate is a cis-element regulating gene transcription likely plays a role in its expression exclusively in the optic primordium and the brain.

Through my expression profiling studies, an additional set of genes, for which little or no expression data was previously available, were shown to be involved in embryonic development. These genes include the AIFL involved in apoptosis, Thap7

involved in transcription repression, and MMP11 involved in the breakdown of the extracellular matrix. While the SLC2A gene family previously was reported to be involved in morphological integrity of head cartilage, now we know that SLC2A11 expression is localized to migrating neural crest cells, an important precursor cells in vertebrate developmental that gives rise to the pigment cells, peripheral neurons and glia, or head cartilage. Genes with previously known biochemical functions but no known physiological function, such as BCR, a serine/threonine kinase which was previously known only for its role in leukemias, SLC7A7 a cationic amino acid transporter , and PPIL a peptidylprolyl isomerase, now has been shown to be involved exclusively in vertebrate embryonic development. In addition, four other newly characterized genes that have no previously known function, C22orf16, LOC391303, KIAA0376, LOC150223 now we know their their specific expression profiles (Table 4.1) during specific zebrafish embryonic developmental stages.

Taken together, the results support my third major conclusion that a high percentage of the genes that are expressed in developing zebrafish embryos in the chromosome 22 region studied are involved in the developmental pathways of vertebrates. These studies thus form the foundation of future experiments aimed at determining the function of these genes using zebrafish as the experimental model.

Gene Name	Whole mount *in situ* expression pattern of human homologous zebrafish genes
SLC2A11	premigratory cranial and trunk neural crest cells, migrating crest cells, chondrocytes
AIFL	forbrain, tectum, cerebellum, midbrain, hindbrain
SLC7A4	Only at 24 hpf, -forebrain, cerebellum, midbrain, and hindbrain
Thap7	24 hpf,-forebrain, retina, cerebellum and the hindbrain. 48 hpf, -optic tectum and the tegmentum 72hpf , - tegmentum.
PPIL2	central nervous system, the retina, and the somites at the trunk area
BCR_Chr8	16 hpf, 20 hpf, -optic primordium, brain rudiment 24 hpf – 36hpf diaencephalon, dorsal midbrain, cerebellum, and hindbrain
BCR_Chr21	16 hpf, 20 hpf, -optic primordium, brain rudiment 24 hpf – 36hpf diaencephalon, dorsal midbrain, cerebellum, hindbrain, myotomes
MMP11	telencephalon, tectum, cerebellum, and hindbrain
C22orf16	brain, retina, and developing somites or myotomes
LOC391303	24 hpf, retina, forebrain, midbrain and the hindbrain. 48 hpf, expression was seen in the tectum, optic tectum and the telencephalon 72 hpf, tegmentum
KIAA0376	24 hpf, forbrain, posterior midbrain, hindbrain, notochord, pectoral fin, myotomes, tail bud and the proctodeum 48 hpf, telencephalon, midbrain, hindbrain, otic vesicle, pectoral fin, lateral line ganglia and the proctodeum72 hpf, heart, the otic vesicle, branchial arches, lateral line ganglia and the proctodeum
LOC150223	Otic vesicle, notochord, liver

Table 4.1 A summary of specific gene expression pattern in developing zebrafish embryos

References

Adoutte, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme, B., and deRosa, R. (2000). The new animal phylogeny, Reliability and implications. Proc. Natl. Acad. Sci. USA **97**, 4453-4456.

Allaman-Pillet, N., Djemai, A., Bonny, C., and Schorderet, D.F. (2001). The 5′ Repeat Elements of the Mouse Xist Gene Inhibit the Transcription of X-Linked Genes. *Gene Expression*, **9,** 93-101.

Allshire, R.C., Dempster, M., Hastie, N.D. (1989). Human telomeres contain at least three types of G rich repeat distributed non-randomly. *Nucleic Acid Res.* **17**, 4611-4627.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Altschul, S.F., Madden, T.L., Schoffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST, A new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402.

Amores, A., A. Force, Y.L. Yan, L. Joly, C. Amemiya, A. Fritz, R.K. Ho, J. Langeland, V. Prince, Y.L. Wang et al. (1998). Zebrafish *hox* clusters and vertebrate genome evolution. *Science* **282**, 1711–1714.

Angata, T., Varki, N.M., and Varki, A. (2001). A second uniquely human mutation affecting sialic acid biology. *J. Biol. Chem.* **276,** 40282-40287.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M.D., Roach, J., Oh, T., Ho, I.Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S.F., Clark, M.S., Edwards, Y.J., Doggett, N., Zharkikh, A., Tavtigian, S.V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen., L, Hood., L, Tan,. YH, Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., and Brenner, S. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes* (2002) . *Science*. **297**, 1301-1310.

Avery, O.T., MacLeod, C.M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.* **98,** 451-460.

Bailey, J.A, Yavor, A.M., Viggiano, L., Misceo, D., Horvath, J.E., Archidiacono, N., Schwartz, S., Rocchi, M., and Eichler, E.E. (2002). Human-Specific Duplication and Mosaic Transcripts: The recent paralogous structure of chromosome 22. *Am. J. Hum*. Genet. **70**, 83-100.

Bailey, J.A., Church, D.M., Ventura, M., Rocchi, M., Eichler, E.E. (2004). Analysis of Segmental Duplications and Genome Assembly in the Mouse. *Genome Res.* **14**,789–801.

Barbazuk W. B., Ian Korf, Candy Kadavi, Joshua Heyen, Stephanie Tate, Edmund Wun, Joseph A. Bedell, John D. McPherson, and Stephen L. Johnson. (2000). The Syntenic Relationship of the Zebrafish and Human Genomes. *Genome Res.* **10**, 1351 – 1358.

Baxendale, S., Abdulla, S., Elgar, G., Buck, D., Berks, M., Micklem, G., Durbin, R., Bates, G., Brenner, S., and Beck, S. (1995). Comparative sequence analysis of the human and pufferfish Huntington's disease genes. *Nature Genet.* **10**, 67-76.

Best, R.G., Diamond, D., Crawford, E., Grass, F.S., Janish, C., Lear,T.L., Soenksen, D., Szalay, A.A., and Moore, C.M. (1998). Baboon/human homologies examined by special karyotyping (SKY): A visual comparison. *Cytogenet. Cell. Genet*. **82**, 83-87.

Benton, M.J. (1997). *Vertebrate Paleontology*. (Chapman & Hall, New York).

Berget, S.M., Moore, C., and Sharp, P. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Nat. Acad. Sci. USA* **74,** 3171-3175.

Boxem M and van den Heuvel S: 2002. *C.* elegans class B synthetic multivulva genes act in G1 regulation. *Current Biology* **12**: 906-911.

Brackett, B. G., Bousquet, D., Nice, M. L., Donawick, W. J., Evans, J. F. and Dressel, M. A. (1982). Normal development following in vitro fertilization in the cow. *Biol. Reprod*. **27**,147-158.

Bradley, A. (2002). Mining the mouse genome. *Nature* **420**, 512-514.

Bray, N., Dubchak, I., and Pachter, L. (2003). AVID, A Global Alignment Program. *Genome Res.* **13(1)**,97-102.

Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B., and Aparicio, S. (1993).
Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. *Nature* **366**, 265-268.

Broman, K. W., Murray, J. C., ShefÆeld, V. C., White, R. L. & Weber, J. L. (1998). Comprehensive human genetic maps, individual and sex-speciÆc variation in recombination. *Am. J. Hum. Genet*. **63**, 861-869.

Brown, W.R.A. (1989). Molecular cloning of human telomeres in yeast. *Nature* **338**, 774-776.

