UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

COMPARATIVE ANALYSIS OF HUMAN CHROMOSOME 22 CES-DGCR SYNTENIC REGIONS IN CHIMPANZEE, BABOON, BOVINE, MOUSE AND ZEBRAFISH AND EXPRESSION PROFILING IN ZEBRAFISH EARLY DEVELOPMENTAL STAGES USING WHOLE MOUNT IN SITU HYBRIDIZATION

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

SHELLY K. OOMMEN

Norman, Oklahoma

2006
COMPARATIVE ANALYSIS OF HUMAN CHROMOSOME 22 CES-DGCR SYNTENIC REGIONS IN CHIMPANZEE, BABOON, BOVINE, MOUSE AND ZEBRAFISH AND EXPRESSION PROFILING IN ZEBRAFISH EARLY DEVELOPMENTAL STAGES USING WHOLE MOUNT IN SITU HYBRIDIZATION

A Dissertation APPROVED FOR THE DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY

____________________________________  ____________________________________________
Bruce A. Roe                                                                                   

____________________________________  ____________________________________________
Paul Cook                                                                                       

____________________________________  ____________________________________________
Ann West                                                                                         

____________________________________  ____________________________________________
George Richter-Addo                                                                             

____________________________________  ____________________________________________
Tyrrell Conway
Acknowledgements

I thank God for all His mercies, guidance and motivation all through my life and helping me to achieve what I am thus far and always.

I would like to express my sincere gratitude and deepest appreciation to my major professor Dr Bruce Roe, whose intellectual and scientific guidance, patience and support has always inspired me to do the best and never limit myself in scientific thinking and thus enabling me to complete my research successfully. I would like to also thank my advisory committee, Drs Paul F Cook, Ann H West, G.B Richter-Addo and Tyrell Conway for their continued support, participation and encouragement during the research. I would like to express special thanks to Dr Han Wang for his guidance, training and collaboration with the zebrafish experiments.

I also would like to extend my thanks to all members of Dr. Roe’s laboratory, as each member has been helpful and supportive during my research and making the lab a great environment to be a part of. I would like to sincerely thank my colleagues, Christopher Lau, Jiangfeng Li, Hung-Chun (James) Yu for their team work and challenging constructive discussions while working as a team investigating comparative genomics and whole mount in situ experiments. I also thank our informatic team, Steve Kenton, Dr Axin Hua, Jim White and Hongshing Lai for their invaluable help and guidance and especially Dr.Fares Najar whose help and guidance has enabled to make graduate life easier. I would like to sincerely thank the former member of this lab Dr. Stephane Deschamps, for his help, guidance and training which laid the foundation for earlier work carried out during the research. I would
like to express my sincere thanks to the sequencing team, the primer making group
and Mounir and his group members for helping in carrying out the research goals
smoothly.

I would like to express my heart felt appreciation and sincere thanks to my
parents K Kunjummen Verghese and Girly Kunjummen for their priceless and
steadfast love, prayers, encouragement, and motivation, which has constantly helped
me to do the best and achieve goals. I would like to sincerely thank my elder sister,
Dr Susan and her husband, Mathew K George for their constant love, support, prayers
and encouragement and my second sister Shylu and her husband, Pradeep Abraham
for their constant love, prayers and encouragement and also to all my neices and
nephews for their love. I would like to express my sincere thanks to my parents in law
A.G Abraham and Chinnamma Abraham and family for their patience and prayers. I
would like to express my thanks to my cousin brother Binu C John and family for
their love and care and all my relatives, church family and friends both back home
and in U.S.A for their prayers, love and encouragement.

I would like to sincerely express my deepest heart felt love and appreciation to
my husband Dr.Arnold Abraham for his love, friendship, encouragement and
invaluable patience that has enabled me to finish my research successfully.
# Table of Contents

ACKNOWLEDGEMENTS ........................................................................................................ iv
TABLE OF CONTENTS ........................................................................................................ vi
LIST OF TABLES ................................................................................................................ vii
LIST OF FIGURES ............................................................................................................... viii
LIST OF ABBREVIATIONS .................................................................................................. x

ABSTRACT .......................................................................................................................... xiv

1. INTRODUCTION ............................................................................................................. 1
   1.1 DNA the Genetic Material ...................................................................................... 1
   1.1.1 Historical Background ................................................................................... 1
   1.1.2 DNA Structure ............................................................................................... 3
   1.2 From Structure to Function ................................................................................ 4
   1.3 DNA Sequencing Technology ........................................................................... 8
   1.4 Human Chromosome 22 Cat Eye and DiGeorge Syndrome Region .......... 9
      1.4.1 Low Copy Repeat Regions in Human Chromosome 22 ....................... 10
   1.5 Comparative Genomics .................................................................................... 13
      1.5.1 Significance of Comparison with Both Distantly Related and Closely Related Species ................................................................................................. 15
      1.5.2 Closely and Distantly Related Species Comparison ............................. 17
         1.5.2.1 Human and chimpanzee (6-8MYA) .................................................. 17
         1.5.2.2 Human and baboon (35-40 MYA) .................................................... 21
         1.5.2.3 Human and cow (85MYA) ............................................................ 23
         1.5.2.4 Human and mouse (60/80MYA) .................................................... 24
         1.5.3 Human and Zebrafish (450MYA) ......................................................... 25
            1.5.3.1 Zebrafish as the model system for whole mount in situ hybridization 25
            1.5.3.2 Global gene duplication in zebrafish (350-400MYA) ............... 27
            1.5.3.3 Mutation studies in zebrafish ....................................................... 30

2. MATERIALS AND METHODS ......................................................................................... 32
   2.1 Sequencing of the Chimp, Baboon, Cow, Mouse and Zebrafish Clones Syntenic to Human Chromosome 22 Cat Eye and DiGeorge Region ........................................... 32
      2.1.2 Random Shotgun Sequencing .................................................................... 33
      2.1.3 Analysis of the Sequences using Different Tools ..................................... 36
   2.2 Whole Mount In Situ Studies using Zebrafish as the Model Organism ........ 38
      2.2.1 Isolation of Total RNA .............................................................................. 38
      2.2.2 RT-PCR and RNA Probe Preparation by In Vitro Transcription ........... 39
      2.2.3 Single Stranded Exon Specific DNA Probes and Primer Design ........... 41
      2.2.4 Whole Mount In Situ Hybridization ....................................................... 43

3. RESULTS AND DISCUSSIONS ......................................................................................... 46
   3.1 Comparative Analysis of Human Chromosome 22 CES-DGCR Region .... 53
      3.1.1 Repetitive Elements ............................................................................... 53
      3.1.2 Locations of Repeat Element Insertions and Their Significance .......... 57
3.2 Significance of Nucleotide and Amino Acid Changes

3.2.1 Substitution Preference between First, Second, and Third Codon Position between Human, Chimpanzee, and Baboon

3.2.2 Transitional Changes versus Transversional Changes

3.2.3 Synonymous (Ks) versus Nonsynonymous (Ka) Changes

3.2.4 Codon Usage Preference between Human, Chimpanzee, and Baboon

3.2.5 Substitutions, Insertions and Deletions (indels): Hydrophobic vs Hydrophilic Amino Acid changes

3.3 Gene Expression Studies in Zebrafish

3.3.1 CES Region Gene Expression Studies

3.3.1.1 ENSG00000138860

3.3.1.2 Interleukin 17 receptor (IL17R)

3.3.1.3 Cat eye syndrome chromosome region, candidate 5 (CECR5)

3.3.1.4 Cat eye syndrome chromosome region, candidate 2 (CECR2)

3.3.1.5 BCL2-like 13 apoptosis facilitator (BCL2L13)

3.3.1.6 O94909/Kiaa0891

3.3.1.6.1 Expression of zf1

3.3.1.6.1 Expression of zf2

3.3.1.6.1 Expression of zf1 on (acerebellar) ace mutant embryo

3.3.1.7 Peroxisome biogenesis factor 26 (Q7Z2D7/PEX26)

3.3.2 DGCR Region Specific Gene Expressions

3.3.2.1 ENSG00000185643

3.3.2.2 Solute carrier family 25(mitochondrial carrier; citrate transporter), member 1) (SLC25A1)

3.3.2.3 Mitochondrial ribosomal protein L40 (MRPL40)

3.3.2.4 T box-1 (TBX1)

3.3.2.5 Armadillo repeat gene deletes in velocardiofacial syndrome (ARVCF)

3.3.2.6 DiGeorge syndrome critical region gene 8 (DGCR8)

3.3.2.7 HpaII tiny fragments locus 9C (Q8IZ69/HTF9C/ NM_022727)

3.3.2.8 Zinc finger, DHHC-type containing 8 (ZDHHC8)

3.3.2.9 ENSG00000182364

3.3.2.10 ENSG00000183099

3.3.2.11 Scavenger receptor class F, member 2 (SCARF2)

3.3.2.12 NM_032775(ENSG00000185214/Q96B68/NP_116164.2/KLH22 (Kelch-like 22))

3.3.2.13 Positive cofactor 2, multiprotein complex) glutamine/Q-rich-associated protein (PCQAP)

3.3.2.14 Phosphatidylinositol 4-kinase, catalytic, alpha polypeptide (PIK4CA)

3.3.2.15 v-crk sarcoma virus CT10 oncogene homolog (avian)-like gene (CRKL)

3.3.3 PHD finger protein 5A (Phf5a)

4. CONCLUSIONS

4. CONCLUSIONS
List of Tables

Chapter 1 ..............................................................................................................................1
Table 1.1 Major advances in molecular biology..............................................................5
Table 1.2 Examples of genes following the DDC model for the maintenance of the extra copies of duplication.................................................................29

Chapter 3 ..........................................................................................................................46
Table 3.1 The chimpanzee BAC clones from the rp43 and ptb libraries syntenic to human chromosome 22 sequenced and their associated marker genes. ..........47
Table 3.2 The baboon BAC clones from the rp41 library syntenic to human chromosome 22 sequenced and their associated marker genes. .................47
Table 3.3 The bovine BAC clones from the rp42 library syntenic to human chromosome 22 sequenced and their associated marker genes. .................48
Table 3.4 The mouse BAC clones from the rp21 library syntenic to human chromosome 22 sequenced and their associated marker genes. .................48
Table 3.5 BAC The zebrafish BAC clones from the ch211 library syntenic to human chromosomes 22, 5, 11 and 6 sequenced and their associated marker genes.....48
Table 3.6 List of genes seen in human, chimpanzee and baboon that contain repeat elements in their intronic, 5'UTR, 3'UTR and exonic regions...........................58
Table 3.7 (a,b,c,d,e,f) List of genes showing the insertions and deletions in human, chimpanzee and baboon and also showing the difference in conservation between different species in chimpanzee..........................................................76
Table 3.8 The table shows the major trends in amino acid substitutions (hydrophilic or hydrophobic) in different genes of human, chimpanzee and baboon............77
Table 3.9 TRANSFAC Web Based software used to identify transcription factor binding sites showing conservation of regulatory elements between human and zf1 and human and zf2..............................................................................92

Chapter 4 .......................................................................................................................126
Table 4.1 List of genes in the CES region with their expression pattern in zebrafish with reference to earlier studies in human and mouse.................................129
Table 4.2 List of genes grouped based on their expression in relation to the phenotypes noticed in CES syndrome. ...........................................................130
Table 4.3 List of genes in the DGCR region with their expression pattern in zebrafish. ..............................................................................................................133
Table 4.4 List of genes grouped based on their expression in relation to the phenotypes noticed in DGCR syndrome...............................................................134
# Table of Figures

**Chapter 3** ......................................................................................................................... 46

- Figure 3.1 Chimpanzee BACs Studied. ............................................................................ 46
- Figure 3.2 Comparison between the nucleotide sequences of the human chromosome 22 CES-DGCR region and syntenic region in chimpanze. ......................................................... 49
- Figure 3.3 Overview of PIP. .......................................................................................... 50
- Figure 3.4 The PIP output highlighting conservation ...................................................... 51
- Figure 3.5 VISTA output highlighting comparison between whole genome shotgun sequence and BAC based sequence. ................................................................. 52
- Figure 3.6 Major repeats ............................................................................................... 54
- Figure 3.7 Unique AluY repeats. .................................................................................. 55
- Figure 3.8 Insertion of SINE repeats in human SNAP29 gene .................................... 59
- Figure 3.9 PIP output of SNAP29 gene highlighting conservation of exons and deletion in chimpanzee. ....................................................................................... 60
- Figure 3.10 PIP output highlighting unique SINE insertion in human Pex26 gene. ....... 61
- Figure 3.11 PIP output of Pex26 gene highlighting conservation of exons and deletion in chimpanzee. ....................................................................................... 61
- Figure 3.12 First, second and third position nucleotide changes ...................................... 63
- Figure 3.13 Substitution rate between human, chimpanzee and baboon. ......................... 64
- Figure 3.14 Transitional vs Transversional changes ........................................................ 65
- Figure 3.15 Codon usage bias in amino acids with six codons. ........................................ 68
- Figure 3.16 Preference for G/C or A/T ending codons .................................................... 68
- Figure 3.17 Phylogenetic tree representation to interpret indels. ..................................... 70
- Figure 3.18 CES region of human chromosome 22 ........................................................ 79
- Figure 3.19 Expression pattern of the ENSG00000138860 gene .................................. 80
- Figure 3.20 Expression pattern of the IL17R gene. ......................................................... 81
- Figure 3.21 Expression pattern of the CECR5 gene. ....................................................... 83
- Figure 3.22 Expression pattern of the CECR2 gene. ....................................................... 85
- Figure 3.23 Expression pattern for the BCL2L13 gene. .................................................. 86
- Figure 3.24 zf1 and zf2 gene comparison to human Kiaa0819 gene .................................. 87
- Figure 3.25 Expression pattern of the zf1 gene. .............................................................. 88
- Figure 3.26 Expression pattern of the zf2 gene. .............................................................. 89
- Figure 3.27 Expression pattern of the zf1 gene in an ace mutant ..................................... 90
- Figure 3.28 Expression pattern of the Pex26 gene. ......................................................... 94
- Figure 3.29 DGCR region of human chromosome 22 ..................................................... 95
- Figure 3.30 Expression pattern of the ENSG00000185643 gene .................................... 96
- Figure 3.31 Expression pattern of the SLC25A1 gene ................................................... 98
- Figure 3.32 Expression pattern of the MRPL40 gene ..................................................... 100
- Figure 3.33 Expression pattern of the TBX1 gene .......................................................... 101
- Figure 3.34 Expression pattern of the ARVCF gene ....................................................... 103
- Figure 3.35 Expression pattern of the DGCR8 gene ...................................................... 105
- Figure 3.36 Expression pattern of the HTF9C gene ....................................................... 106
- Figure 3.37 Expression in the somites and developing myotomes ................................... 107
- Figure 3.38 Expression of the ZDHHC8 gene ............................................................... 108
Figure 3.39 Expression pattern of the ENSG00000182364 gene ................. 110
Figure 3.40 Expression pattern of the ENSG00000183099 gene .................... 112
Figure 3.41 Expression pattern of the SCARF2 gene .................................. 113
Figure 3.42 Expression pattern of the NM_032775/ENSG00000185214 gene ..... 115
Figure 3.43 Enlarged view of the eye photoreceptor cells ............................. 116
Figure 3.44 Expression pattern of the PCQAP gene ...................................... 117
Figure 3.45 Expression pattern of the PIK4CA gene .................................... 119
Figure 3.46 Expression pattern of the CRKL gene ........................................ 121
Figure 3.47 VISTA output of Phf5a gene ...................................................... 122
Figure 3.48 PIP output of Phf5a gene .......................................................... 123
Figure 3.49 Expression pattern of gene Phf5a .............................................. 124
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALU</td>
<td>a human DNA repeat element with homology to 7SL RNA</td>
</tr>
<tr>
<td>ARVCF</td>
<td>armadillo repeat gene deletes in velocardiofacial syndrome</td>
</tr>
<tr>
<td>B1</td>
<td>a family of SINEs found in rodents</td>
</tr>
<tr>
<td>BAC</td>
<td>a family of SINEs found in rodents</td>
</tr>
<tr>
<td>BCL2L13</td>
<td>BCL2-like 13 (apoptosis facilitator)</td>
</tr>
<tr>
<td>BCR</td>
<td>Breakpoint cluster region</td>
</tr>
<tr>
<td>BID</td>
<td>BH3 interacting domain death agonist</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>CDC45L</td>
<td>CDC45 cell division cycle 45-like (S. cerevisiae)</td>
</tr>
<tr>
<td>CDCREL-1</td>
<td>Cell division control related protein 1</td>
</tr>
<tr>
<td>CECR1</td>
<td>cat eye syndrome chromosome region, candidate 1</td>
</tr>
<tr>
<td>CECR2</td>
<td>cat eye syndrome chromosome region, candidate 2</td>
</tr>
<tr>
<td>CECR5</td>
<td>cat eye syndrome chromosome region, candidate 5</td>
</tr>
<tr>
<td>CECR6</td>
<td>cat eye syndrome chromosome region, candidate 6</td>
</tr>
<tr>
<td>CES</td>
<td>Cat Eye Syndrome</td>
</tr>
<tr>
<td>CLDN5</td>
<td>claudin 5 (transmembrane protein deleted in velocardiofacial syndrome)</td>
</tr>
<tr>
<td>COMT</td>
<td>catechol-O-methyltransferase</td>
</tr>
<tr>
<td>CRKL</td>
<td>v-crk sarcoma virus CT10 oncogene homolog (avian)-like</td>
</tr>
<tr>
<td>DGCR</td>
<td>DiGeorge syndrome Critical Region</td>
</tr>
<tr>
<td>DGCR14</td>
<td>DiGeorge syndrome critical region gene 14</td>
</tr>
<tr>
<td>DGCR2</td>
<td>DiGeorge syndrome critical region gene 2</td>
</tr>
<tr>
<td>DGCR6L</td>
<td>DiGeorge syndrome critical region gene 6 like</td>
</tr>
<tr>
<td>DGCR8</td>
<td>DiGeorge syndrome critical region gene 8</td>
</tr>
<tr>
<td>ENSG000</td>
<td>Ensembl Gene name representation</td>
</tr>
<tr>
<td>ERV1</td>
<td>Endogenous retrovirus</td>
</tr>
<tr>
<td>GNB1L</td>
<td>guanine nucleotide binding protein (G protein), beta polypeptide 1-like</td>
</tr>
<tr>
<td>GRB2</td>
<td>growth factor receptor-bound protein 2</td>
</tr>
<tr>
<td>HIRA</td>
<td>HIR histone cell cycle regulation defective homolog A (S. cerevisiae)</td>
</tr>
<tr>
<td>HTF9C</td>
<td>HpaII tiny fragments locus 9C</td>
</tr>
<tr>
<td>IL17R</td>
<td>interleukin 17 receptor</td>
</tr>
<tr>
<td>KCNMB3L</td>
<td>potassium large conductance calcium-activated channel, subfamily M, beta member 3-like</td>
</tr>
<tr>
<td>LDLa</td>
<td>Low Density Lipoprotein Receptor Class A</td>
</tr>
<tr>
<td>LINE</td>
<td>Long Interspersed repeat element, such as L1</td>
</tr>
<tr>
<td>LTR</td>
<td>long terminal repeats</td>
</tr>
<tr>
<td>MaLR</td>
<td>mammalian apparent LTR retrotransposons</td>
</tr>
<tr>
<td>MER</td>
<td>Medium Reiteration frequency Repeat</td>
</tr>
<tr>
<td>MICAL3</td>
<td>microtubule associated monoxygenase, calponin and LIM</td>
</tr>
<tr>
<td>Term</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>MIL1</td>
<td>Bcl-2-like 13 protein (MIL1 protein) (Bcl-rambo)</td>
</tr>
<tr>
<td>MIR</td>
<td>Mammalian-wide Interspersed Repeat</td>
</tr>
<tr>
<td>MLT1H</td>
<td>an old retrovirus-like LTR</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger ribonucleic acid</td>
</tr>
<tr>
<td>MRPL40</td>
<td>mitochondrial ribosomal protein L40</td>
</tr>
<tr>
<td>PCQAP</td>
<td>PC2 (positive cofactor 2, multiprotein complex) glutamine/Q-rich-associated protein</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PEX26</td>
<td>peroxisome biogenesis factor 26</td>
</tr>
<tr>
<td>Phred/Phrap</td>
<td>Phil’s read editor/ Phil’s read assembly program</td>
</tr>
<tr>
<td>PIK4CA</td>
<td>phosphatidylinositol 4-kinase, catalytic, alpha polypeptide</td>
</tr>
<tr>
<td>PRODHL</td>
<td>proline dehydrogenase (oxidase)-1 like</td>
</tr>
<tr>
<td>RANBP1</td>
<td>RAN binding protein 1</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>Rnase</td>
<td>ribonuclease</td>
</tr>
<tr>
<td>SCARF2</td>
<td>scavenger receptor class F, member 2</td>
</tr>
<tr>
<td>SERPIND1</td>
<td>serpin peptidase inhibitor, clade D (heparin cofactor), member1</td>
</tr>
<tr>
<td>SINE</td>
<td>Short Interspersed Element</td>
</tr>
<tr>
<td>SLC25A1</td>
<td>solute carrier family 25, member 1</td>
</tr>
<tr>
<td>SNAP29</td>
<td>synaptosomal-associated protein, 29kDa</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>STK22A</td>
<td>serine/threonine kinase 22A</td>
</tr>
<tr>
<td>TBX1</td>
<td>T-box 1</td>
</tr>
<tr>
<td>TOMM40</td>
<td>translocase of outer membrane 40 homolog (yeast)</td>
</tr>
<tr>
<td>TUBA8</td>
<td>tubulin, alpha 8</td>
</tr>
<tr>
<td>TXNRD2</td>
<td>thioredoxin reductase 2</td>
</tr>
<tr>
<td>UFD1L</td>
<td>ubiquitin fusion degradation 1-like</td>
</tr>
<tr>
<td>USP18</td>
<td>ubiquitin specific peptidase 18</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
<tr>
<td>X-gal</td>
<td>5-bromo-4-chloro-indolyl-b-D-galactosidase</td>
</tr>
<tr>
<td>ZDHHC8</td>
<td>zinc finger, DHHC-type containing 8</td>
</tr>
<tr>
<td>ZNF74</td>
<td>zinc finger protein 74</td>
</tr>
</tbody>
</table>
Abstract

Comparative genomic analysis is a powerful tool that can illuminate the genomic sequence features that result in the changes that drive evolution. In this dissertation, the 4.5 Mb region proximal to the centromere of human chromosome 22 that encodes the contiguous Cat Eye Syndrome and DiGeorge-Velocardiofacial Syndrome (CES-DGCR/VCFS) Critical Regions and the orthologous regions from chimpanzee, baboon, cow, mice and zebrafish have been sequenced and compared. Overall the human and chimpanzee sequences were ~98.5% identical and the human-baboon sequences were ~92% identical at the nucleotide level. A high degree of conservation was observed in both the gene order and the coding region sequences for these synteny regions, with a lower degree of conservation in the intronic and intergenic regions. The conserved structural features likely represent conserved functional properties while the observed differences must be responsible for portions of the human and primate specific phenotypes. The region studied was slightly larger in humans than in chimpanzees and baboons, since the human lineage had a higher insertion frequency relative to the other primates (or the other primates have a higher deletion frequency compared to humans). By comparing the sequenced regions of the chimpanzee genome from three different individual chimpanzees, Clint (ch251), Donald (rp43) and Gon (ptb), the first major conclusion from this dissertation research is that these three chimpanzees differ from each other by ~1.2%, almost as much as humans differ from chimpanzees.
The majority of the amino acid substitutions in humans, chimpanzees, baboons and bovines are changes from hydrophilic to hydrophilic amino acids. The observed human and chimpanzee substitution rate was 1.2% and that between humans and baboons was 2.6%, with Ka/Ks ratios for human and chimpanzee at 0.44 and human and baboon at 0.48. Thus, the second major conclusion from this present work is that at least in the case of humans vs primates, the genes are evolving by purifying selection.

The final series of experiments were based on the earlier observation that 16 genes in the human chromosome 22 CES-DGCR region had reported expression but no detailed expression profiles while 6 others had no known expression profiles. Through the comparative sequencing and subsequent whole mount in situ studies reported in this dissertation, expression of these 22 genes was observed to occur during zebrafish development, mainly during early developmental stages followed by either decreased or no expression in later stages in the brain, ear, eyes, heart, pharyngeal arches, liver, and kidney, all organs related to anomalies resulting in phenotype observed in CES-DGCR patients. Therefore, the third major conclusions from this present work is that contrary to prior studies pointing to single gene alterations resulting in these diseases, it now is clear that both CES and DGCR are multigene-based diseases.
1. Introduction

1.1 DNA the Genetic Material

1.1.1 Historical Background

The quest to find the relationship between structure and function in living things started as early as the 17th century with the discovery of the circulatory system by anatomist William Harvey. However, it was not until the late 1830s that Matthias Schleiden and Theodor Schwann proposed the cell theory. In the 1860s it was suggested that hereditary transmission is through sperm and egg, and in 1868, Haeckal postulated that the nucleus was important for heredity because the sperm consisted largely of nuclear material. In 1896, Edmund Beecher Wilson suggested that ‘nuclein’ was the substance of inheritance, and in 1874, Johanan Fredrich Miescher isolated nuclei from human pus cells and salmon sperms. In 1859, Charles Darwin’s publication of the Origin of Species marked the first step in describing how heredity operates in large populations of living things but it lacked an explanation of how traits were transmitted from generation to generation. In 1865, Mendel’s work with the garden pea (*Pisum sativum*) provided a theoretical model of inheritance, as he demonstrated that inherited characteristics are carried by discrete units or factors (now known as genes as coined by Johannsen) that are resorted in different ways in each generation. In the first decade of the 20th century, based on Mendel’s theory of inheritance and the cell biologist’s knowledge of chromosomes, the science of genetics was born. In 1902, Walter Sutton and Boveri independently observed that chromosomes and genes occur as pairs that separated in a similar fashion during gamete formation, and paired again during fertilization. They also suggested that genes are located on the chromosomes.
This chromosomal theory of inheritance postulated by Sutton in 1903 was refined during the period 1910 through 1940 by the classical genetic studies by Thomas Hunt Morgan and his students C.B. Bridges, H.J. Muller and A.H. Sturtevant. In 1910, Morgan’s study of the common fruit fly *Drosophila melanogaster* (Morgan, 1910) proved not only the chromosomal theory of inheritance but also helped in understanding the various aspects of meiosis. By the 1940s, the field of molecular genetics began to grow rapidly with the discovery and chromosomal location of several genes. However, little was known about the chemical makeup of the material that constitutes a gene or how genes functioned, although there was significant speculation prior to 1940. Here French geneticist L.H. Cuenot had speculated about gene function as early as 1903 when he proposed that due to the action of different genes, differences in coat color were inherited in mice. Then in 1909, as a result of his studies on the human disease alkaptonuria and other human diseases categorized as “inborn errors of metabolism”, Garrod proposed the relationship between gene and enzyme to perform a biochemical reaction. He speculated that the gene controlled the synthesis of the enzyme and a defective gene inherited at birth can disrupt the enzyme function (Garrod, 1908). Later, in 1928, Fred Griffith’s transformation experiment with smooth (S) and rough (R) strains of Pneumococcus showed that some “transforming principle” transferred from virulent dead S strain that converted the avirulent R strain to a virulent strain by synthesizing a new polysaccharide coat (Griffith, 1928). However, it was not until 1941 that G.W. Beadle and E.L. Tatum clearly showed the correlation between gene and enzyme with their work on X-ray induced mutants of the bread mold *Neurospora crassa* and proposed the “one gene-one enzyme” theory (Beadle and Tatum, 1941). In 1944, Avery, Macleod, and McCarty chemically characterized that the “transforming principle” was DNA through
a series of protease and nuclease digestion experiments (Avery et al., 1944) showing that DNA was the genetic material. DNA was confirmed to be the heredity material in 1952, when Hershey and Chase further solidified this concept by demonstrating that DNA is a component of T2 bacteriophage (Hershey and Chase, 1952).