Brownlie, A., Donovan, A., Pratt, S.J., Paw, B.H., Oates, A.C., Brugnara, C., Witkowska, H.E., Sassa, S., and Zon, L.I. (1998). *Nat. Genet.* **20**, 244–250.

Burge C, Karlin S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. **268**, 78-94.

Butler, B. (2000). Wellcome Trust funds bid to unravel zebrafish genome *Nature* **408**, 503-511.

Caccone, A. and Powell, J.R. (1989). DNA divergence among hominoids. *Evolution* **43,** 925-942.

Carey, K.D., and Rice, K.S (1996) The aged female baboon as a model of menopause. Paper presented at the annual meeting of the American Society of Primatologists, Madison, August, 11-16.

Chen, F-C., and Li, W-H. (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**, 444-456.

Chou, H. H., Takematsu, H., Diaz, S., Iber, J., Nickerson, E., Wright, K. L., Muchmore, E. A., Nelson, D. L., Warren, S. T., and Varki, A. (1998). *Proc. Natl. Acad. Sci. U. S. A*. **95,** 11751–11756.

Collins, J. E., Goward, M.E., Cole, C. G., Smink, L. J., Huckle, E. J., Knowles, S., Bye, J. M., Beare, D. M. and Dunham, I. Reevaluating Human Gene Annotation: A Second-Generation Analysis of Chromosome 22. *Genome Res.* **13**, 27 – 36.

Collip, J. B. (1925). The extraction of parathyroid hormone which will prevent or control parathyroid tetany and which regulates the level of blood calcium. *J. Biol. Chem*. **63**,395-438.

Copeland, N. G., Jenkins, N. A. & Court, D. L. (2001). Recombineering a powerful new tool for mouse functional genomics. *Nature Rev. Genet*. **2**, 769–779.

Cox, L.A., Birnbaum, S. and VandeBerg J.S. (2002). Identification of candidate genes regulating HDL cholesterol using a chromosomal region expression array. *Genome Res*. **12**, 1693-1702.

Crick, F.H.C., Barnett, L., Brenner, S., and Watts-Tobin, R.J. (1961). General nature of the genetic code for proteins. *Nature* **192,** 1227-1232.

Crick, F.H.C. (1966). Codon anti-codon pairing, the wobble hypothesis. *J. Mol. Biol.* **19,** 548-555.

Deiss, L. P., Feinstein, E., Berissi, H., Cohen, O., and Kimchi, A. (1995) Identification of a novel serine/threonine kinase and a novel 15-kD protein as potential mediators of the gamma interferon-induced cell death. *Gene Dev.* **9,** 15–30.

Dib, C. et al. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152-154.

Dunham, I., Shimizu, N., Roe, B.A., et al. (1999). The DNA Sequence of human chromosome 22. *Nature* **402**, 489-495.

Davis, M., Malcolm, S. and Rabbitts, T.H. (1984) Chromosome translocation can occur on either side of the c-*myc*oncogenein Burkitt lymphoma cells. *Nature*, **308**, 286-288

DiGeorge, A. (1965) A new concept of the cellular basis of immunity. *J. Pediatr.*, **67**, 907.

Driever, W., L. Solnica-Krezel, A.F. Schier, S.C.F. Neuhauss, J. Malicki, D.L. Stemple, D.Y.R. Stainier, F. Zwartkruis, S. Abdelilah, Z. Rangini et al. (1996). A genetic screen for mutations affecting embryogenesis in zebrafish. *Development* **123**, 37–46.

Donovan, A., A. Brownlie, Y. Zhou, J. Shepard, S.J. Pratt, J. Moynihan, B.H. Paw, A. Drejer, B. Barut, A. Zapata et al. (2000). Positional cloning of zebrafish *ferroportin1* identifies a conserved vertebrate iron exporter. *Nature* **403**, 776–781.

Edwards, J. H. (1994). Comparative genome mapping in mammals. *Curr. Opin. Genet. Devel.* **4,** 861-867.

Eggen, A. and Fries, R. (1995). An integrated cytogenetic and meiotic map of the bovine genome. *Anim. Genet.* **4,** 215–236.

Ermert, K., Mitlohner, H., Schempp, W., and Zachau, H. G. (1995).The immunoglobulin kappa locus of primates. *Genomics* **25,** 623–629.

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

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

Edelmann, L., Pandita, R.K., Spiteri, E., Funke, B., Goldberg, R., Palanisamy, N., Chaganti, R.S.K., Magenis, E., Shprintzen, R.J., and Morrow, B.E. (1999). A common molecular basis for rearrangement disorders on chromosome 22q11. *Hum. Mol. Gen.* **8**, 1157-1167.

Emanuel, B.S., Selden, J.R., Wang, E., Nowell, P.C. and Croce, C.M. (1984) *In situ*hybridization and translocation breakpoint mapping. I. Nonidentical 22q11 breakpoints for the t(9;22) of CML and the t(8;22) of Burkitt lymphoma. *Cytogenet. Cell Genet.*, **38**, 127-131.

Eppig, J. T. (1996). Comparative maps, adding pieces to the mammalian jigsaw puzzle. *Curr. Opin. Genet. Devel*. **6,** 723-730.

Field, K.G., Olsen, G.J., Lane, D.J., Giovannoni, S.J., Ghiselin, M.T., Raff, E.C., Pace, N.R., and Raff, R.A. (1988). Molecular phylogeny of the animal kingdom. Science **239,** 748-753.

Finch, J.T., et al. (1977). Structure of nucleosome core particles of chromatin. *Nature* **269,** 29-36.

Fisher, R.A. (1935) The sheltering of lethals. *Am. Nat*. **69**, 446-455.

Force, A et al. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*. **151**,1531-1545.

Friedman, R. & Hugbes, A.L. (2001). *Mol. Biol. Evol*. **18,** 89-93

Frazer, K.A., Chen, X., Hinds, D.A., Pant, P.V.K., Patil, N., and Cox, D.R. (2003). Genomic DNA Insertions and Deletions Occur Frequently Between Humans and Nonhuman Primates. *Genome Res.* **13**,341–346

Fujiyama A, Watanabe H, Toyoda A, Taylor TD, Itoh T, Tsai SF, Park HS, Yaspo ML, Lehrach H, Chen Z, Fu G, Saitou N, Osoegawa K, de Jong PJ, Suto Y, Hattori M, Sakaki Y (2002). Construction and analysis of a human-chimpanzee comparative clone map. *Science* **295**,131-4.

Gale, M., Jr.; Blakely, C. M.; Hopkins, D. A.; Melville, M. W.; Wambach, M.; Romano, P. R.; Katze, M. G. (1998). Regulation of interferon-induced protein kinase PKR: modulation of P58(IPK) inhibitory function by a novel protein, P52(rIPK). *Molec. Cell. Biol.* **18**: 859-871.

Gates, M.A., L. Kim, E.S. Egan, T. Cardozo, H.I. Sirotkin, S.T.Dougan, D. Laskari, R. Abagyan, A.F. Schier, and W.S. Talbot.(1999). A genetic linkage map for zebrafish, Comparativeanalysis of genes and expressed sequences. *Genome Res.* **9**, 334–347.

Gagneux, P. and Varki, A. (2001). Genetic differences between humans and great apes. *Mol. Phylogenet.Evol.,* **18,** 2-13.

Gellner, K., and Brenner, S. (1999). Analysis of 148 kb of Genomic DNA Around the *wnt1* Locus of *Fugu rubripes. Genome Res.* **9**, 251-258.