1.1.2. DNA Structure

The biological information in a genome is encoded in the nucleotide sequence of its DNA or RNA molecule. The DNA (deoxyribonucleic acid) of an organism is the hereditary material of all living organisms that makes up its genome (with the exception that viruses have an RNA genome) and the primary unit of this genetic information is the genes. However, it was not until 1953, when Watson and Crick reported that DNA has a double helical structure consisting of two complementary polynucleotide chains that form a right handed helix in an antiparallel fashion, that the field of molecular biology truly began (Watson and Crick, 1953). This work was based on interpreting the x-ray diffraction data from DNA fibers obtained by M.H.F. Wilkins and Rosalind Franklin (Wilkins, 1951) and Erwin Chargaff’s observation in 1950 that the number of molecules of adenine equals to the number of thymine and the number of molecules of guanine equals to that of cytosine (Chargaff, 1951). It thus was established that polynucleotide chain consists of four nucleotides, two purines, namely adenine (A) and guanine (G) and two pyrimidines cytosine (C) and thymine (T), and the distance between the adjacent nucleotides is 3.4 Å (10 nucleotides per turn), with an A base pairing with T and G base pairing with C via 2 and 3 hydrogen bonds respectively. Each nucleotid in a DNA sequence consists of a pentose sugar (deoxyribose) with either a purine or pyrimidine attached at position 1 of the
sugar and a phosphate covalently linked to the sugar by asymmetrical 5’-3’ phosphodiester bonds. The direction of the phosphodiester bonds determines the character of the molecules, the sequence from 5’→3’ being different from the sequence 3’→5’. DNA sequence by convention is written in the order they are transcribed i.e. from the 5’ end (upstream direction) to the 3’ end (downstream direction) (Micklos et al., 1990; Cooper, 1994).

The DNA content in the haploid genome is a characteristic of each living organism. Genome size (Cavalier-Smith, 1985) is roughly correlated with the complexity and there is high variability not only in the DNA content but also in the number and the size of the chromosomes. There is a lack of correlation between total genome size and organism complexity; this phenomenon, termed the C-value paradox, occurs because of the presence of variable amount of repetitive DNA in different genomes (Gall, 1981, Gregory, 2001). Furthermore, with our present day knowledge of several eukaryotic genomic sequences, it is clear that these genomes mostly consist of noncoding DNA with repeated elements as the major component (Britten and Kohne, 1968).

1.2. From Structure to Function

In 1958, Francis Crick proposed the central dogma of Molecular Biology in which DNA and the subsequent protein sequences are colinear and the DNA directs its own replication as well as its transcription to RNA (RNA which specifies protein synthesis is mRNA) and RNA being translated to proteins. The more recent advances in our understanding of “central dogma”, shown in the table 1.1, both confirm and extend the hypothesis that genetic information flows from DNA to RNA to proteins.
| Charles Yanofsky and Seymour Benzer (1954) | Mutations in the genes of E.coli and T4 bacteriophage produced parallel changes in amino acid sequence |
| Zamecnik and Hoagland (1958) | Showed that tRNA was the adaptor molecule responsible to attach to amino acids during translation and order it on the template mRNA |
| Robert Holley (Holley et al., 1965). | Proposed the cloverleaf structure for tRNA |
| Benjamin Hall and Sol Spiegelman (1964) | Showed specific DNA sequences are transcribed into complementary mRNAs |
| S. Weiss and J. Hurwitz (1960) | Discovery of DNA dependant RNA polymerase |
| William (1977) | Showed binding of RNA polymerase to DNA in prokaryotes |
| Crick and Sydney Brenner (1961) | Triplet genetic code for each amino acid |
| Nirenberg and Khorana (1966) | Complete cracking of the genetic code |
| A. Claude (1930) and P. Zamecnik (1950) | Ribosomes as the site of protein synthesis |

**Table 1.1** Major advances in molecular biology

Transcription, the process of RNA synthesis, is catalyzed by the enzyme RNA polymerase. There are three eukaryotic RNA polymerases: RNA polymerase I, localized in the nucleolus and synthesizes rRNA (where rRNA is processed and assembled into ribosomes), RNA polymerase II, localized in the nucleoplasm and synthesizes heterogeneous nuclear RNA (hnRNA) the precursor for mRNA; and RNA polymerase III, localized in the nucleoplasm and synthesizes tRNAs, 5S rRNAs, 7SL RNA, U6 small nuclear (sn)RNA and other small stable RNAs involved in RNA processing (Hamsey and Reinberg, 1999; Paul and White, 2000). Transcription initiation by RNA polymerases requires several transcription factors that recognize specific consensus sequences located either upstream or downstream of the transcription start site. Hence there are specific consensus sequences in promoter region, such as the TATA or Hogness box, with a consensus sequence of TATAA that is located 25-30 base pair upstream from the transcription start site. The TATA box is recognized by TFIID and facilitates the correct positioning of RNA polymerase II for...
transcription initiation. Further upstream from the TATA at -75 is the CAAT box (GGNCAATCT) that plays a role in determining the efficiency of the promoter (Fickett and Hatzigeorgiou, 1997). An additional enhancer sequence found further upstream at -90 is the GC box (GGGCGG) that often is present in multiple copies in the promoter and can occur in either orientation (Werner, 1999). One or more of these GC enhancer elements are required for proper functioning of almost all promoters and are used to control gene transcription in both a temporal (developmental stage specific) and a spatial (tissue specific) framework. Two classes of transcription factors are recognized by the core and enhancer promoter elements. The general or basal transcription factors (general transcription factors: TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH) bind to the core promoter close to the transcription initiation site while gene specific activators (SP1, NF1, Oct-1) binds to the enhancer elements upstream of the transcription start site at varying distances (Wray et al., 2003). Together, these transcription factors and RNA pol II contribute to the assembly of a transcription preinitiation complex of transcription (Dermitzakis and Clark, 2002). The primary transcripts, termed heterologous nuclear RNAs, are formed as a result of transcription, but are not immediately functional in protein synthesis. The production of mature, functional mRNA requires RNA processing at the 5’ ends, and polyadenylation, i.e. the addition of approximately 20-250 adenine residues, at the 3’ end of the message. All eukaryotic mRNA have a three distinct domains, a 5’ untranslated region (5’UTR), a coding region (amino acid coding triplet codons) and a 3’ untranslated region (3’UTR). The 5’ and 3’ UTR of the mRNA play an important role in post-transcriptional regulation of gene expression as they control transport of mRNA from the nucleus to the cytoplasm,
translation efficiency, and the final subcellular localization and mRNA stability (Proudfoot, 2000).

Ribosomes are responsible for the translation of mRNA into proteins. The amino acid sequence of the final product is dictated by the nucleotide sequence of the mRNA whose codon are recognized by tRNA molecules with a specific amino acid esterified to the 3’ free hydroxyl group by a specific aminacyl-tRNA synthetase. The anticodon of the tRNA, base pairs with the codons of the mRNA to specify the amino acids to add to the growing peptide chain. Proteins are synthesized from the amino to carboxyl direction in three stages of synthesis namely initiation, elongation and termination. Initially the 40S subunit binds with the initiator aminoacyl tRNA and the initiation factors (IF) at the 5’ end cap of the mRNA and slides along till the start codon AUG is recognized and 60S subunit joins to form the 80S ribosome, that then reads the mRNA in the 5’—3’ direction. The Kozak sequence GCCRCCAUGG (Kozak, 1989) where R represents two purines that flank AUG, often is necessary for the recognition of the correct AUG initiation codon (Kozak, 1999). Elongation then occurs when the amino acids are incorporated successively into the growing polypeptide chain by the condensation reaction catalyzed by peptidyl-transferase activity (present as a part of large ribosomal RNA between the amino group of the incoming amino acid and the carboxyl group of the last amino acid incorporated), thus forming a peptide bond between successive residues. Translation continues until a termination codon (UAA, UAG, UGA) lacking a cognate tRNA is reached. This results in the binding of release factor followed by disruption of the multi-subunit ribosome structure, the removal of
the N-terminal methionine, and the release of the completed protein into the cytoplasm. (Li, 1997; Saccone and Pesole, 2003)

1.3. DNA Sequencing Technology

Within the past decade, DNA sequencing has been transformed into a convenient, routinely performed technique. The discovery of a large number of type II restriction endonucleases, initially provided the means to isolate small homogenous pieces of DNA for sequence analysis. Later improvements in polyacrylamide gel separation techniques (Sanger and Coulson, 1978) along with the discovery of DNA base specific chemical cleavage methods (Maxam and Gilbert, 1977) and improvements in DNA strand synthesis methods (Sanger et al., 1977), have contributed to the development of rapid DNA sequencing. Since the original Sanger sequencing method was introduced, advancements resulting in improvements in DNA sequencing (Mardis and Roe, 1989) sensitivity and efficiency that include new enzymes for DNA sequencing (Mead et al., 1991; Tabor et al., 1987; Stenesh and Roe, 1972), optimizing of the fragment set separation (Sanger and Coulson, 1978; Bankier and Barrell, 1989; Brown, 1984), fluorescence labeling detection methods and cycle sequencing (Smith et al., 1986; Gaxton, 1991) and improved shotgun sequencing strategy (Deininger et al., 1983; Bodenteich et al., 1994), automated instruments for large scale sequencing (McBride et al., 1989; Ansorge et al., 1987; Craxton, 1991) and improved sequence assembly and sequence alignment programs (Hyatt and Uberbacher, 2002; Galas and McCormack, 2002). These improvements have lead to increased yields and reproducibility in both the template preparations and sequencing reaction products and
the development of the high throughput DNA sequencing instruments. The resulting increased amount of sequence data now are catalogued in large genomic databases and annotated using both computer based and biological analysis approaches.

1.4. Human Chromosome 22 Cat Eye and DiGeorge Syndrome Region

The DNA is packaged into chromosomes and in the 23 human chromosomes, six are acrocentric (13, 14, 15, 21, 22, and Y), five are metacentric (1, 3, 16, 19 and 10), and the remaining are submetacentric. Human chromosome 22, one of the five human acrocentric chromosomes, is the second smallest and comprises 1.6-1.8% of the genomic DNA. The short arm of chromosome 22p contains multiple rRNA genes but lacks protein coding genes. In contrast, the long arm, 22q, is the gene rich euchromatic region (Dunham et al., 1999). Several congenital anomaly disorders (McDermid et al., 2002) are associated with the regions of 22q. These include DiGeorge and Velocardiofacial syndromes (VCFS), diseases that result from microdeletions that occur once in every 4000 births (Wilson et al., 1993). These microdeletions affect cardiovascular, neurological, psychiatric, endocrinological, immunological, palatal abnormalities and characteristic facial features. Over 90% of the patients with VCFS typically delete a 3Mb region, and 7% have a 1.5 Mb nested deletion with the proximal breakpoint region similar to that of the 3Mb deletion (Carlson et al., 1997). Cat Eye syndrome is caused as a result of an inverted duplication of the proximal 22q11 region leading to a bisatellited supernumerary chromosome and a partial tetrasomy that spans the p-arm and a part of the 22q11 (Footz et al., 2001; McDermid et al., 1986). CES is
characterized by congenital defects including anal atresia, ocular coloboma, preauricular tags/pits, heart and kidney defects, dysmorphic facial features and mental retardation. Derivative 22, or Der 22, is a genetic disease caused by a partial trisomy of both 11q23-qter and 22q11-qter that results in mental retardation and often multiple anomaly disorders (Fraccaro et al., 1980; Zackai et al., 1980). Low copy repeats (LCRs) are found in the vicinity of all three 22q11 breakpoint regions with their size ranging from 40-350kb and an identity of 97-98%. These LCRs differ in gene content and organization but often include 11 known genes or pseudogenes that have the potential for recombination leading to duplication or deletion of the above discussed chromosomal region (Shaikh et al., 2001; Edelmann et al., 1999).

1.4.1. Low Copy Repeat Regions in Human Chromosome 22

Chromosomal rearrangement (recurrent and nonrecurrent) breakpoints are found mainly in pericentromeric and subtelomeric regions, and often occur within intervals containing LCRs or AT rich palindromes or pericentromic repeats. LCRs constitute 5% of the human genome, 10-500kb in size with a sequence identity of >95% (Stankiewicz and Lupski, 2002) and are involved in meiotic non-allelic homologous recombination (NAHR), the mechanism that underlies genome rearrangements associated diseases. Thus, in most of the human genomic disorders such as Charcot Marie Tooth disease type 1A (Wise et al., 1993), neurofibromatosis type 1 (NF1) (Regnier et al., 1997), Williams-Beuren (WBS) (Robinson et al., 1996; Francke et al., 1999), Smith Magnesia (SMS) (Greenberg et al., 1991; Juyal et al., 1996), dup(17)(p11.2p11.2), Cat Eye and
DiGeorge syndromes, azoospermia and male infertility (as a result of Y chromosome deletion) (Repping et al., 2002), are caused by chromosomal deletions or duplications with LCRs acting as recombination hotspots.(Christine and Lupski, 2004) . Reciprocal deletions and duplications are the result of nonallelic homologous recombination (NAHR) in LCRs on the same chromosome in direct orientation, but inversions take place when a NAHR is between LCRs in inverted orientation. These reciprocal translocations often occur as result of NAHR between LCRs on different chromosomes as seen between chromosome 21q21.3-qter and recipients 13 and 18qter (Stankiewicz et al., 2003). Since the LCRs resulted from segmental duplications of specific gene rich regions of the genome that then were mutated to gene fragments and/or pseudogenes or repeat gene clusters, they recently have been termed as duplicons (Ji, Y et al., 2000).

The main biological outcomes of genomic duplications are that they provide the potential for genetic diseases based on spontaneous chromosomal rearrangements or they can alter gene evolution pathways by the interstitial deletions and reciprocal translocations that occur within them (Ji, Y et al., 2000).

For example the 3Mb typically deleted chromosome 22 DGCR region includes four copies (Edelmann et al., 1999) of chromosome 22 specific LCRs, namely LCR-A ~350kb, LCR-B ~135kb, LCR-C ~ LCR-D ~250kb (Shaikh et al., 2000; Shaikh et al., 2001; Spiteri et al., 2003). LCR-A and LCR-D are near or at the end points of 3Mb typically deleted in DGS/VCFS/CAFS syndromes (Emanuel et al., 1998) and CES duplications localize to especially LCR-A and also LCR-D (McTaggart et al., 1998).

Other LCRs also contain repetitive LINEs (long interspersed repetitive elements) and SINEs (short interspersed nuclear elements). The SINE family of
transposable elements include Alu elements that form 10% of the human genome (Chen et al., 2002; Jurka et al., 1993), usually is 280-300 bp long (Lander et al., 2001) and may be transcribed by RNA polymerase III (Duncan et al., 1979). Alu repeats are dimeric free left alu monomer and free right alu monomer fusions that originated from a primitive monomeric alu which is a DNA complementary to a 141bp internally deleted processed 7SL RNA. Alu sequences thus have two \((G + C)\) rich similar fragments, that are linked by a variable A-rich region and that terminates with a variable length 3’-poly(A) tail (Quentin, 1992; Jurka and Zuckerkandl, 1991; Ullu and Tschudi, 1984).