Gibbs, R., Weinstock, G., Kappes, S., Schook, L., Skow, L., and Womack, J. (2002). Bovine Genomic Sequencing Initiative, Cattle-izing the Human Genome. *A White Paper for Bovine Genome Sequence.*

Glazko, G.V., Nei, M. (2003). Estimation of divervence times for major lineages of primate species. *Mol. Biol. Evol.* **20**, 424-424.

Godsave, S., Dekker, E.J Holling, T., Pannese, M., Boncinelli, E and Durston, A. (1994) Expression patterns of Hoxb genes in the Xenopus embryo suggest roles in anteroposterior specification of the hindbrain and in dorsoventral patterning of the mesoderm. *Dev. Biol*. **166**, 465-476.

Gordon, D., Abajian, C., and Green, P. (1999). Consed, A graphical tool for sequence finishing. *Genome Res*. **8**, 195-202.

Gothel S.F., and Marahiel M.A. (1998). Peptidyl-prolyl cis-trans isomerases, a superfamily of ubiquitous folding catalysts. *Cell. Mol. Life Sci.* **55,** 423–436.

Goodman, M. (1999). The genomic record of humankind's evolutionary roots. *Am. J. Hum. Genet.* **64**, 31-39.

Graves, J.A.M., Wakefield, M.J., Peters, J., Searle, A.J., Archibald, A., O'Brien, S.J., Womack, J.E. (1995) Report of the Committee on Comparative Gene Mapping. In: Cuticchia, A.J., Chipperfield, M.A., Foster, P.A (comps) Human Gene Mapping. Johns Hopkins University Press Baltimore 1351-1408.

Green, P. and Ewing B. (copyright (1993)-(1996) Phred documentation.

Green, P. (copyright (1994)-(1996) Phrap documentation.

Groves, C.P. (1997). Taxanomy and phylogeny of primates. *Molecular Biology and Evolutionof Blood Group and MHC Antigens in Primates,* pp. 3-23, Springer-Verlag Publishers, Berlin.

Haffter, P., M. Granato, M. Brand, M.C. Mullins, M. Hammerschmidt, D.A. Kane, J. Odenthal, F.J.M. van Eeden, Y.-J. Jiang, C.-P. Heisenberg et al. (1996). The identification of genes with unique and essential functions in the development of the zebrafish, *Danio rerio. Development* **123**, 1–36.

Haldane, J.B.S. (1933) The part played by recurrent mutation in evolution. *Am. Nat*. **67**, 5-9.

Hardison, R.C. (2000). Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet. 16, 369-372.

Harrison, P.M, Hegyi H., Balasubramanian S., LuscomMbe N.M., Bertone P., Echols N., Johnson T. (2002) Gerstein Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* 2002 **12(2)**, 272-80.

Harrison, P.M., Zheng, D., Zhang Z., Carriero, N., and Gerstein, M. (2005). Transcribed processesed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nuc. A. Res.* **33,** 2374-2383.

Higley, J.D., Thompson, W.W., Champoux, M., Goldman, D., Hasert, M.F., Kraemer, G.W., Scalon, J.M. et al. (1993) Paternal and maternal genetic and environmental contributions to cerebrospinal fluid monoamine metabolites in rhesus monkeys. *Arch Gen Psychiatry*. **50**, 615-623.

Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., Yagami, K., Wynshaw-Boris, A. and Yoshiki, A. (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature,* **423,** 91–96.

Hogan, B., Beddington, R., Costantini, F. & Lacy, E. (1994). *Manipulating the Mouse Embryo, A Laboratory Manual,* pp. Cold Spring Harbor Laboratory Press, New York.

Holland, P.W.H., Garcia-Fernandez, J., Williams, J.A. & Sidow, A. (1994). Gene duplicationsand the origins of vertebrate development. *Development Suppl. 125-133*

Hughes A.L., da Silva, J., Friedman, R. (2001) Ancient genome duplications did not structure the human Hox-bearing chromosomes. *Genome Res*. **5,** 771-780.

Hubbard, T., Andrews, D., Caccamo, M., et al. (2005) Ensembl 2005. *Nucleic Acids Res.* **33,** 447-453.

HGPI (Human Genome Project Information). www.doegenomes.org

IHGSC (International Human Genome Sequencing Consortium). (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Ijdo, J.W., Baldini, A., Ward, D.C., Reeders, S.T., and Wells, R.A. (1991). Origin of human chromosome 2: An ancestral telomere-telomere fusion. *Proc. Natl. Acad. Sci. USA*, **88**, 9051-9055.

Irie, A., Koyama, S., Kozutsumi, Y., Kawasaki, T., and Suzuki, A. (1998). The Molecular Basis for the Absence of *N*-Glycolylneuraminic Acid in Humans *J. Biol. Chem.* **273,** 15866 – 15871.

Iritani, a. and Niwa, K. (1977). Capacitation of bull spermatozoa and fertilization in vitro of cattle follicular oocytes matured in culture. *J. Reprod. Fertil*. **50**, 119-121.

Jayashankar, L., Brasky, M.K., Ward, A.J., and Attanasio, R. (2003) Lymphocyte modulation in a baboon model of immunosenescence. *Clinical and Diagnostic Laboratory Immunology*, **10**, 870-875.

Jerome, C.P, Kimmel, D.B., McAlister, J.A., Weaver, D.S. (1986) Effects of ovariectomy on iliac trabecular bone in baboons. *Cacif Tissue Int*. **39**, 206-208.

Johnson, L. A., Flook, J. P., Look, M. V. and Pikel, D. (1987). Flow sorting of X- and Y-bearing spermatozoa into populations. *Gamete Res*. **16**, 1-9.

Joyner, A. L. (1999) *Gene Targeting, A Practical Approach.* Oxford Univ. Press, New York.

Kaplan, J.R, Fontenot, M.B., Berard. J., Manuck, S.B., Mann, J.J. (1995) Delayed dispersal and elevated monoaminergic activity in free-ranging rhesus monkeys. Am J Primatol. **35**, 229-234.

Kaessmann H., Wiebe V., Weiss G., Paabo S. (2001). Great ape DNA sequences reveal a reduced diversity and and expansion in humans. *Nat Genet* **27**,155-6.

Kammerer, C.M., Sparks, M.L., and Rogers, J. (1995) Effects of age, sex and heredity on measures of bone mass in baboons. *J Med Primatol*. **24**, 236-242.

Kawasaki K, Minoshima S, Shimizu N. (2000). Propagation and maintenance of the 119 human immunoglobulin Vlambda genes and pseudogenes during evolution. *J. Exp. Zool.* **288(2):**120-34.

Kent, W.J. (2002). BLAT—The BLAST-like alignment tool. *Genome Res.* **12,** 656–664.

Kimmel, C.B. (1989). Genetics and early development of zebrafish. *Trends Genet.* **5**, 283–288.

Klein, J., and Takahata, N. (2002). Where do we come from? The molecular evidence for human descent. Publishers, Springer-Verlag Berlin Heidelberg New York.

Koop, B.F., Tagle, D.A., Goodman, M., and Slightom, J.L. (1989). A molecular view of primate phylogeny and important systematic and evolutionary questions. *Mol. Biol. Evol.* **6,** 580–612.

Kornberg, R.D. (1974). Chromatin structure, a repeating unit of histones and DNA. *Science* **184,** 868-871.

Kornberg, R.D., and Lorch, Y., (1992). Chromatin structure and transcription. *Ann. Rev. Cell. Biol.* **8**, 563-587.

Korneev, S.A., Park,J.H. and O'Shea, M. (1999). Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNAtranscribed from an NOS pseudogene. *J. Neurosci.* **19,** 7711–7720.

Kruglyak, S., Durrett, R. T., Schug,M. D. & Aquadro, C. F. (1998). Equilibrium distribution of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc.Natl Acad. Sci. USA* **95**, 10774-10778.

Kumar, S., and Hedges, B. (1998). A molecular timescale for vertebrate evolution. *Nature* **329**, 917-920.

Kumar, S., Tamura, K., and, Nei, M. (2004). MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Briefings in Bioinformatics* **5**:150-163.

Lee, C., et al. (1997). Human centromeric DNAs. *Hum Genet.* **100**, 291-304.

Li W.H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol* **36**, 96–99.

Liu G, Zhao S, Bailey JA, Sahinalp SC, Alkan C, Tuzun E, Green ED, Eichler EE. (2003). Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* **13**,358-68.

Luke, S., and Verma, R.S. (1993). Telomeric repeat [TTAGGG]n sequences of human chromosomes are conserved in chimpanzee (*Pan troglodytes). Mol. Gen. Genet.* **237**, 460-462.

Lynch, M & Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics*. **154**, 459-473.

Malik, H. S., Henikoff, S. and Eickbush, T. H. (2000) Poised for contagion, evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res.* **10**, 1307-1318.

Martin, R.D. (1993). Primate origins, plugging the gaps. *Nature* **363,** 223-233.

Martin, A. (2001). Is tetralogy true? Lack of support for the "one-to-four rule". *Molecular Biology and Evolution* **18,** 89-93.

Martin, J.L., Mahaney, C.M., Bronikowski, M.A., Carey, D.K., Dyke, B., Comuzzie, G.A. (2002) Lifespan in captive baboons is heritable. Mechanisms of Ageing and *Development*. **123**, 1461-1467.

Maxam, A. M., and Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Nat. Acad. Sci.* **74**, 560-564.

Mayor C., Brudno M., Schwartz J. R., Poliakov A., Rubin E. M., Frazer K. A., Pachter L. S. and Dubchak I. (2000) VISTA, Visualizing Global DNA Sequence Alignments of Arbitrary Length. *Bioinformatics*, **16**,1046-1047.

McClintock, J.M., Kheirbek, M.A., and Prince, V.E. (2002) Knockdown of duplicated zebrafish hoxb1 genes reveals distinct roles in hindbrain patterning and a novel mechanism of duplicate gene retention. *Dev*. **129**, 2339-2354.

Mighell A.J., Smith, N.R., Robinson, P.A., Markham, A.F. (2000) Vertebrate pseudogenes. *FEBS Lett*. **468(2-3)**, 109-14.

Muller, S., and Wienberg, J. (2001). "Bar-coding" primate chromosomes, molecular cytogenetic screening for the ancestral hominoid karyotype. *Hum. Genet.* **109**, 85-94.

Myers, E.W. and Miller, W. (1988). Optimal alignments in linear space. *Comput. Appl. Biosci.* **4,** 11–17.

Mayor C., Brudno M., Schwartz J. R., Poliakov A., Rubin E. M., Frazer K. A., Pachter L. S. and Dubchak I. (2000) VISTA, Visualizing Global DNA Sequence Alignments of Arbitrary Length. *Bioinformatics*, **16**,1046-1047.

McGinnis, W and Krumlauf, R. (1992). Homeobox genes and axial patterning. *Cell*. **68**, 283-302.

Mouse Genome Sequencing Consortium (MGSC). (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562.

Muchmore, E. A., Diaz, S., and Varki, A. (1998). A Structural Difference Between the Cell Surfaces of Humans and the Great Apes *Am. J. Phys. Anthropol.* **107,** 187–198.

Mullis, K.B. and Faloona, F.A., "Specific synthesis of DNA *in vitro* via a polymerase catalyzed chain reaction", *Meth. Enzymol.* **155**, 335-350 (1987).

Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang, Y.P., Ryder, O.A., O'Brien, S,J. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature* **409,** 614-618.

Nadeau, J. H. et al. (1995). A rosetta stone of mammalian genetics. *Nature* **373,** 363-365.

Nickerson, E. and Nelson, D.L. (1998). Molecular definition of pericentric inversion breakpoints occurring during the evolution of humans and chimpanzees. *Genomics* **50,** 368–372.

Ning, Z., Cox, A.J., Mullikin, J.C. (2001). SSAHA, a fast search method for large DNA databases. *Genome Res*. **11**,1725-1729.

Nowell, P.C. and Hungerford, D.A. (1960) A minute chromosome in human chronic granulocytic leukemia. *Science*, **132**, 1497-1499.

O'Brien, S. J. et al. (1993). Anchored reference loci for comparative genome mapping in mammals. *Nature Genet*. **3,** 103-112.

O'Brien, S.J., J.F. Eisenberg, M. Miyamoto, S.B. Hedges, S. Kumar,D.E. Wilson, M. Menotti Raymond, W.J. Murphy, W.G.Nash et al. (1999). Genome maps 10. Comparative genomics. Mammalian radiations. Wall chart. *Science* **286**, 463–478.

Ohara O., Nagase,T., Ishikawa,K.-I., Nakajima,D., Ohira,M., Seki,N. and Nomura,N. (1997) Construction and characterization of human brain cDNA libraries suitable for analysis of cDNA clones encoding relatively large proteins. *DNA Res*. **4,** 53–59.

Ohno, S. (1970) Evolution by Gene Duplication, George Allen and Unwin Publishers, London.

Olson, M. V., Eichler, E. E., Varki, A., Myers, R. M., Erwin, J. M., and McConkey, E. H. (2002). A White Paper Advocating Complete Sequencing of the Genome of the Common Chimpanzee, *Pan Troglodytes.*

Oxtoby E, Jowett T. (1993) Cloning of the zebrafish krox-20 gene (krx-20) and its expression during hindbrain development. *Nucleic Acids Res*. **21(5):**1087-95.

Okada, N., Hamada, M., Ogiwara, I. & Ohshima, K. (1997). SINEs and LINEs share common 39 sequences, a review. *Gene* **205**, 229-243.

Page, S.L. and Goodman, M. (2001) Catarrhine Phylogeny: Noncoding DNA Evidence for a Diphyletic Origin of the Mangabeys and for a Human–Chimpanzee Clade. *Mol. Phy. Evo.* **18,** 14-25.

Paigen, K. (1995). A miracle enough, the power of mice. *Nature Med*. **1**, 215–220.

Pennacchio, L.A. and Rubin, E.M. (2001). Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* **2**, 100-109.

Phillips, P. H. (1939). Preservation of bull semen. *J. Biol. Chem*. **130**,415.

Polge, C., Smith, A. and Park, A. S. (1949). Revival of spermatozoa after vitrification and dehydration at low temperature. *Nature* **164**, 666.

Postlethwait, J.H. and W.S. Talbot. (1997). Zebrafish genomics, From mutants to genes. *Trends Genet.* **13,** 183–190.

Postlethwait, J.H., Woods, I.G., Ngo-Hazelett, P., Yan, Y., Kelly, P.D., Chu, F., Huang, H., Hill-Force, A., and Talbot, W.S. (2000). Zebrafish Comparative Genomics and the Origins of Vertebrate Chromosomes *Genome Res*., **10**, 1890-1902.

Pushkarsky T, Yurchenko V, Vanpouille C, Brichacek B, Vaisman I, Hatakeyama S, Nakayama KI, Sherry B, Bukrinsky MI. (2005) Cell Surface Expression of CD147/EMMPRIN Is Regulated by Cyclophilin 60*. *J Biol Chem.* **280(30)**, 27866-27871.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D. (1997) GeneCards: encyclopedia for genes, proteins and diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center (Rehovot, Israel). World Wide Web URL: http://www.genecards.org/

Reddy K.C. and Villeneuve A.M. (2004). *C. elegans* HIM-17 links chromatin modification and competence for initiation of meiotic recombination. *Cell* **118**, 439-452.