Alu elements occur mostly in the introns, often in the 3’UTR rather than the 5’UTR, but rarely occur in the gene coding region (Jun et al., 2004) (Chen et al., 2002). As a result, intron gain or loss is greatly influenced by Alu repeats and depending on the class of Alu element inserted, the insertion is either of recent or distant evolutionary origin. Chromosome 22q11 LCRs have Alu SINEs sequences at the junctions of genes (or pseudogenes) found within them (Babcock et al., 2003) and the younger subfamilies AluY and AluS are noticed near or within the junctions of segmental duplications throughout the human genome (Bailey et al., 2003). The oldest primate subfamily of alu sequence is AluJ and both LCRs are believed to have evolved 35 million years ago at the same time of the burst of primate Alu retroposon activity (Bailey et al., 2003) and segmental duplications during primate speciation that resulted in multiple LCRs duplications known (Ji, Y et al., 2000). It also has been shown that two pericentric inversion breakpoints in chimpanzee and (4;19) translocation in gorilla are localized in
the LCRs of the orthologous chromosomal regions (Stankiewicz et al., 2001; Locke et al, 2003).

Thus, there have been numerous studies aimed at understanding and describing the content and evolution of the LCR regions of human chromosome 22q11 that indicate that these LCR regions are predisposed to rearrangements resulting in congenital anomaly disorders, and that the expansion of these regions can cause both genome rearrangements and gene amplification. Orthologs of human chromosome 22 in non-human primates, i.e. chimpanzee, gorilla, and rhesus monkey also have duplications containing elements similar to the chromosome 22 specific LCRs. Studying the variation in both number and organization of these LCR regions between human and primates will help us understand the evolution and predisposition to duplication mediated rearrangements associated with human disease (Babcock et al., 2003; Shaikh et al., 2001, Shaikh et al., 2000; Shaw et al., 2004).

1.5. Comparative Genomics

With the discovery of the double helical structure of DNA, Watson and Crick anticipated that the arrangement of the order of the four DNA bases determined the genetic makeup of all living organisms. This gave rise to the field of molecular biology and saw a growing number of scientists interested in comparing nucleic acids and proteins of one species to another. Interestingly, studies in many different biology disciplines such as anatomy, biochemistry, pharmacology, immunology, and cell biology, are rooted in comparative biology (Nobrega and Pennachio, 2003). Genomics is a recent branch of biology that uses comparison to analyze individual genomic
sequences and provides information into both genomic structure and genome function (Ohtsuka et al., 2003). With the completion of the human genome sequence (Lander et al., 2001; Venter et al., 2001), efforts have been made to completely annotate and understand the coding regions (Batzoglou et al., 2000; Alexandersson et al., 2003; Korf et al., 2001), the functional noncoding (Loots et al., 2000; Stojanovic et al., 1999; Wasserman et al., 2000) and the quarter of the genome that contains noncoding intronic regions (Hare and Palumbi, 2003; Williams et al., 2003) as well as intergenic and regulatory regions (Gottgens et al., 2000; Williams et al., 2003; Kent and Zahler, 2000). Comparative analysis is useful for annotating evolutionarily conserved genomic regions because functionally important regions generally are well conserved across different species, and non-functional regions accumulate mutations or identify totally unknown genes that are specific to that species. Functional regions do not accumulate mutation as rapidly as non-functional regions because there is evolutionary pressure to maintain the important encoded biological function (Onyango et al., 2000). Comparative genomics is based on the hypothesis that genomes being compared have a common evolutionary ancestor. Therefore, cross-species comparison spanning wide evolutionarily distant species and closely related species can help distinguish functionally important sequences such as regulatory elements and coding sequences conserved over evolutionary time that have a shared ancestry (Loots et al., 2000; Thomas et al., 2003; Rijnkels et al., 2003; Dubchak et al., 2000).

1.5.1. Significance of Comparison with Both Distantly Related and Closely Related Species
Comparing the sequences of distantly related species genomes often can identify functional sequence elements if they have been conserved over evolutionary time. In contrast, comparative genomics of closely related species also provides unique genotype to phenotype correlations. However because of their high sequence similarity in orthologous regions over a relatively short time since they shared a common ancestor, there is very little sequence divergence and only small species specific changes will be observed. This makes it difficult to identify the functional elements (Boffelli et al., 2004; O’Brien et al., 1999; Nobrega et al., 2003; Thomas et al., 2002). Therefore, for a clear annotation of the human genomic sequence, comparisons with an increasingly wide variety of species between the range of extremely close (primates) will help identify the regions where divergence is tolerated by looking specifically for differences rather than similarities. In evolutionarily distant vertebrates and more distant invertebrate species similarities will help identify highly functionally conserved sequences. For example, the strictly functionally conserved sequences other than sequence conserved by chance were highlighted in studies of 12Mb genomic sequence from 12 species, in region orthologous to a 1.8MB region in human chromosome 7. This 1.8MB region includes the cystic fibrosis (CFTR) gene (Thomas et al., 2003), and comparative analysis of CFTR genes of human, cows and pigs helped to identify intronic regulatory elements that were not possible with the distant human-fugu comparison or the evolutionarily closer human-sheep comparison (Williams et al., 2003). Recent comparative genomic studies on analysis of stem cell leukemia SCL loci also indicates the benefits of extending comparison between the phylogenetic distances of the human comparison to chicken (Gottgens et al., 2002) as additional functional
enhancers easily could be identified. Earlier studies of the SCL locus in human and mice identified four new genes as well as several conserved regulatory elements (Gottgens et al., 2001). In contrast, a human–marsupial comparison was effective in identifying regulatory elements in genes paralogous to SCL and LYL1 (lymphoblastic leukemia derived sequence-1) (Chapman et al., 2003) while a human-fish comparison was beneficial in identifying cis-regulatory modules of Homeobox (HOX) genes (Aparicio et al., 1995), a human-chicken comparison helped identify the collagen XII gene promoter that further explained the regulation of the expression pattern of the extracellular matrix component (Chiquet et al., 1998). In addition, a recent pig-human-mouse comparison analyzed the porcine INS-IGF2-H19 gene cluster, an important Quantitative Trait Locus (QTL) primarily affecting the development of muscles (Amarger et al., 2002). Cross species sequence comparison helped analyze and identify the regulatory elements of interleukins 4, 13 and 5 (Loots et al., 2000). More recently, an alternative method but similar method for analysis of regulatory modules in closely related species, termed phylogenetic shadowing, was described (Boffelli et al., 2003).

1.5.2. Closely and Distantly Related Species Comparison

1.5.2.1 Human and chimpanzee (6-8MYA)

Human and chimpanzee are estimated to share a common ancestor approximately 6-8 MYA (millions years ago) (Chen and Li, 2001; Glazko and Nei, 2003). A comparison of the human and chimpanzee genomes show differences from our closest relative, help to understand human specific diseases (Olson and Varki,
2003), and provide population genetics information (Chen and Li, 2001; Ruvolo, 1997). In their 1975 study, King and Wilson reported that there is 99% similarity between average human and chimpanzee based on DNA hybridization experiments (King and Wilson, 1975). Although highly similar at the molecular level, humans and chimpanzees differ far more than two humans. Protein coding region changes can give rise to new genes by small insertions and deletions and nucleotide substitutions as well as by gene duplications, exon shuffling, retro transposition and gene fusions. The primate genome sequence will provide evidence about the sequences of these events that occurred during the time since humans and chimpanzee shared a common ancestor (Long et al., 2003). The availability of sequences for human and chimpanzee genomic sequences already have shown that small sequence changes occur within coding regions and that only 95% of the genomic chimpanzee DNA could be aligned to humans (Britten, 2002) in comparison to earlier estimates (Goodman, 1990). These newer difference estimates result mainly from insertions and deletions, i.e. indels, (3.4%), rather than from single base pair change (1.4%). For example, comparison of human (Major Histocompatibility Complex) MHC class I region to that of chimpanzee showed deletions of 95kb between human (major histocompatibility complex class I chain related (MIC-A and B functional transcripts)) MICA and MICB genes when compared to the single hybrid chimpanzee MIC gene (Anzai et al., 2003). Other recent studies have observed additional genomic variations that include insertion of human endogenous retrovirus K provirus (HERV-K) in chimpanzee, bonobo and gorilla, but not in human (Barbulescu et al., 2001) indicating that chimpanzee, bonobo and gorilla are closer to each other in evolutionary time than they are to humans.
Other differences have been observed in repetitive elements (Bailey et al., 2003; Nergadze, G.S et al., 2004, Hamdi, et al.,1999) and retroviral or transposon sequences (Sverdlov, 2000), as well as in multigene families such as the immunoglobulin and olfactory receptor gene families in human that are duplicated and chromosomally rearranged (Nei et al.,1997). Differences in cell surface of sialic acid (N-glycolyl-neuraminic acid (Neu5Gc) between human and chimpanzee also has been reported. It has been postulated that the changes in Neu5Gc observed in chimpanzee makes them less susceptible to pathogens or alters how it recognizes pathogens and cellular signals (Muchmore et al., 1998; Varki, 2000; Olson and Varki, 2003). Thus, the differences between human and chimpanzee can be seen in chromosomal organization, copy numbers, locations and functional status of a small number of individual genes within multigene families (Nei et al., 1997), a few altered coding regions that result in different gene products in both the species (Frazer et al., 2003; Hacia, 2001), differences in intron loss or gain (Cho et al., 2004; Coghlan et al.,2004; Babenko et al., 2004; Fedorov et al., 2003), differences in the number and distribution of interspersed repeats (Smit,1999; Kim et al., 1999) and different transposable elements in the human genome (Gagneux and Varki. 2001). Thus, although comparing the primate and human genomes should provide an understanding of the species specific characteristics based on gene structure, further gene expression studies will be needed to fully understand genome function (Karaman et al., 2003; Khaitovich et al., 2004). Additional studies into diseases that differ in frequency and severity between chimpanzee and human (AIDS, Alzheimer’s, cancer, malaria and reproductive disorders) and understanding the
genetic factors in these differences will be important in developing approaches to control and eventually cure these diseases (Varki, 2000).

Another aspect of human-chimpanzee diversity is variation in the position and sequence of repeats. Segmental duplications constitute about 5% of the human genome (Bailey et al., 2001) and either are intrachromosomal (seen in chromosome 15, 16, 17, 21 and 22) or interchromosomal duplications. In human, segmental duplications occur with at least two >90 identical copies of an ~100kb to 200kb per chromosome (Bailey et al., 2004). Segmental duplications also are hotspots for chromosomal rearrangements such as a chimpanzee specific inversion ((Locke et al., 2003), a human specific inversion of chromosome 18 (Goidts et al., 2004) and the human chromosome 2 fusion region (Fan et al., 2002). Segmental duplications have occurred several different times during primate evolution. For example, humans and chimpanzees have two copies of Charcot-Marie-Tooth neuropathy type 1A repeat sequence while gorilla has only one (Keller et al., 1999). Similarly human and African apes have 2 copies of creatine transporter SLC6A8 and the adrenoleukodystrophy gene at the Xq28 as well as the HS.135840 gene at 4q24 while only a single copy is present in orangutans and old world monkeys (Courseaux et al., 2001; Lupski et al., 1998, Eichler et al., 2001).

Information about nucleotide substitutions also has been obtained by sequence comparisons between closely related species (Webster et al., 2003; Hellmann et al., 2003a; Hellmann et al., 2003), and insertions and deletions (indels) are the commonly reported sequence differences between humans and chimpanzee (Watanabe et al., 2004). Because it has been reported that human chromosome 21 is 9% smaller than chimpanzee chromosome 22 (Frazer et al., 2003), it has been proposed that this
difference is due to insertion of mobile elements, deletion of regions by unequal recombination, or different mutation rates introduced during DNA replication (Ebersberger et al., 2002). Except for the human chromosome 21 and the chimpanzee chromosome 22 studies described above, much of the comparative primate genomic sequence data is based on relatively small sequenced regions.

The main differences between human and primates include single nucleotide replacement; insertions, deletions and larger duplications, and differing distribution and abundance of other repetitive elements and endogenous retroviruses. Through the studies described in this dissertation, the comparison of additional genomic regions will allow for a more in depth genomic comparison.

1.5.2.2 Human and baboon (35-40 MYA)

Human and baboon (old world monkey) have been estimated to have shared a common ancestor approximately 35-40 MYA (Stewart and Disotell, 1998; Goodman, M 1999). The sequence information of the genome of more distantly related primate will determine which of the differences between human and chimpanzee occurred due to mutations in the human lineage conserved by natural selection and thus provide an outgroup for human and chimpanzee comparison. It also has been observed that the duplications seen in human happened during the radiation of the ape lineage and since these duplications are not seen in old world monkeys, their sequence will give the evidence to the ancestral segments that lead to human genome. DNA sequences
(Caccone and Powell 1989), specific gene sequences, genetic loci arrangement on chromosomes (Graves et al., 1995), wide range of physiological characteristics that will help in the analyses of the interaction between gene-gene and its environment (Blanjero, 1993; Van deBerg and Williams-Blangero, 1996;) neurophysiological function (Carey and Rice, 1996; Kaplan et al., 1995; Higley et al., 1993) show great similarity between human and baboon, thus indicating genetic similarity and the close evolutionary relationship between baboons and human. The genomes of human and baboon are larger compared to chimpanzee due to the presence of more Alu repeat elements and L1 elements and also due to the increased rate of insertion in baboon compared to chimpanzee (Liu et al., 2003). Seven human autosomes have the same loci order as their baboon homologs and a recent evolutionary fusion unique to human gave rise to human chromosome 2 that exists independently in two baboon chromosomes 12 and 13 (Ijdo et al., 1991). In contrast, baboon chromosome 3 was formed by the fusion of baboon ortholog of human chromosomes 7 and 21 (Best et al., 1998). Similarly, baboon chromosome 7 is homologous to both human chromosome 14 and 15 (van Oorshot and VandeBerg, 1991), while baboon chromosome 10 is a fusion of human chromosome 20 and 22 (Rogers et al., 2000). The close evolutionary relationship between human and baboon also has been confirmed by other studies. For example, when a 29,920 human cDNA filter array was used to compare human and baboon bone marrow CD34+ cell global expression, the results showed a similar expression pattern and gene abundance with a difference of less than 3% (Gomes et al., 2001). Baboons also have been useful in studies involved in cholesterol metabolism and a region in baboon chromosome 18 has been associated with HDL levels (Cox et al., 2002).
Similarly, other experiments have reported the genetic component responsible for an atherogenic response to environmental stimuli (MacCleur et al., 1988; Blangero et al., 1990), cortical bone thickness, peak bone density (Kammerer et al., 1995) and relative organ weight (Mahaney et al., 1993). Atherosclerosis has been studied in baboon as the genes encoding various serum apolipoproteins and the LCAT and LPL genes encoding lipid metabolizing enzymes are 97-98% similar in the coding region between baboon and human (Hixson et al., 1993a to b). Baboons also are important models for human osteoporosis (Jerome et al., 1986), for human aging (Martin et al., 2002; Jayashankar et al., 2003) as well as for the development of vaccines for meningitis, HIV, influenza, tuberculosis, hepatitis B, and several other infectious diseases.