(RGSPC) Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521.

Rogers, J., and Hixson, E.J. (1997) Insights from model systems: Baboons as an animal model for genetic studies of common human disease. *Am. J. Hum. Genet*, **61**, 489-493.

Rogers, J., Mahaney, C.M., Witte, M.S., Nair, S., Newman, D., Wedel, S., Rodriguez, A.L., Rice, S.K., Slifer, H.S., Perelygin, A., Slifer, M, Palladino-Negro, P., Newman, T., Chambers, K., Joslyn, G., Parry, P., and Morin, A.P.(2000) A genetic linkage map of the baboon (Papio hamadryas) genome based on human microsatellite polymorphisms. *Genomics*, **67**, 237-247

Rogers, J and VandeBerg, L, J (2001) Gene Maps of Nonhuman Primates. *ILAR Journal*, **39**, 1-12.

Rossant, J. & McKerlie, C. (2001). Mouse-based phenogenomics for modelling human disease. *Trends Mol. Med*. **7**, 502–507

Robl, J. M., Prather, R. S., Branes, F., Eyestone, W., Northey, D., Gilligan, B. and First, N. L. (1987). Nuclear transplantation in bovine embryos. *J. Anim. Sci*. **64**,642-647.

Robledo R, Bender P, Leonard J, Zhu B, Osoegawa K, de Jong P, Xu X, Yao Z, Roe B. (2004). The immunoglobulin lambda variable light-chain region in primates has been shaped by multiple, independent, small-scale and large-scale insertion/deletion events. *Genomics.* **4**, 678-685.

Roe, B.A. *Protocols for Recombinant DNA isolation, cloning and sequencing*. (The University of Oklahoma, Norman, (1997) URL: http,//genome.ou.edu/proto.html

Ruvolo, M. (1997). Molecular phylogeny of the hominoids, inferences from multiple independent DNA sequence data set*s. Mol. Biol. Evol.* **14,** 248-265.

Roest, C. H. et al. (2000). Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat. Genet*. **25,** 235-238.

Roussigne, M., Cayrol, C., Clouaire, T., Amalric, F., and Girard, J. P. (2003). THAP1 is a nuclear proapoptotic factor that links prostate-apoptosis-response-4 (Par-4) to PML nuclear bodies. *Oncogene* **22**, 2432–2442

Salamov, A.A., and Solovyev, V.V. (2000). Ab initio gene finding in *Drosophila* genomic DNA. *Genome Research* **10**, 516-522.

Sanger, F., Thompson, E. O. P., and Katai, R. 1955. The amide groups of insulin. *Biochem J*. **59**,509-514.

Sanger, F. 1959. Chemistry of Insulin. *Science* **129**,1340-1344.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhihitors. *Proc. Nat. Acad. Sci.* **74**, 5463-5467.

Schinzel, A., Schmid, W., Fraccaro, M., Tiepolo, L., Zuffardi, O., Opitz, J.M., Lindsten, J., Zetterqvist, P., Enell, H., Baccichetti, C., Tenconi, R. and Pagon, R.A. (1981) The `cat eye syndrome', dicentric small marker chromosome probably derived from a no. 22 (tetrasomy 22pter>q11) associated with a characteristic phenotype.*Hum. Genet.*, **57**, 148-158.

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W.(2000). PipMaker—A Web Server for Aligning Two Genomic DNA Sequences *Genome Res*., **10**, 577–586.

Shi J, Xi H, Wang Y, Zhang C, Jiang Z, Zhang K, Shen Y, Jin L, Zhang K, Yuan W, Wang Y, Lin J, Hua Q, Wang F, Xu S, Ren S, Xu S, Zhao G, Chen Z, Jin L, Huang W. (2003). Divergence of the genes on human chromosome 21 between human and other hominoids and variation of substitution rates among transcription units. *Proc Natl Acad Sci U S A*. **100**, 8331-8336.

Shprintzen, R.J., Goldberg, R.B., Lewin, M.L., Sidoti, E.J., Berkman, M.D., Argamaso, R.V. and Young, D. (1978). A new syndrome involving cleft palate, cardiac anomalies, typical facies, and learning disabilities, velo-cardio-facial syndrome. *Cleft Palate J.*, **15**, 56-62.

Shiraishi, S., Yokoo, H., Kobayashi, H., Yanagita, T., Uezono, Y., Minami, S., Takasaki, M., and Wada, A. (2000) *Neurosci. Lett.* **293,** 211–215.

Sibbald, B. (2000). The unraveling of chromosome 22, start saying goodbye to medicine as you know it. *Canadian Medical Association Journal* **162,** 252-253.

Sidow, A. (1996). Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev*. **6,** 715–722.

Silver, L. M. (1995). Mouse Genetics, Concepts and Practice. Oxford Univ. Press, New York.

Smit, AFA & Green, P. RepeatMasker at http,//repeatmasker.org

Smit, A. F. (1996). The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev*. **6**, 743- 748.

Smith, T.F. and Waterman, M.S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197.

Solomon, K.S., Logsdon, J.M. Jr., and Fritz, A. (2003) Expression and Phylogenetic analyses of three zebrafish Fox1 class genes. Developmental *Dynamics.* **228**, 301-307.

Solovyev, V.V., Salamov, A.A, and Lawrence C.B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Research* **22**, 5156-5163.

Spring, J. (1997). Vertebrate evolution by interspecific hybridization- are we polypoid? *FEBS Lett.* **400**, 2-8.

Stahmann, M. A., Huebner, C. F., and Link, K. P. 1941. Studies of the hemorrhagic sweet clover disease. V. Identification and synthesis of the hemorrhagic agent. *J. Biol. Chem.* **138**, 513-527.

Stewart C-B, Disotell, TR. (1998) Primate evolution - In and out of Africa. *Current Biology* **8**, 582-587.

Thornton, J.W. (2001). Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proc Natl Acad. Sci. USA* **98**, 5671-5676.

Thompson J., Higgins D., Gibson T. (1994). CLUSTAL W, improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22**, 4673-4680.

Toth, G., Gaspari, Z. & Jurka, J. (2000). Microsatellites in different eukaryotic genomes, survey and analysis.*Genome Res*. **10,** 967-981.

Trower, M.K., Orton, S.M., Purvis, I.J., Sanseau, P., Riley, J., Christodoulou, C., Burt, D., See, C.G., Elgar, G., Sherrington, R., Rogaev, E.I., St George-Hyslop, P., Brenner, S., and Dykes, C.W. (1996). Conservation of synteny between the genome of the pufferfish (Fugu rubripes) and the region on human chromosome 14 (14q24.3) associated with familial Alzheimer disease (AD3 locus). *Proc. Natl. Acad. Sci. U.S.A.* **93**,1366-1369.

Vanin, E.F. (1985) Processed pseudogenes, characteristic and evolution. *Annu. Rev. Genet.* **19**, 253-272.

Varki, A. (2000). A chimpanzee genome project is a biomedical imperative. *Genome Res.* **10,** 1065-1070.

Venkatesh, B. & Brenner, S. (1997) *Gene* **187**, 211–215.

Venter, J.C. et al. (2001). The sequence of the human genome. *Science* **291**, 1304-1351.

Wang, H., Q. Long, S.D. Marty, S. Sassa, and S. Lin. (1998). *Nat.* Genet. **20**, 239–243.

Warburton, P.E., et al. (1993). Nonrandom localization of recombination events in human alpha satellite repeat unit variants, implications for higher order structural characteristics within centromeric heterochromatin. *Mol. Cell. Biol.* **13**, 6520-6529.

Warburton, P.E., et al. (1996). Characterization of a chromosome-specific chimpanzee alpha satellite subset, evolutionary relationship to subsets on human ch romosomes. *Genomics* **33**, 220-228.

Watanabe H, Fujiyama A, Hattori M, Taylor TD, Toyoda A, Kuroki Y, Noguchi H, BenKahla A, Lehrach H, Sudbrak R, Kube M, Taenzer S, Galgoczy P, Platzer M, Scharfe M, Nordsiek G, Blocker H, Hellmann I, Khaitovich P, Paabo S, Reinhardt R, Zheng HJ, Zhang XL, Zhu GF, Wang BF, Fu G, Ren SX, Zhao GP, Chen Z, Lee YS, Cheong JE, Choi SH, Wu KM, Liu TT, Hsiao KJ, Tsai SF, Kim CG, OOta S, Kitano T, Kohara Y, Saitou N, Park HS, Wang SY, Yaspo ML, Sakaki Y. (2004). DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429**, 382-388.

Weinstein, B.M., D.L. Stemple, W. Driever, and M.C. Fishman. (1995). *gridlock*, a localized heritable vascular patterning defect in the zebrafish. *Nat. Med.* **1**, 1143–1147.

Wenskink, P. et al. (1974). A system for mapping DNA sequences in chromosomes of *D. melanogaster. Cell* **3,** 315-325.

Westerfield, M., (2000) Ed. 4. *The Zebrafish Book*. http// zfin.org /zf_info /zfbook

Wiltbank, J. N., Rothisberger, J. A. and Zimmerman, D. R. 1961. Effect of human chorionic gonadotropin on maintenance of the corpus luteum and embryo survival in the cow. *J. Anim. Sci*. **64**,642-647.

Womack, J.E. and Kata, S. (1995). Bovine genome mapping, Evolutionary inference and the power of comparative genomics. *Curr. Opin. Genet. Dev*. **5,** 725–733.

Yan, Y.-L., Gates, M.A., Horne, S., Amores, A., Brownlie, A., Donovan, A., Egan, E.S., Force, A., Gong, Z. et al.(1998). Vertebrate genome evolution and the zebrafish genemap. *Nat. Genet.* **18**, 345–349.

Yan, Y.-L., Miller, C.T., Nissen, R., Singer, A., Liu, D., Kirn, A., Draper, B., Willoughby, J., Morcos, P.A., Amsterdam, A., Chung, B.-c., Westerfield, M., Haffter, P., Hopkins, N., Kimmel, C., and Postlethwait, J.H. (2002) A zebrafish sox9 gene required for cartilage morphogenesis. *Development.* **129**, 5065-5079.

Yu, Y., Bradley, A. (2001). Engineering chromosomal rearrangements in mice. *Nature Rev. Genet*. **2**, 780–790.

Yunis, J.J. and Prakash, O. (1982). The origin of man, a chromosomal pictorial legacy. *Science* **215,** 1525–1530.

Zackai, E.H. and Emanuel, B.S. (1980) Site-specific reciprocal translocation t(11;22)(q23;q11), in several unrelated families with 3,1 meiotic disjunction. *Am. J. Med. Genet.* **7**, 507-521.

Zhong, T.P, Rosenburg, M., Mohideen, M.P.K., Weinstein, B. and Fishman, M.C. (2000). *Gridlock*, an HLH gene required for assembly of the aorta in zebrafish. *Science* **287**, 1820–1824.

Zhu O.S. Heisterkamp, N., Groffen, J. (1990) Unique organization of the human BCR gene promoter. *Nucleic Acids Res*. **18(23)** 7119- 7125.

Appendix

Sanger* Human Gene ID	Ensembl** Human Gene ID	Gene* Name	Category*	Homology* Description	Function**,#	Ensembl** Zebrafish Gene ID	Expression**,&,@
AC002472.11		AC002472.11	partial_gene	Matches ETSs			
AC002472.7	ENSG00000183773	AIFL	coding_gene	Similar to apotosis-inducing factor	Involved in apoptosis	ENSDARG00000002125	In developing zebrafish embryos. Brain and somites
AC002472.12		AC002472.12	Non coding gene	Matches ESTs			
AC002472.2	ENSG00000099949	LZTR1	coding_gene	leucine-zipper-like transcriptional regulator 1	Predicted to be dna binding transcriptional regulator	ENSDARG00000015905	Unspecific in developing zebrafish embryos
AC002472.8	ENSG00000184436	Thap7	coding_gene	Homo sapiens cDNA: FLJ20956 fis clone ADSE02049. alpha tubulin-like	Transcriptional repressor, involved in apoptosis	ENSDARG00000027585	In developing zebraifsh embryos. Brain.
AC002472.6		AC002472.6	*pseudogene*	Similar to Sw:P05214 mouse TUBULIN ALPHA-3ALPHA-7 CHAIN			
AC002472.1	ENSG00000099957	P2RXL1	coding_gene	Purinergic receptor P2X-like 1 orphan receptor. Antisense to SLC7A4 over 298 bp.	Receptor for ATP that acts as a ligand gated ion channel.		In human tissues. Predominantly in skeletal muscle
AC002472.5	ENSG00000099960	SLC7A4	coding_gene	Solute carrier family 7 (cationic amiN/A acid transporter y+ system) member 4	Involved in the transport of the cationic amino acid	ENSDARG00000026245	In developing zebrafish embryos. Brain, otic vesicle, pectoral fin, branchial arches.
AC002472.4	ENSG00000174164	P2RXL2	partial_gene	Similar to Sw:O15547 human P2X PURIN/ACEPTOR 6 (ATP RECEPTOR).			

AC002472.10		AC002472.10	*pseudogene*	Similar to TR:Q9Y627 HYPOTHETICAL 66.2 KDA PROTEIN.			
AC002472.3		AC002472.3	*pseudogene*	Similar to Sw:P05214 mouse TUBULIN ALPHA-3ALPHA-7 CHAIN			
AC002472.9	ENSG00000169674	AC002472.9	*pseudogene*	Similar to AK024408 Homo sapiens cDNA FLJ14346 fis			
AP000550.9	ENSG00000169675	AP000550.9	*pseudogene*	Similar to AL117485 POM-like			
AP000550.6	ENSG00000169676	AP000550.6	*pseudogene*	BCR related sequence			
AP000550.5	ENSG00000169677	AP000550.5	*pseudogene*	Similar to Tr:Q13116 human MEMBRANE PROTEIN-LIKE PROTEIN			
AP000550.1	ENSG00000169678	AP000550.1	*pseudogene*	Similar to Tr:O75461 Human E2F TRANSCRIPTIONAL REPRESSOR PROTEIN E2F-6			
AP000550.8	ENSG00000169679	AP000550.8	partial_gene	Matches ESTs - N/Avd LCR Gene			
AP000550.4	ENSG00000169680	AP000550.4	*pseudogene*	GGT related			
AP000550.2	ENSG00000169681	AP000550.2	*pseudogene*	gamma-glutamyltransferase			
AP000550.10	ENSG00000169682	AP000550.10	partial_gene	Part of Homo sapiens mRNA for KIAA0466 protein			
AP000552.1	ENSG00000169683	AP000552.1	*pseudogene*	Similar to Tr:O75461 Human E2F TRANSCRIPTIONAL REPRESSOR PROTEIN E2F-6			
AP000552.2	ENSG00000169684	AP000552.2	*pseudogene*	Similar to Tr:Q13116 human MEMBRANE PROTEIN-LIKE PROTEIN			