1.5.2.3 Human and cow (85MYA) humchr22: cow BTA17 and 5

It has been estimated that human and cow shared a common ancestor ~85 MYA and because of this recent divergence, the cow has played very important role in studying human diseases (Fries and Ruvinsky, 1999; Sanger, 1959; Stahmann et al., 1941; Collip, 1925; Wiltbank et al., 1961; Evans and Long, 1921). The genomic sequences of bovine-human orthologous regions can help in extrapolating information from veterinary science to human medicine and also in annotating human genome for conserved genes and conserved noncoding regulatory elements (Thomas and Touchman, 2002; Williams et al., 2003).

The first generation comparative genetic maps of human and bovine created by radiation hybrid mapping and EST sequencing identified 768 genes of which 638 had
human orthologs, 105 conserved segments in between the two genomes and the location of 41 translocations and 54 internal rearrangements (Band et al., 2000). The second generation map showed 20% improvement to the first map with 1463 genes, 195 conserved segments of which 31 are newly conserved (Wind et al., 2004). The comparison between human and cattle was informative in providing insights into mammalian chromosome evolution as seen in the highly conserved gene order of the distal part of bovine chromosome 18 and Human chromosome 19q, which was different in mice due to rearrangements in the proximal part of mouse chr7, locus order between human and bovine in the chromosomal segments containing genes FANCA, CDK10, SPG7, APRT, GALNS and SLC7A5 also was conserved but inverted (Goldammer et al., 2002). Complete synteny conservation is seen between four cattle chromosomes and their human orthologs (BTA12 and HSA13, BTA19 and HSA17, BTA24 and HSA 18, and BTAX and HSAX). The cattle human comparative map will help in understanding segment boundaries (Larkin et al., 2003), centromere repositioning and distribution of genes in known breakpoint regions (Pevzner and Tesler, 2003). Although seven human chromosomes (1,3,4,5,10,11,X) have repositioned centromeres in the cattle genome located within large conserved syntenic blocks without any visible gene rearrangements, the centromere of human chromosome 22, repositioned in bovine chromosome17. Thus, of the 29 total cattle autosomes, 20 have either complete homology with human chromosome or homology with p- or q-arm or homology to genes of a single human chromosome (Wind et al., 2004).

1.5.2.4 Human and mouse (60/80MYA)
It is estimated that humans and mice shared a common ancestor approximately 80-100 MYA (Li and Gaur, 1991; Kumar and Hedges, 1998). Since most human and mouse coding regions (~1.5% of their genomes) have a high degree of sequence similarity and to some degree in the noncoding regions, sequence comparisons readily can reveal orthologous genes (Touchman et al., 2001). Although the noncoding regions usually show great divergence because they are not under selective constraints, approximately 1.5% of the human and mouse genome are conserved in noncoding regions.

Numerous studies have investigated synteny breakpoints between human and mouse and it has been observed that there is an enrichment of segmental duplications between human and mouse genomes (Armengol et al., 2003; Bailey et al., 2004). Additional comparison of human and mouse sequences has identified several novel regulatory elements (Hardison et al., 1997) as seen in the analysis of stem cell leukemia (SCL) loci (Gottgens et al., 2001), cystic fibrosis transmembrane conductance regulator gene (CFTR) (Ellsworth et al., 2000), Bruton’s tyrosine kinase (BTK) (Oeltjen et al., 1997) and the α-like and β-like globin genes (Flint et al., 2001). Comparative analysis of human (autoimmune regulator (autoimmune polyendocrinopathy candidiasis ectodermal dystrophy)) AIRE gene with its mouse ortholog gene (Aire) also showed high conservation of both gene structure and expression pattern (Blechschmidt et al., 1999). It is well documented that the mouse is a good model for studying human diseases (Bedell et al., 1997) and in particular for understanding cancer such as human acute myeloma leukemia. As part of this dissertation research, sequencing and identifying genes involved in radiation induced leukemogenesis in mouse chromosome
2 region orthologous to human chr11p11-13 and human chr15q11-15 (both human chromosomes have putative tumor suppressor genes) was done in an effort to obtain a more detailed understanding of human radiation induced AML (Hayata et al., 1983; Rithidech et al., 1997; Rithidech et al., 2002).

1.5.3 Human and Zebrafish (450MYA)

1.5.3.1 Zebrafish as the model system for whole mount in situ hybridization

The last common ancestor of present day humans and fish is estimated to have lived 400-450MYA (Kumar and Hedges, 1998). This makes fish one of the most distant non-mammalian vertebrates with an available sequence that can be compared to humans to identify conserved functional sequences. Several fish species have been sequenced, two pufferfish genomes, Fugu rubripes and Tetradon nigroviridis and genome of Oryzias latipes (medaka) (Ishikawa, 2000) essentially are completed and a working draft of the Danio rerio (zebrafish) genome (Grunwald, 2002; Vogel, 2000) is available. Comparison of human genome and the compact 365 Mb of Fugu genome has identified 1000 new unidentified human genes, and other conserved sequences that do not correspond to genes (Aparicio et al., 2002). The zebrafish is an ideal model system for both comparison with a distant species and also for gene expression studies since most human genes are conserved in fish (Dodd et al., 2000; Dooley et al., 2000; Fishman, 2001; Postlethwait et al., 2000). In 1981, George Streisinger introduced zebrafish as a model genetic organism (Streisinger et al., 1981, 1986). This led Kimmel
(Kimmel et al., 1989, 1991a, 1991b, 1995), Westerfield (Westerfield et al., 1990, 1992), Eisen (Eisen et al., 1991, 1993), and others to study zebrafish cell fate, lineage, and patterning during early embryonic development and nervous system development. The zebrafish genome is about 2900cM and 1.7x10^9 bp (Hinegardner et al., 1972) or about half the physical size of the human genome. The embryos are small (3-4cm long) and transparent, with one female capable of producing over 200 embryos per week. Embryonic development is rapid (5 days) and external, with the heart beating by the end of the first day and most organs, or at least their primordia, in place by five days after fertilization (Kimmel et al., 1995). The availability of the genetic and physical maps (Postlethwait et al., 1994; Pijnacker et al., 1995; Johnson et al., 1995; Knapik et al., 1996; Postlethwait et al., 1998, Geisler et al., 1999; Hukriede et al., 2001), cloning resources (Amemiya et al., 1999; Collins et al., 1995; Woods et al., 2000; Hukriede et al., 1999; Kelly et al., 2000; Zhong et al., 1997), gene expression screens provide a background to understand developmentally regulated genes during zebrafish embryogenesis (Kudoh et al., 2001). Recent studies have shown that synteny between the zebrafish and human genome is conserved (Donovan et al., 2000) for some gene groups such as MHC class III region (Sultmann et al., 2000), but gene order along the syntenic chromosomes often differ and include inverted and/or transposed sequences (Barbazuk et al., 2000).

During the past 20 years, development biology, when supplemented with molecular and genetic information, has established that common genetic systems are conserved if they control the formation of organs during embryonic development of different animals. Thus, the genes that control early development in human embryos
can be studied utilizing the information obtained by the detailed analysis of the model genetic organisms such as zebrafish (Haffter et al., 1996) and mouse (Neidhardt et al., 2000). The recent availability of 15,000 unique zebrafish EST and microarray data has provided information to control zebrafish embryogenesis by gene coordination in a genetic network and thus help identify many organogenesis specific genes (Lo, J et al., 2003).

1.5.3.2 Global gene duplication in zebrafish(350-400MYA)

A gene duplication occurred during chordate evolution that plays an important role in the evolution of both genomes and organisms, leading to gene function diversification and biological diversity (Ohno 1970; Lynch and Conery, 2000). Zebrafish has two orthologs for many mammalian genes which were introduced by whole genome or chromosomal duplications that occurred in bony fish (including zebrafish and fugu) after their divergence from the tetrapod lineage that includes humans (Van de Peer et al., 2002; Woods et al., 2000; Taylor et al., 2001).

Two models have been put forward to explain the survival of two or more genes in a genome, the classical model (Fisher, 1935; Haldane, 1933) and the duplication-degeneration-complementation (DDC) model (Force et al., 1999; Lynch and Force, 2000). The classical model predicts that one copy of the duplicate will accumulate null or deleterious mutation and usually degenerate to a pseudogene within a few million years and the other duplicate will retain the original function as gene loss is permissible in one duplicate as only one gene is required to maintain function similar to the ancestral single copy. However, this model fails to explain the duplicates present in
different genomes such as the human genome with at least 15% duplicated genes with 5.2% being segmental duplications (Li et al., 2001), 30-75 % of which are duplicated in the tetraploid fish lineages (Allendorf et al., 1975; Ferris and Whitt, 1979). Half of these duplicated genes have been maintained for over 30 million years in Xenopus (Hughes and Hughes, 1993) and 20% of the duplicated genes in zebrafish lineage has been retained for over 110 million years (Postlethwait et al., 2000).

In contrast, in the DDC model, the duplicated genes survive by evolving in two phases. In the first phase, one copy of the gene undergoes nonfunctionalization (complete loss of regulatory regions), while the other copy undergoes neofunctionalization, mainly by changes in regulatory regions or subfunctionalization, by complementary changes in regulatory regions. This model is similar to the classical model but differs in that during the second phase, the duplicate genes that acquire neofunctionalization or subfunctionalization are maintained by removal of redundant subfunction (Lynch et al., 2001). A few examples of duplicated genes in zebrafish maintained by DDC model are listed in table 1.2.

| **Hox clusters:** single cluster in invetebrates, four clusters in tetrapod vertebrates (mouse ad xenopus), more than four clusters in teleosts | Amores et al., 1998; McGinnis and Krumlauf et al., 1992; Greer et al., 2000; Godsave et al., 1994; McClintock et al., 2002 |
| Microphthalmia associated transcription factor (Mitf) | Altschmied et al., 2002 |
| SOX9 (developmental regulator) One copy in mouse, two copy in zebrafish show subfunctionalization | Yan et al., 2002 |
| Duplicated nodal class genes shown by mutants ndr1 (nodal related 1) and Cyclops (cyc) | Sampath et al., 1998 |
| Engrailed (four copies) | Amores et al., 1998 |
| Msx genes (duplicated) | Ekker et al., 1997 |
| Fox1 (three copies) | Solomon et al., 2003 |
Two copies of cytochrome P450 aromatase (Cyp 19) differentially expressed in brain and ovary | Chiang et al., 2001

Two copies of JAK family of protein tyrosine kinase. Jak2a expressed in nervous system and developing buds and Jak2b expressed in developing lens and nephritic ducts | Oates et al., 1999

Duplicates of nkx genes | Lee et al., 1996

| Table 1.2 Examples of genes following the DDC model for the maintenance of the extra copies of duplication |

Thus, many zebrafish genes and other multigene families, such as the immunoglobulins and T cell receptors in vertebrates, come into existence or expand as a result of tandem duplication and inter chromosomal transpositions (Lewin 2000) and survive as a result of positive selection for both gain of function of each new family member, the classical model, and or partition of function among the members, the DDC model.

1.5.3.3 Mutation studies in zebrafish

In identifying genes involved in a wide variety of developmental processes, zebrafish has become an ideal model for molecular and genetic studies with the recent introduction of large scale mutagenesis screens (Amsterdam et al., 1999). Both forward genetic mutagenesis screens (Mullins et al., 1994) and reverse genetic gene expression knock down by morpholino injection (Nasevicius and Ekker, 2000) have been done in the zebrafish. The two main strategies for random saturation mutagenesis are chemical mutagenesis with ethyl-nitrosurea (ENU) that has been to generate over 7000 point mutagenesis (Grunwald and Streisinger, 1992; Mullins et al., 1994; Solnica-Krezel et
al., 1994) and insertional mutagenesis (Kidwell 1986; Gridley et al., 1987) mediated by retroviral insertion (Amsterdam et al., 1999; Golling et al., 2002; Schier et al., 1996; Amsterdam et al., 1999; Postlethwait and Talbot, 1997). Some zebrafish organs are morphologically and functionally similar to human; it has been proposed that large scale mutagenic screens could help in understanding the mechanisms and pathways directly relevant to human disease and therapy (Stainier and Fishman, 1994; Stainier et al., 1996; Pack et al., 1996; Ransom et al., 1996; Weinstein et al., 1996; Neuhauss, et al., 1996; Whitfield et al., 1996; Bennett et al., 2001). The efficiency of the mutant is determined by resemblance of the mutant phenotype to mimic the particular disease. Many human diseases with gene homologs in zebrafish include Huntington’s disease gene homolog (Karlovich et al., 1998); steroid imbalance associated disease (Lai et al.,1998) and Alzheimer’s disease (Leimer et al., 1999); Hotl-Oram syndrome caused by mutation of the Tbx5 gene (Garrity et al., 2002); the von gogh mutant formed by mutation in tbx1 that causes defects in ear, pharyngeal arches, absence of thymus, fusion, and loss of neural crest derived pharyngeal cartilages; and reduction in endodermal pouches and aortic arches, a phenotype resembling DiGeorge syndrome (Piotrowski et al., 2003).
2. Materials and Methods

2.1 Sequencing of the Chimp, Baboon, Cow, Mouse and Zebrafish

Clones Syntenic to Human Chromosome 22 Cat Eye and DiGeorge Region

BAC clones syntenic to human chromosome 22 sequenced during the course of this dissertation research were from the rp43 (Donald), ch251 (Clint) and ptb (Gon) chimpanzee BAC libraries, rp41 baboon library, rp42 bovine library, ct7 and rp23 mouse libraries and the ch211 zebrafish BAC library. Chimpanzee BACs were selected by locating the BAC end sequences on human chromosome 22 using BLAST (Altschul et al, 1990, 1997) and later confirmed by comparative genomic maps constructed by an alignment between BAC end sequences of chimpanzee with the human genomic sequences (Fujiyama et al, 2002). Clones for mouse were acquired from Jim Lund (mouse chromosome 16) and Kanokporn Rithidech (mouse chromosome 2); baboon and bovine were acquired from Chris Amemiya and Harris Lewin, respectively; and zebrafish clones were chosen from the chori 211 library using gene specific probe
screening of primary and secondary pools of the clones. Ten primary pools of the zebrafish BAC library \( (76 \times 384 = 29184/10 = 2918 \text{ BAC’s in each of the 10 pools}) \) were screened by PCR using gene specific primers for the presence of genes of interest. When a pool showed a PCR product, the entire secondary plate was screened via PCR for that particular BAC clone. When a primer pair indicated more than one of the primary pools, due to the presence of 2 to 3 copies of the genes, after the identification of the plate from the positive primary pools, the secondary pool was rescreened by PCR to determine the exact column and row of the individual clone carrying the gene of interest.