AP000552.8	ENSG00000169685	AP000552.8	*pseudogene*	Similar to AL117485 POM-like			
AP000552.3	ENSG00000169686	AP000552.3	*pseudogene*	BCR related sequence			
AP000552.4	ENSG00000169892	LOC391303	partial_gene	Matches ESTs - N/Avel LCR Gene	No known function	ENSDARG00000027240	In developing zebrafish embryos. Brain and retina
AP000552.7	ENSG00000169688	AP000552.7	*pseudogene*	Similar to AK025647 Homo sapiens cDNA: FLJ21994 fis			
AP000552.6	ENSG00000169689	AP000552.6	partial_gene	Similar to Em:X91348 DGCR5 in geN/Amic Em:AC000095			
AP000552.5	ENSG00000169690	AP000552.5	partial_gene	Similar to Em:AF029571 Homo sapiens peripheral benzodiazepine receptor interacting protein.			
AP000557.1	ENSG00000169635	HIC2	coding_gene	KIAA1020 protein	Transcriptional repressor	ENSDARG00000038298	Unspecific in developing zebrafish embryos. In human tissue. Brain, highest level in cerebellum, testis
AP000557.2		AP000557.2	coding_gene	N/Aval gene			
AP000557.3		AP000557.3	coding_gene	Similar to Em:AF012872 human phosphatidyliN/Asitol 4-kinase 230			
AP000557.4		AP000557.4	coding_gene	Similar to Em:AB002316 KIAA0612			
AP000553.1		UBE2L3	coding_gene	ubiquitin-conjugating enzyme E2L3			
AP000553.6	ENSG00000161179	LOC150223	coding_gene	Matching ESTs	No known function	ENSDARG00000002884	In developing zebrafish embryos. Otic vesicle, notochord, liver.

AP000553.7	ENSG00000161180	AP000553.7	partial_gene	Similar to SW:Q9D5J4 4930432J16RIK PROTEIN.			
AP000553.4	ENSG00000128228	SDF2L1	coding_gene	Stromal cell-derived factor 2-like 1	Function unclear	ENSDARG00000008998	In human tissue. High in testis, moderate in pancreas spleen, prostate small intestine, colon. Low in brain, skeletal muscle.
AP000553.5		AP000553.5	partial_gene	Matches EST cluster			
AP000553.2	ENSG00000100023	PPIL2	coding_gene	Peptidylprolyl isomerase (cyclophilin)-like 2	Catalyzes the cis-trans isomerization of proline imidic peptide bonds in oligopeptides	ENSDARG00000002016	In developing zebrafish embryos. Brain, retina, somites. In human tissue. High in thymus, pancreas and testis. Also in heart, placenta, lung, liver, skeletal muscle, kidney, spleen prostate ovary small intestine and colon.
AP000553.3	ENSG00000100027	YPEL1	coding_gene	Homo sapiens unkNAwn mRNA	Belong to the YIPPEE family based on sequence	ENSDARG00000035630	N/A
AP000555.1	ENSG00000100030	MAPK1	coding_gene	Mitogen-activated protein kinase 1	Phosphorylates microtubule associated protein-2, myelin basic protein and ELK1	ENSDARG00000027552	Unspecific in developing zebrafish embyos.
AP000555.2		AP000555.2	*pseudogene*	Similar to AB047842 Macaca fascicularis brain cDNA			

EM:D86995.1	ENSG00000100034	PPM1F	coding_gene	KIAA0015	Promotes apoptosis	ENSDARG00000005786	Unspecific in developing zebrafish embryos.
D87012.1	ENSG00000100038	TOP3B	coding_gene	Topoisomerase (DNA) III beta	Possess negatively supercoiled DNA relaxing activity	ENSDARG00000027586	Unspecific in developing zebrafish embryos. In human tissues. Isoform1: testis, heart, skeletal muscle. Isoform2: thymus, kidney, pancreas
AC000029.2	ENSG00000100218	RTDR1	coding_gene	Homo sapiens rhabdoid tumor deletion region protein 1 (RTDR1)	Function unclear	ENSDARG00000017983	Unspecific in developing zebrafish embryos.
AC000029.1		GNAZ	coding_gene	guanine nucleotide binding protein (G protein) alpha z polypeptide			
AC000102.1	ENSG00000100228	RAB36	coding_gene	RAB36 member RAS oncogene family	Oncogene	ENSDARG00000014058	Unspecific in developing zebrafish embryos.
U07000.1	ENSG00000186716	BCR	coding_gene	Active BCR-related gene	Function unclear	ENSDARG00000042474 ENSDARG00000028844	In developing zebrafish embryos. Brain, retina, peripheral neurons
U07000.2		FBXW3	*pseudogene*	Similar to Tr:Q9UKB7 F-BOX PROTEIN FBW3			
U07000.3		U07000.3	*pseudogene*	Similar to AL117485 POM-like			
AP000343.2		AP000343.2	*pseudogene*	Similar to Em:U14972 Human ribosomal protein S10 mRNA			
AP000344.1		AP000344.1	partial_gene	Similar to Tr:Q16859 Human CARBOXYLESTERASE			
AP000344.2		AP000344.2	coding_gene	Matching EST cluster			

AP000344.6		AP000344.6	partial_gene	Drawn using Incyte ESTs (Not all perfect match).			
AP000344.7		AP000344.7	partial_gene	Matches ESTs			
AP000344.3		AP000344.3	*pseudogene*	Similar to Tr:Q13312 human TXBP181			
AP000345.2		AP000345.2	Non_coding_gene	Matches ESTs			
AP000345.1		IGLL1	partial_gene	ImmuN/Aglobulin lambda-like polypeptide 1			
AP000346.8		AP000346.8	*pseudogene*	Similar to BC008986 Homo sapiens rhabdoid tumor deletion region protein 1			
AP000346.6c		AP000346.6	coding_gene	Homo sapiens mRNA			
AP000346.5		AP000346.5	partial_gene	Similar to Tr:O70122 Mus musculus SODIUM GLUCOSE COTRANSPORTER (FRAGMENT)			
AP000347.2		IGLL2	partial_gene	immuN/Aglobulin lambda-like polypeptide 2			
AP000346.1		ASLL	partial_gene	arginiN/Asuccinate lyase-like			
AP000346.2		AP000346.2	*pseudogene*	GGT related			
AP000347.3		AP000347.3	partial_gene	Similar to Sw:Q03385 mouse GUANINE NUCLEOTIDE DISSOCIATION STIMULATOR RALGDS FORM A			
AP000347.1		AP000347.1	*pseudogene*	Similar to BETA-GLUCURONIDASE PRECURSOR (Sw:P08236)			
AP000348.1		ZNF70	partial_gene	zinc finger protein 70 (Cos17)			
AP000348.2		VPREB3	coding_gene	Homo sapiens pre-B lymphocyte protein 3 (VPREB3) mRNA			
AP000348.3		AP000348.3	coding_gene	Matches EST cluster			