### 2.1.2 Random Shotgun Sequencing

The selected BAC clones were sequenced using the shotgun method (Chissoe et al., 1991; [http://www.genome.ou.edu](http://www.genome.ou.edu); Roe et al., 1995, 1996). The target BAC DNA was isolated using a modified cleared lysate procedure (Birnboim and Doly, 1979) ([http://www.genome.ou.edu/DblacetateProcV3D.html](http://www.genome.ou.edu/DblacetateProcV3D.html)) from one liter or 200 ml cultures of each clone harboring the BAC of interest after late log growth in LB medium with appropriate antibiotic at 37°C. A portion of this large scale DNA then was physically sheared by nebulization at 8 psi or hydroshearing into 1.5 kb-4 kb fragments. Shearing generates DNA fragments with single-stranded end, that then are converted to blunt ends using Klenow DNA polymerase and phosphorylated at their 3’ blunt ends using T4 polynucleotide kinase at 37°C in a water bath for 30 minutes. Two to four kb DNA fragments then were size selected by electrophoresis on 0.8% low
melting point agarose gel, and extracted from the gel using TE-saturated phenol, TE-saturated phenol-chloroform, chloroform, and finally a water-saturated ether extraction. After ethanol precipitation, the pellet was dried and dissolved in ddH₂O. These size selected DNA fragments then were ligated into PUC18/SmaI vectors (blunt ended and dephosphorylated at 5’ ends) using T4 DNA ligase and subsequently, electroporated into E.coli XL1 Blue MRF’ electrocompetent cells using a Bio Rad pulser. This transformation mixture containing X-gal (24mg/ml in DMF) for blue and white colony selection was plated onto 24.3 x 24.3 cm petri dishes and incubated at 37°C overnight. The white colonies were picked into 384 microtiter plates using Q-Pix (Genetix) colony picker and incubated in a HiGro oxygenated shaker incubator (Gene Machines Inc.) for 22 hrs at 37°C at 520rpm. The template for sequencing (subclone DNA) then was isolated from individual colony growths using an automated alkaline lysis protocol (modified Birnboim and Doly, 1979) on Zymark (Twister II and Sciclone subunit) and V-prep robots. The cycle sequencing reaction was pipetted on the V-prep robot by adding the isolated subcloned DNA, forward or reverse universal primer, and ABI BigDye or Amersham ET terminator mix (contains fluorescent-labeled ddNTP terminators, dNTPs, Taq DNA polymerase and MgCl₂). This was followed by incubation for 60 cycles consisting of 95°C for 10 seconds, 50°C for 5 seconds, and 60°C for 4 minutes in a Perkin Elmer Gene Amp PCR system 9600 thermocycler. The resulting nested fragment sets were purified by 95 % ethanol/0.12N sodium acetate, pH 4.5, precipitation that removed the unincorporated dye terminators. Following drying, the DNA pellet was resuspended in 20-50 µl of sterile ddH₂O and electrophoretically loaded onto an ABI 3700 capillary sequencer capillary containing POP-5™ polymer.
and electrophoresed for 3 hrs at 6.5 kV. Sequence reads then were transferred to a local
SUN workstation where the base sequences were determined by Phred (Ewing et al.,
1998) followed by sequence reads assembly into large contiguous regions (contigs)
with the Phrap assembly program (P. Green, pers. comm.). The consensus sequence
from the assembly is created using a program Consed (Gordon et al., 1998). An average
of 6-8 fold coverage of the shotgun sequence data, with typical read length of from 4-
500 bases, from an initial random assembly generated contigs that usually covered 95-
99.9% of the cloned sequence (http://www.genome.ou.edu/poisson_calc.html) (Lander
and Waterman, 1988). Directed sequencing either with custom synthesized primers off
the shotgun sub-clone or by PCR produced template from the BAC clones were
sequenced to close either captured or uncaptured gaps. The gaps occurred because a
region might have a lower coverage or no coverage at all by shotgun reads, by the
presence of low quality or shorter than average reads on the ends of the contigs, or by
secondary structures due to high GC content or long repeat regions. Often different
nucleotides, such as dGTP or dITP for GC rich gaps, or dRhodamine dye terminators
for AT rich gaps, were used to close gaps in especially difficult regions. However,
when gaps could not be closed by sequencing from subclones or PCR amplification
from the original BAC, the desired region of the target BAC often would be amplified
using a PCR mix in which the dGTP was replaced by 7-deazadGTP (GC rich gaps)
(Deirick et al., 1993; Mizusawa et al., 1986). Under these conditions, the resulting
template was more easily denatured during the cycle sequencing reaction, since 7-
deazadGTP only forms 2-H bonds with dCTP. In addition, 5-10% dimethyl sulfoxide
(DMSO) (Winship, 1989), 1M betaine (Henke et al, 1997), or formamaide (Zhang et al,
1991) could be added to the cycle sequencing reaction to improve denaturation of secondary structures and prevent formation of primer dimers. This approach improved sequence quality and read length, using custom synthetic primers and either the Amersham E.T mix, the ABI BigDye mix, the dGTP mix or the dRhodamine mix. In regions where BAC end reads were not found, primers were made specific to the BAC vectors and PCR or MPCR amplified with different sets of primers designed at the ends of the contigs to correct any mis-joining contig-end regions. Typically, a combination of several of the above approaches were needed to close all gaps and reduce the error rate of a sequenced BAC insert to less than one uncertain base per 10,000 bases.

2.1.3 Analysis of the Sequences using Different Tools

Analysis of the resulting DNA sequences was required to locate the biologically relevant information embedded in the sequence. To define the basis of molecular evolution and phylogenetic relationships between organisms, alignment of the sequences were revealed differences and similarities between the sequences in both coding and noncoding regions. For shorter regions with high sequence conservation, Align (Myers and Miller, 1989) or Lalign (Huang and Miller, 1991) (if two sequences are compared) and Clustalw (Higgins et al., 1994) or Clustalx (Jeanmougin et al., 1998), or T-coffee (Notredame et al., 2000) (if more than two sequences were compared) were used. The clustal programs pairwise align the sequences by calculating the best match between sequences such that differences, similarities, and identities become evident, thus defining biological meaning to alignments (Eddy, 1995). For
larger sequences, both local alignment programs that determine high quality alignments between sequences irrespective of their order or orientation and global alignment programs that produce a single optimized comparison on the entire length of the sequences were used. Here two local alignment programs, Pipmaker (two genomic sequences) (Elinitski et al., 2002; Schwartz et al., 2000) and Multipipmaker (multiple genomic sequences) (Schwartz et al., 2003), were used. These programs are based on performing a blastz local alignment and generating a percentage identity plot between the aligned sequences according to their position with reference to one of the sequences that is chosen as the base sequence. In these alignments, only matches 50% and above are shown, and an ungapped alignment is plotted. The global alignment program Vista (Visualization Tool for Alignment) (Mayor et al, 2000) uses the Avid (a global alignment program) to determine the percent identity between one or more sequences using a fixed sliding window. The resulting plot shows the alignments greater than 10% in relation to the position within the sequences. In contrast to multipipmaker, vista shows not only alignments of sequences but also helps identify conserved elements. zPicture, another local alignment and visualizing tool based on blastz, can identify more obscure evolutionarily conserved regions such as transcription factor binding sites and other regulatory elements (Ocharenko et al., 2004). Repeatmasker compares the input sequence with a database of repetitive elements and can identify these repeats in a DNA sequence (Smith and Waterman, 1981). eShadow can compare both nucleotide and amino acid sequences of closely related species and is useful in predicting conserved functional elements (Ocharenko et al, 2004). REPuter can analyze repeats on a genomic level and identify both unique sequences and low copy repeat regions (Kurtz...
et al., 2001). The prediction of conserved motifs was done using both ProfileScan (http://www.isrec.isb-sib.ch/software/PFSCAN_form.html) and Motif (http://motif.genome.ad.jp), software that compares amino acid sequence to the protein motifs characterizing protein families and domains in the Prosite and Pfam (Bateman et al, 2002) databases respectively (Gribskov et al, 1987; Bucher et al, 1996). Potential transcription factor binding sites were determined using TFsearch (www.cbrc.jp/research/db/TFsearch.html) and Motif, programs that compare an input sequence with a collection of vertebrate transcription binding sites. Toucan, another program that can identify regulatory regions, such as known transcription binding regions and predicted binding site motifs, also was useful in visualizing repeated sequences (Aerts et al., 2003).

2.2 Whole Mount In Situ Studies using Zebrafish as the Model Organism

2.2.1 Isolation of Total RNA

Isolation of total RNA from zebrafish embryos at different developmental stages was done using the Trizol method (Chomczynski and Sacchi, 1987). In this method, the embryos were dechorinated using pronase (30 mg/ml). 1ml of trizol (phenol in saturated buffer 38%, guanidine thiocyanate 0.8 M, ammonium thiocyanate 0.4 M, sodium acetate 0.1 M (pH=5), glycerol) was added to a batch of 100 embryos and vortexed, vigorously to homogenize the embryos. The light red colored supernatant was collected after centrifugation for 10 min at 14000 rpm. The lower dark red layer again was treated with trizol, vortexed, and centrifuged, and the upper lighter red color supernatant solution again was collected. This solution was divided into tubes
containing 1.3 ml each. 0.2 ml of chloroform was added to each tube, followed by centrifugation for 20 min. Then, the upper liquid layer was transferred into a new tube and equal volume of isoproponol was added, centrifuged for 10 min in the cold room. The resulting pellet was washed with 70% ethanol. After a quick spin the ethanol was discarded and the pellet was air dried for 10 min, dissolved in 100 µl of autoclaved DEPC ddH2O by vortexing, and incubated for 5 min at 60°C. 10 µl of the isolated RNA was checked on a formaldehyde gel and stored at -80°C.

2.2.2 RT-PCR and RNA Probe Preparation by In Vitro Transcription

Single-stranded cDNA was synthesized from the total RNA pool isolated from a combined embryos collected at 61, 72 and 77 hours post-fertilization (hpf). A 25 µl RT-PCR reaction contained 5x (avian myeloblastosis virus) (AMV) buffer, 25 mM MgCl2, 10 mM dNTP’s, individual primer pairs (200 pmoles), 1.5 units of Taq polymerase, 20 units of AMV-RT (Roche) and 2 µg of total embryonic RNA. The first cycle of PCR was incubated at 48°C for 45 min during which the RNA was reverse transcribed by reverse transcriptase AMV-RT into first single stranded cDNA (Sambrook et al., 1989) at a site that was determined by the primer (See Appendix table 5 for a listing of PCR primers designed corresponding to the 5’UTR and 3’UTR of the mRNA of interest). This immediately was followed by a denaturation step at 94°C for 2 min for complete inactivation of the transcriptase and a second round of regular PCR of 40 cycles (94°C for 30 sec, 60°C for 45 sec, 68°C for 45 sec and 68°C for 10 min, 10°C for 24 min followed by 4°C) for the second strand cDNA synthesis. Since the first PCR
product sometimes contained multiple length DNAs, they were size fractionated on a 1% agarose gel. The band corresponding to the expected size PCR product was excised, the DNA was eluted (freeze and squeeze method described earlier), and served as the template for a second round of PCR using a pair of nested primers. Typically, this treatment resulted in a single product that then was purified by treating it with shrimp alkaline phosphatase and exonuclease I, followed by a phenol/chloroform extraction, ethanol/acetate precipitation, and a 70% ethanol wash. The pelleted DNA then was dried and resuspended in 30 ml of sterile distilled deionized water. After end repair, the PCR product was cloned into Smal site of pUC18 and sequence verified. The cloned product then was excised from the pUC18 vector by EcoRI and BamH1 restriction enzyme digestion. It then was sub-cloned into the pBlueScript SK− at the EcoRI and BamH1 sites. Again, the insert was sequenced verified. Plasmid DNA was isolated by miniprep double-stranded DNA isolation (Roe et al., 1995) and purified by passage through a Zymo-spin column (DNA clean and concentrator kit). A single digestion with BamH1 was carried out to linearize the DNA and, after an preparative agarose gel and elution, the insert was purified using the Glass Milk gene clean procedure. After confirming the DNA concentration by agarose gel electrophoresis, a 20 μl of the in vitro transcription reaction containing linearized plasmid DNA, T7 RNA polymerase, 10 x transcription buffer, 10 x nucleotide mix with digoxigenin-labeled UTP (Roche), and RNase inhibitor (40 U/μl, cat#300152-51, Stratagene) was incubated at 37°C for 2 hrs followed by RNAase free DNase treatment for 25 min. The RNA probe then was precipitated with 0.5 M EDTA, 5 M LiCl (pH 5.0) and 2 volumes of cold ethanol and the pellet was washed with 70% ethanol. The dried pellet was
dissolved in diethylpyrocarbonate (DEPC) treated H₂O containing 1 µl of RNase inhibitor. The RNA probe was electrophoresed on a freshly prepared 1% agarose gel to confirm the correct size of the DIG labeled RNA probe.