AP000348.4	ENSG00000138869	C22orf16	coding_gene	Similar to yeast Sw:Q03667 and C. elegens SW:Q09254	No known function	ENSDARG00000010717	In developing zebrafish embryos. Brain, retina, somites, myotomes.
AP000349.1	ENSG00000099953	MMP11	coding_gene	Matrix metalloproteinase 11 (stromelysin 3)	Breakdown of extracellular matrix	ENSDARG00000026293	In human tissues. Stromal cells of breast carcinoma
AP000349.2		SMARCB1	coding_gene	SWISNF related matrix associated actin dependent regulator of chromatin subfamily b member 1	Trancription factor	ENSDARG00000011594	N/A
AP000350.1		AP000350.1	coding_gene	Similar to Homo sapiens CGI-101 protein mRNA (AF151859).			
AP000350.2	ENSG00000133460	SLC2A11	coding_gene	Similar to glucose transporters SW:P22732	Glucose transporter	ENSDARG00000034501	In developing zerbafish embryos. Premigratory and migrating neural crest cells, cranium
AP000350.3		MIF	coding_gene	Macrophage migration inhibitory factor (glycosylation-inhibiting factor)	May be regulating fuction of macrophage	ENSDARG00000014445	Unspecific in developing zebrafish embryos. In human tissues. Sites of inflamation
AP000350.4		AP000350.4	*pseudogene*	Similar to Tr:Q14145 human KIAA0132 and Wp:CE05435 RING CANAL PROTEIN LIKE TR:Q1867			
AP000350.5		AP000350.5	partial_gene	GSTT3-1 similar to Glutathione S-transferases.			
AP000350.6	ENSG00000185487	AP000350.6	*pseudogene*	GSTT4-1 similar to Glutathione S-transferases.			
AP000350.7	ENSG00000133433	GSTT2	coding_gene	glutathione S-transferase theta 2	Function unclear	ENSDARG00000017388	N/A

AP000351.1		AP000351.1	partial_gene	DDT2 Similar to Sw:O35215 D-DOPACHROME TAUTOMERASE.			
AP000351.2	ENSG00000099977	DDT	coding_gene	D-dopachrome tautomerase			
AP000351.3		AP000351.3	coding_gene	GSTT2-2 similar to Glutathione S-transferases.			
AP000351.4		AP000351.4	partial_gene	GSTT4-2 similar to Glutathione S-transferases.			
AP000351.5	ENSG00000185831	AP000351.5	coding_gene	GSTT3-2 similar to Glutathione S-transferases.		ENSDARG00000009085	N/A
AP000351.6		AP000351.6	*pseudogene*	Similar to Tr:P03886 human NADH-UBIQUINVANE OXIDOREDUCTASE CHAIN 1			
AP000351.13		AP000351.13	*pseudogene*	Similar to Tr:Q13541 human 4E-BINDING PROTEIN 1.			
AP000351.7		AP000351.7	partial_gene	GSTT5-1 similar to Glutathione S-transferases.			
AP000351.8		AP000351.8	partial_gene	GSTT4-3 similar to Glutathione S-transferases.			
AP000351.9		AP000351.9	partial_gene	DDT4 Similar to Sw:O35215 D-DOPACHROME TAUTOMERASE.			
AP000351.10	ENSG00000099991	GSTT1	coding_gene	glutathione S-transferase theta 1		ENSDARG00000042428	N/A
AP000351.11		AP000351.11	partial_gene	GSTT3-3 similar to Glutathione S-transferases.			
AP000351.12		AP000351.12	*pseudogene*	GSTT5-2 similar to Glutathione S-transferases.			

AP000352.1	ENSG00000099991	CAB1	coding_gene	Homo sapiens calcineurin binding protein cabin 1 mRNA	It may serve as a negative regulator of T cell receptor signalling.	ENSDARG00000006098	Unspecific in developing zebrafish embryos. Widely expressed in different human tissues
AP000353.2		AP000353.2	partial_gene	Matches EST cluster			
AP000354.3	ENSG00000099998	GGTLA1	coding_gene	Gamma-glutamyltransferase-like activity 1		ENSDARG00000007929	N/A
AP000354.1		AP000354.1	*pseudogene*	gamma-glutamyltransferase 3			
AP000354.7	ENSG00000128262	AP000354.7	*pseudogene*	Similar to AL117485 POM-like			
AP000354.4		AP000354.4	*pseudogene*	BCR related sequence			
AP000354.6		AP000354.6	partial_gene	Similar to Sw:Q07254 frog 40S RIBOSOMAL PROTEIN S10			
AP000354.2	ENSG00000100014	AP000354.2	coding_gene	Em:HSAB2374 : Human mRNA for KIAA0376 gene partial cds.	No known function	ENSDARG00000006719	In developing zebrafish embryos. Brain, notochord, somites, pectoral fin, lateral line ganglia, proctodeum.
AP000355.1		ADORA2A	coding_gene	adenosine A2a receptor	Receptor for Adenosine, mediated by G proteins which activate adenyl cyclase	ENSDARG00000018790	Unspecific in developing zebrafish embryos.
AP000355.2	ENSG00000100024	UPB1	coding_gene	Matches EST cluster	3rd/final step in the catabolism of the pyrimidine bases	ENSDARG00000011521	Unspecific in developing zebrafish embryos.
AP000356.6	ENSG00000138867	AP000356.6	coding_gene	Matches EST cluster			

AP000356.7	ENSG00000100028	SNRPD3	coding_gene	Human SnRNP core protein Sm D3 mRNA	small nuclear ribonuclearprotein Sm D3	ENSDARG00000005825	Unspecific in developing zebrafish embryos.
AP000356.8	ENSG00000178026	Q8N0S9	coding_gene	Matches EST cluster			
AP000356.4	ENSG00000100031	GGT1	coding_gene	Gamma-glutamyltransferase 1	Gamma-glutamyl cycle, synthesis and degradation of glutathione	ENSDARG00000013504	In human tissue. Intestine and kidney
AP000356.10		AP000356.10	*pseudogene*	GGT related			
AP000356.5		BCRL6	*pseudogene*	Breakpoint cluster region-like 6			
AP000356.12		AP000356.12	*pseudogene*	Similar to AL117485 POM-like			
AP000356.9	ENSG00000180060	POM121L1	*pseudogene*	Similar to Tr:Q13116 human MEMBRANE PROTEIN-LIKE PROTEIN			
AP000357.3		AP000357.3	*pseudogene*	Similar to TR:Q9Y689 Homo sapiens ARF-family of Ras related GTPases			
AP000357.2		AP000357.2	*pseudogene*	Similar to AF006082 Homo sapiens actin-related protein Arp2			
AP000358.2		AP000358.2	*pseudogene*	Similar to AJ404615 Homo sapiens mRNA for fructosamine-3-kinase			
AP000358.1		AP000358.1	*pseudogene*	Similar to Human cysteine-rich heart protein (hCRHP) mRNA	Contains 1-Lim Zinc Binding domain		
AP000359.2		AP000359.2	partial_gene	Similar to Sw:O95404 human hiwi mRNA.	Contains PAZ domain, may be involved in RNAi		

AP000359.1		TOP1P2	*pseudogene*	Topoisomerase (DNA) I pseudogene 2			
dJ930L11.1	ENSG00000167037	dJ930L11.1	coding_gene	Similar to KIAA0397 Em:AB007857	No known function	ENSDARG00000028857	N/A
bA9F11.1		bA9F11.1	coding_gene	Matches ESTs			
bK221G9.4		bK221G9.4	coding_gene	Matches EST cluster			
bK243E7.3		bK243E7.3	partial_gene	Matches EST sequences			
bK221G9.1		bK221G9.1	*pseudogene*	Stathmin family SCG10-like SCLIP Neuroplasticin-2 NCP2 pseudogene			

* Information from Sanger Chromosome 22 genes and predicted genes: http://www.sanger.ac.uk/cgi-bin/hgp/chr22/gene_table (Collins et al. 2003)

** Information from Ensembl Genome browser: http://www.ensembl.org/ (Hubbard et al. 2005)

Information from GeneCards: http://www.genecards.org/ (Rebhan et al. 1997)

@ Zebrafish whole mount *in situ* hybridization data