2.2.3 Single Stranded Exon Specific DNA Probes and Primer Design

The expression pattern of the known and unknown human chromosome 22 genes by systematically identifying the expression of zebrafish orthologs in zebrafish embryos. A reciprocal BLAST comparison of the predicted human chromosome 22 genes to the recently available compilation of zebrafish predicted genes from the zebrafish whole genome shotgun assembly gave the sequence of the zebrafish orthologs of the human chromosome 22 genes (exons). Exon specific PCR primers were picked using PrimOU and then crossmatched to the entire zebrafish genome sequences to confirm that the primers were unique. These unique primers (20mers) then were used to amplify the gene of interest using the zebrafish genomic DNA as the PCR template. For this purpose zebrafish genomic DNA was extracted using the protocol described by A. Fritz in ZFIN. After removing the embryo from the chorion, 50 embryo’s (7 days old) were transferred into 1.5 ml Eppendorff tubes and rinsed with water. After the water was completely removed, 1ml of extraction buffer (10 mM Tris pH 8.2, 10 mM EDTA, 200 mM NaCl, 0.5% SDS, 200 µg/ml proteinase K) then was added to 50 embryos and incubated overnight at 50°C. This resulted in completely resuspending the embryos in the extraction buffer. The embryonic genomic DNA was extracted by phenol/chloroform extraction followed by ethanol/salt precipitation. The resulting DNA
pellet was dried and dissolved in 50-100 µl of water and stored at -20°C after
determining the concentration by measuring the absorbance at 260 nm. The
concentration of the genomic DNA (1 A<sub>260</sub>=50 ng/µl) obtained typically was between
the range 100-1000 ng/µl. The PCR reactions were performed with 95°C hold for 5
min, followed by first 10 cycles, denaturation at 94°C for 30 sec, annealing at 55°C for
30 sec and elongation at 72°C for 30 sec and then for 30 cycles, denaturation at 94°C
for 1 min, annealing at 60°C for 2 min and elongation at 72°C for 1 min, for a total of
40 cycles. An alternative touch-down cycling PCR program also was used with the
following cycling conditions: 94°C hold for 2 min followed by 30 cycles, denaturation
at 94°C for 1 min, annealing at range of temperatures 65°C-50°C with an decrement of
2°C (65°C for 2 cycles, 63°C for 3 cycles, 61°C for 4 cycles, 60°C for 4 cycles, 58°C
for 4 cycles, 56°C for 4 cycles, 54°C for 4 cycles, 52°C for 3 cycles and 50°C for 2
cycles), elongation at 72°C for 2 min followed by 10 cycles of 94°C for 1 min, 58°C
for 1 min and 72°C for 2 min and an additional extension step at 72°C for 5 min (Don
et al., 1991; Weinholds et al., 2003).

In cases where multiple bands were obtained, nested PCR was performed to
enrich the specific band of interest using nested primer pairs inside the first set of
primers. Amplified PCR products then were loaded onto a 3% low melt agarose gel and
after electrophoresis, the specific sized DNA was extracted from the gel using the
freeze and squeeze method, precipitated with ethanol/salt, washed with 70% ethanol,
dried, and dissolved in ddH2O. The purified PCR product then was used as a template
for unidirectional PCR (Fuhrmann et al, 1994, Vanholme et al., 2002; Knuchel et al.,
2000; Kitazawa et al., 1999) with a single primer (sense or antisense) at time in the
presence of DIG-dUTP (PCR DIG labeling mix. Roche) to produce the single stranded DNA sense and antisense probes (Patel and Goodman, 1992; Seydoux and Fire, 1995).

The cycling conditions for unidirectional PCR were: hold for 5 min at 95°C, followed by 10 cycles denaturation at 94°C for 1 min, annealing at 58°C for 1 min and elongation at 72°C for 1 min and then 30 cycles denaturation at 94°C for 1 min, annealing at 60°C for 1 min and elongation at 72°C for 1 min, for a total of 40 cycles. The DIG labeled single stranded DNA probe were precipitated by ethanol/salt, washed with 70% ethanol, and dissolved in hybridization buffer.

2.2.4 Whole Mount In Situ Hybridization

Zebrafish were raised and maintained under standard conditions of 27°C (Westerfield, 1994). Fish embryos from 12, 16, 24, 48 and 72 hpf developmental stages were collected and incubated in 28.5°C (Kimmel et al, 1995) in the presence of PTU (phenylthiourea-0.003%) to prevent melanization of the embryos and then, after dehydration, stored at -20°C in 100% methanol. The mutant embryos, HI1373 (Golling et al, 2002) and Ace mutant (Brand et al, 1996) similarly were isolated and noted for the morphological defects. Prior to hybridization, the dehydrated zebrafish embryos stored in 100% methanol were rehydrated by suspension in 1xPBST (Phosphate Buffered Saline Tween-20) immediately prior to a hybridization experiment. The embryos in 1x PBST were treated with proteinase K (10 mg/ml) by
incubation at 20°C for 1-20 min. Here the short time incubation was for early embryonic stages and longer incubation was for later stages, as previously determined by pilot studies that determined the incubation time needed for maximum permeability with minimal embryo disruption. The embryos then were treated with glycine (25 mg/ml) to inhibit proteinase k and after washing with PBST, were fixed in 4% paraformaldehyde (PFA), washed again with PBST and distributed into separate wells of a flat bottom 96-well microtiter plate. In some instances, heating the embryos in PBST before the prehybridization step reduced the background hybridization to achieve a clearer expression pattern. Pre-hybridization and hybridization was performed at different temperatures ranging from 50 to 70°C (depending on the probe length and binding specificity to reduce the background hybridization). The pre-hybridization was carried out for 1h in hybridization solution (50% formamide, 5X SSC, heparin (50 μg/ml), yeast RNA (500 μg/ml ), 0.2% Tween-20, 1M citric acid pH=6) followed by hybridization in 50% hybridization solution with DIG labeled ssDNA probes for 16 h. A series of gradient washes involving high salt (2x SSC), high temperature (ranging from 48-70°C depending on the hybridization temperature used earlier during probe binding). A low salt (0.2x SSC), high temperature (48-70°C) highly stringent wash, followed by a low salt (0.2x SSC), room temperature wash, automated on the Robins Hydra 66 robot, was used to remove any unbound probe and reduce the background. After these washes, the embryos were treated with 2% blocking solution (BSA-100 mg, sheep serum-1 ml, 1X PBST) for 1 hour and then incubated overnight by shaking at 4°C with anti-DIG antibody (1:10,000) (anti-digoxigenin-AP, Fab fragments, Roche). Again after a series of washes with 1x PBST and NTMT (5 M NaCl, 1 M MgCl₂, 1 M
Tris pH=9.5, 20% Tween-20) on the Hydra, the color was enhanced by incubating at room temperature in a substrate color reaction mixture (NBT (nitroblue tetrazolium)/BCIP (5-bromo-4-chloro-3-indolyl phosphate). Once a stable color was maintained, the reaction was terminated by washing twice with 1x PBST. The zebrafish then were visualized and photographed on a Leica MZFLIII Stereofluorescence microscope with an attached 35mm camera and Leica DM4500B DigitalMicroscope with an attached high resolution digital CCD color camera (MicroPublisher 3.3 RTV) with high speed real time viewing.
3. Results and Discussion

The 25 chimpanzee, 9 baboon, 6 bovine, 2 mouse and 3 zebrafish BAC clones sequenced, and combined with additional syntenic clones sequenced by others, prior to the analysis and comparison to the human chromosome 22 CES-DGCR region genes, are listed in tables 3.1, 3.2, 3.3, 3.4 and 3.5 below, with the corresponding marker genes annotated for each clone and a pictorial representation of the chimpanzee chromosome 22 clones in the region syntenic to the CES-DGCR region of human chromosome 22 is shown in figure 3.1.

**Figure 3.1 Chimpanzee BACs Studied.** The black boxes represent the chimpanzee BACs syntenic to human chromosome 22 sequenced during this dissertation research. The cones represent the gaps between the sequenced projects, the blue outlined boxes represent the projects that require sequencing to close the gaps and the numbers above each black rectangle represent the marker genes listed in table 3.1.

<table>
<thead>
<tr>
<th>Chimpanzee clones</th>
<th>Marker genes encoded</th>
</tr>
</thead>
<tbody>
<tr>
<td>rp43-118p7</td>
<td>TOMM40, L1H 3’ homolog, FKSG30</td>
</tr>
<tr>
<td>Baboon Clones</td>
<td>Marker genes encoded</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>1 rp41-88j2</td>
<td>Dgcr2, carbonic anhydrase 15, kiaa0649, HSPC009 protein, DnaJ protein, POM121 kiaa4041, PRODH, DGCR6</td>
</tr>
<tr>
<td>2 rp41-5C11</td>
<td>USP18, XP_372891, RVT,dgcr6,prodh,kiaa4041,POM121 membrane glycoprotein like 1</td>
</tr>
<tr>
<td>3 rp41-177K13</td>
<td>XP_372891, USP18, tuba8, Pex26, IS10-RIGHT TRANSPOSASE (Tn10)</td>
</tr>
<tr>
<td>4 rp41-367I20</td>
<td>CECR1, CECR5, CECR6, IL17R</td>
</tr>
<tr>
<td>5 rp41-110d13</td>
<td>XP_372981, Putative 150(LIH3'), Dgcr6, Prodh, NgR(Nogo66- reticulon 4 receptor precursor)</td>
</tr>
<tr>
<td>6 rp41-30f24</td>
<td>PRODH, DGCR6, Putative 150(LIH3'), XP_515001(Pan), XP_372891,BCR protein, bcr kiaa3017, Znf74, scarf2</td>
</tr>
<tr>
<td>7 rp41-36c24</td>
<td>CRKL, SNAP29, PIK4CA, SERPIND1</td>
</tr>
<tr>
<td>8 rp41-53o7</td>
<td>PIK4CA, SERPIND1, PIK4CA(shorter), HIC2(hypermethylated in cancer 2 protein-related to gene on chromosome 22, similar to hypothetical protein XP_525533.1(Pan), hypothetical protein-DKFZp43H177.1 (Kiaa1666)</td>
</tr>
<tr>
<td>9 rp41-2d18</td>
<td>Predicted similar to Na/H exchanger 3, KIAA1418, PCQAP, Kelch like (kelch like variant)</td>
</tr>
</tbody>
</table>

Table 3.2 The baboon BAC clones from the rp41 library syntenic to human chromosome 22 sequenced and their associated marker genes.

<table>
<thead>
<tr>
<th>Bovine clones</th>
<th>Marker genes encoded</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 rp43-94i21</td>
<td>Repeats similar to human CES genomic DNA</td>
</tr>
<tr>
<td>3 rp43-130h22</td>
<td>FKSG30, POTE2A, similar to Nek2 (NimA), similar to NF1</td>
</tr>
<tr>
<td>4 rp43-26n14</td>
<td>Similar to NF1, POTE2B, NimA</td>
</tr>
<tr>
<td>5 rp43-11m16</td>
<td>Similar to phosphoinositide 3-kinase catalytic gamma polypeptide, FKSG72</td>
</tr>
<tr>
<td>6 rp43-11n14</td>
<td>XK related protein 3, similar to immunoglobulin kappa light chain, KCNM3BL, Kiaa1629 Zinc finger protein 532</td>
</tr>
<tr>
<td>7 rp43-3m22</td>
<td>GRB2</td>
</tr>
<tr>
<td>8 rp43-24a4</td>
<td>IL17R, CECR5, CECR6,CECR1</td>
</tr>
<tr>
<td>9 ptb-23e10</td>
<td>CES region specific repeat</td>
</tr>
<tr>
<td>10 ptb-71c8</td>
<td>Cecr2, Trf2</td>
</tr>
<tr>
<td>11 rp43-131i22</td>
<td>CEKR2, Kiaa1740, CEKR2, Kiaa1740</td>
</tr>
<tr>
<td>12 ptb-142b7</td>
<td>MIL1, BCL2L13, ATP6E1, SLC25A18, CEKR2</td>
</tr>
<tr>
<td>13 rp43-48a6</td>
<td>MIL1, BID, Kiaa0819, Kiaa1374, Mical-3</td>
</tr>
<tr>
<td>14 rp43-14h17</td>
<td>Similar to MICAL-3</td>
</tr>
<tr>
<td>15 rp43-96g17</td>
<td>PEX26, TUB8, USP18</td>
</tr>
<tr>
<td>16 rp43-75h20</td>
<td>SLC25A1,DGCR14, STK22A,DGCR2, POM121 membrane glycoprotein like 1,</td>
</tr>
<tr>
<td>17 ptb-33n14</td>
<td>HIRA</td>
</tr>
<tr>
<td>18 rp43-55o12</td>
<td>HIRA, UFD1F, CDC45, CLDN5</td>
</tr>
<tr>
<td>19 ptb-31f12</td>
<td>CLDN5, CDC45</td>
</tr>
<tr>
<td>20 rp43-14n19</td>
<td>SEPTIN5(CDREC-1), TBX1, GNB1L, TXNRD2, TR</td>
</tr>
<tr>
<td>21 rp43-16g14</td>
<td>ZDHHC8,RANBP1, Hpal9c, T10,ARVCf, COMT,</td>
</tr>
<tr>
<td>22 ptb-34l22</td>
<td>T10, DGCR8, Hpal9c</td>
</tr>
<tr>
<td>23 rp43-48d22</td>
<td>PRODH, DGCR6L, USP18, ZNF74, SCARF2, KELCH-like protein 12, Q96B68</td>
</tr>
<tr>
<td>24 rp43-9f16</td>
<td>ZNF74, SCARF2</td>
</tr>
<tr>
<td>25 rp43-47m18</td>
<td>LZTR-1, CRKL, SNAP-29</td>
</tr>
</tbody>
</table>

Table 3.1 The chimpanzee BAC clones from the rp43 and ptb libraries syntenic to human chromosome 22 sequenced and their associated marker genes.
Table 3.3 The bovine BAC clones from the rp42 library syntenic to human chromosome 22 sequenced and their associated marker genes.

<table>
<thead>
<tr>
<th>Mouse clones</th>
<th>Marker genes encoded</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Rp21-598k13</td>
<td>HIRA(histone cell cycle regulation defective homolog A) chromosome 22.</td>
</tr>
<tr>
<td>2. Rp21-594l4</td>
<td>Glutathione-S-transferase, Degradation in ER protein 3(DER1 like protein3)</td>
</tr>
</tbody>
</table>

Table 3.4 The mouse BAC clones from the rp21 library syntenic to human chromosome 22 sequenced and their associated marker genes.

<table>
<thead>
<tr>
<th>Zebrafish clones</th>
<th>Marker genes encoded</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Ch211-31g3</td>
<td>Zdhhc8(36-186bp-isoform2), XP_373945, HpaI9c Dgcr8(variant), Zdhhc8(377-709bp), Dgcr8(WW domain), T10 (syntenic to human chromosome 22)</td>
</tr>
<tr>
<td>3. Ch211-31g8</td>
<td>Trace amine associated receptor 9(TAAR9) (syntenic to human chromosome 6)</td>
</tr>
</tbody>
</table>

Table 3.5 The zebrafish BAC clones from the ch211 library syntenic to human chromosomes 22, 5, 11 and 6 sequenced and their associated marker genes.

As shown below in figure 3.2, the chimpanzee clones sequenced covered almost all genes syntenic to human chromosome 22 that were listed in table 3.1 above except for those in the 325 kb region between clones rp43-96g17 and rp43-75h20 that encode the genes DGCR6 and PRODH, the 451 kb region between clones rp43-48d22 and rp43-47m18 encoding the genes PCQAP, PIK4CA and SERPIND1, and the 466 kb region upstream of CES as also shown in figure 3.1.
Figure 3.2 Comparison between the nucleotide sequences of the human chromosome 22 CES-DGCR region and syntenic region in chimpanzee. The dot plot shows the comparison and match between the nucleotide sequences of the human chromosome 22 CES-DGCR region (Y axis) and the sequenced syntenic region in chimpanzee. Each dot represents match of 50 bases and continuous dots form lines representing significant matches. The numbers in between the lines represent the gap sizes and the genes listed are those missing because of the sequence gap.

Although many of the marker genes present in chimpanzee, baboon, bovine, mouse and zebrafish have conserved synteny when compared to human chromosome 22, each species encodes additional genes and unique repeats such as, for example, the bovine LDLa, pregnancy associated glycoprotein and a unique Alu bucentaur repeat absent in humans. The PIP comparison shown in figures 3.3(a) and 3.4(b) for a 389kb representative region encoding genes TBX1 to ZDHHC8 from 20 different organisms; human, chimpanzee, macaca, mouse, rat, dog, cow, opposum, chicken, frog, zebrafish, fugu, tetrodon, ciona, fruitfly, mosquito, honey bee, C. elegans, and yeast clearly shows conservation in the coding region for all species compared. As can be seen in this
figure, the DNA sequence from two chimpanzee BAC libraries (sometimes three BAC libraries, see Methods) also were included in this comparison when it was available.

**Figure 3.3 Overview of PIP.** This figure gives an overview of the PIP in figure 3.4 below. The red color indicates $\geq 70\%$ strong alignment for 100bp without gap and green color indicates lower percentage alignment. A trend in the coloring pattern showing the high degree of alignment at the top for closely related species (red) to decreasing degree of alignment (green) for distant species in the evolutionary time scale.

Large conserved areas of the noncoding regions are observed in all mammals with a greater degree of conservation seen in primates. The zebrafish homology with human chromosome 22 is shorter when compared to the mammals and mainly shows a high degree of identity in the exon coding regions. However, it is interesting to note that the gene order over this large region surprisingly is conserved over the species compared considering the large evolutionary distance between common ancestors.
Figure 3.4  The PIP output highlighting conservation. Clear conservation at the coding level between all species compared (cow and yeast missing orthologs for this gene) and conservation at the level of noncoding regions at a greater degree in primates than to other mammals with respect to humans.

A VISTA plot figure of the human, whole genome shotgun chimp-Clint, BAC-based chimp-Donald, whole genome shotgun macaca and BAC-based mouse sequences over the same 389kb region is shown in figure 3.5.
Figure 3.5 VISTA output highlighting comparison between whole genome shotgun sequence and BAC based sequence. The (A) section of VISTA shows three highlighted regions namely (1) which shows a missing exon in macaca while present in human, chimp-Clint, chimp-Donald and mouse and (2) missing noncoding region in chimp-Clint and macaca but present in human and chimp-Donald. (3) shows missing noncoding region in macaca and (4) shows missing noncoding region in chimp-Clint. Section (B) shows missing exon in macaca(5); the missing region in chimp-Donald in this section is due to gap in sequence. Section C shows in region (6) missing exons in chimp-Clint, (7) shows missing noncoding region in Clint and macaca, (8) shows missing noncoding region in Clint and (9) shows missing noncoding and exon in Clint. Section D shows in region (10) missing noncoding region in Clint, Donald and macaca but present in human and mouse. Region (11) shows missing exon in chimp-Donald and macaca but present in human, chimp-Clint and mouse.

Here it can be seen that there is unique conservation of both coding and noncoding regions between the mammalian species compared, although regions are missing in whole genome shotgun data sequence that are present in the BAC based sequences, especially in the noncoding regions. This observation points to one of the major shortcomings of the whole genome shotgun approach when
compared to a BAC by BAC ordered sequencing strategy as the whole genome shotgun approach seems to be lacking sequences that could encode important information further defining the differences seen between human and their close primate relatives that could be significant both at the phenotypic level and in the differences in their susceptibility to diseases.

3.1 Comparative Analysis of Human Chromosome 22 CES-DGCR Region

A comparison between humans and other organisms, in particular chimpanzee and baboon, revealed differences in their repeat elements (both type and number) and in their base composition because of base substitutions that often resulted in differing codon usage.

3.1.1 Repetitive Elements

Previous studies have shown that syntenic regions in species other than chimpanzee and baboon have lesser and shorter repeats than human (Aparcio et al., 2002; Mural et al., 2002) and that the increase in the repeats in humans mainly is due to the reduced deletion rate efficiency (Dehal et al., 2001). These repeat elements also have been suggested to facilitate the rate at which mutations can take place (Yoshio Miki, 1998).

Comparison of the gene repeats (as shown in Appendix table 1 and figure 3.6) between human, chimpanzee and baboon, reveals that human and baboon has a significant increase in the number of retroelement insertions (Liu et al., 2003).
Figure 3.6 Major repeats. Percentage distribution of the major repeats noticed in human, chimpanzee and baboon.

This specially is the case with Alu (SINE) repeats, since chimpanzee has fewer Alu sequences (56%) when compared to both humans (62%) and baboons (65%). It also is evident that the major group of repeats found in humans, chimpanzees, and baboons belong to the Alu, L1 and LTR repeat families. In humans, a majority of the Alu insertions belonged to the young AluY (originating less than 15Myrs ago) subfamilies of recent origin (Dagan et al., 2004; Roy et al., 1999), while others belong to the intermediate AluS family (20Myrs), predominately the AluSx subfamily, followed by AluSg, AluSq and AluSc subfamilies. For example, AluYa5 subfamilies are observed in the intron2 of the gene ATP6V1E1 and intron9 of the gene Kiaa0819, 297 and 309bp, respectively. Another human specific Alu repeat is the 300-317 bp AluYb9 repeat seen in the introns of genes SLC25A18, BID, MICAL3 and DGCR8 (Batzer et al., 1990, 1991, 1995; Caroll et al., 2001; Roy et al., 1999; Roy et al., 2001; Deninger et al., 1999; Matera et al., 1990; Jurka et al., 1993).

In chimpanzee, the Alu repeat families differed from those of humans (Hedges et al., 2004), and commonly belonged to oldest families of Alu repeats, the AluJo and
AluJb that entered the genome 65-40Myrs ago. Chimpanzee Alu repeat families also had unique insertions of recent new AluY subfamilies, AluYg and AluYc3, as observed in intron1 of the BCL2L13 gene, AluYc5 in the intron1 of the BID gene and AluYh9 in intron2 of the DGCR2 gene. In baboon, the repeats belonged to the intermediate family of AluS repeats (AluSx and AluSg). However, the major class of repeats were of the AluY family of repeats, with unique insertions of recent Alu subfamilies such as AluYa5 as observed in the intron 11 of the CECR1 gene and AluYg in intron1 of the Pex26 gene. The unique subfamilies of AluY repeats seen in human, chimpanzee and baboon are summarized in figure 3.7.

![Figure 3.7 Unique AluY repeats. Major classes of unique young AluY family present in human, chimpanzee and baboon.](image)

Other families of “old” Alu sequences, included members of the FLAMC/A and MIR families, that are present in significant numbers in both humans and baboons but is present in lower amounts in chimps. The Alu families present in baboons showed similarity both to humans and chimpanzees, but greatly exceeded in number than those present in humans (Liu et al., 2003).
In addition to Alu repeats, other conserved repeats also were observed. In humans, LINEs in the category L2, L1 (active class of LINEs) and L3 category with lesser numbers of L1M, L1PA and HAL1 elements. A similar trend also was noticed in chimpanzees, but in baboons the LINE repeats show major increases in the L1M, HAL1 and L2 elements over those present in humans. Overall, baboons show greater increase in LINE elements similar to those seen with Alu repeats. The number of LINEs present in baboons, humans, and chimpanzees almost correspond to the number of Alus. Earlier studies showed that the correlation between Alus and LINEs may be because Alus are dependant on the reverse transcriptase and endonucleolytic activity of the LINE retroelements, rendering their distribution very similar (Jurka et al., 2004). In the category of LTR repeats, the major elements belong to MaLR (LTR, MLT1H) and ERV1 (MER) classes in all three species and are greater in number in baboons and less in humans and chimpanzees (Smit et al., 1993; 1999). Other repeats such as the MER1 and MER2 types, occur more often in humans than chimpanzees and baboons. In the category of low complexity repeats, the GC rich repeats are seen more often in humans and chimpanzees while AT rich repeats are seen more often in baboons. Other simple repeats also are significantly increased in baboons and are decreased in both humans and chimpanzees. The major repeats in the bovine genome are of the bucentaur type, a bovine Alu–driven family that encodes a novel 280 amino acids protein in the Alu linked repetitive sequence region (Nobukuni et al., 1997). In the few clones sequenced in mouse and zebrafish in this dissertation, the major repeats seen in mouse mainly belonged to the LINE family both in the exons and introns in comparison to SINEs (mainly belonging to the old Alu subfamilies) and other low complexity and simple
repeats. The major repeats seen in zebrafish belonged to the low complexity and simple repeats followed by SINEs (old Alu subfamilies), LINEs (L2), and few belonging to the MER1, MER2 and also LTR/ERV1 repeats.

3.1.2 Locations of Repeat Element Insertions and Their Significance

Earlier studies have shown that Alu elements play an important role in differential gene regulation by alternate splicing. When present in the 5’UTR and 3’UTR Alu elements can play the role of an enhancer or silencer. Intronic Alu also can inactivate or change the function of a gene product by either creating an alternate splice site or interfering the splicing mechanism. Thus, they can alter the pre-mRNAs processing (for example, exon skipping) (Ricci et al., 2002; Ganguly et al., 2003), leading to altered gene products, when present within introns, 5’UTRs, and 3’UTRs (Mitchell et al., 1991; Knebelmann et al., 1995; Oh et al., 2001; Groover et al., 2003; Hayakawa et al., 2001).

In humans, chimpanzees, and baboons, the insertions of most repeats occur in all portions of a protein coding gene, i.e. in introns, 5’UTR, 3’UTR, and exons (Chen et al., 2002; Minghetti et al., 1993), as listed in the Table 3.6. Low complexity GC rich repeats, Alu, and LINE repeats are observed in these regions.

<table>
<thead>
<tr>
<th>Locations of repeat insertions</th>
<th>Genes of human chromosome 22 and homologous</th>
</tr>
</thead>
</table>

56
Table 3.6 List of genes seen in human, chimpanzee and baboon that contain repeat elements in their intronic, 5’UTR, 3’UTR and exonic regions.

<table>
<thead>
<tr>
<th>Region</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic</td>
<td>GRB2, IL17R CECR5, CECR1, CECR2, SLC25A18, ATP6V1E1, DGCR6, PRODH, Q9BYB2, DGCR2, DGCR14, HIRA, CDC45L, TBX1, GN1BL, USP18, TUBA8, PEX26, MICAL3, KIAA0819, Q8N1L1, BID, BCL2L13, TXNRD2, COMT, Q8IZ69, RANBP1, ZDHCC8, PIK4CA, ZNF74, SCARF2, PCQAP, SNAP29, CRKL</td>
</tr>
<tr>
<td>5’UTR</td>
<td>CLDN5, STK22B, DGCR6, RANBP1, Q8IZ69, ARVCF, CECR1</td>
</tr>
<tr>
<td>3’UTR</td>
<td>CECR1, BCL2L13, PEX26, Q8IW05/T10, RANBP1, CRKL, GNB1L</td>
</tr>
<tr>
<td>Exonic</td>
<td>HIRA, SLC25A1, GSCL, Q9BYB2, PRODH (in baboons), PCQAP and SCARF2 (repeat absent in chimpanzee and baboon exon), ZNF74, DGCR6L, ZDHCC8, TXNRD2 (repeat absent in chimpanzee exon).</td>
</tr>
</tbody>
</table>

Alu elements also play an important role in genome diversity by facilitating homologous recombination events, resulting in duplication, deletion, or translocation (Dagan et al., 2004). The presence of Alu sequences as well as other retrotransposable elements (B1, B2, MIR, LINEs) in pre-mRNA can influence and affect the polyadenylation of transcripts and translation efficiency, as well as the expression patterns (Rowold et al., 2000). These changes often result in serious genetic disorders (Wallace et al., 1991, Knebelmann et al., 1995). As a result, understanding more about the nature of the presence of these repeat elements and their location is very important to understand the evolutionary relationship between repeat elements and genetic disorders. The chimpanzee has two copies of the SNAP29 gene, both of which contain
2 exons. As seen in figure 3.8, copy 1 (chimptrans1-SNAP29) encodes exons 1 and 2 of the human gene while copy 2 (chimptrans2-SNAP29) encodes exons 4 and 5.

The ortholog of the human SNAP29 exon 3 is not present in either of the chimpanzee gene. Intron 2 of the human SNAP29 gene, is made almost entirely of SINEs, and particularly the AluY repeat that is the bridging intron for an additional 83 bp exon that connects the two separate transcripts in chimpanzee to a single transcript in human. The non-repeat regions in this intron showed sequences matching a portion of the cDNA of SNAP29 indicating that the intron was expanded by the addition of an exon likely driven by the presence of SINEs. Interestingly, the human SNAP29 gene organization of a single 5 exon gene in the genome is conserved from human through mouse and zebrafish as shown in the figure 3.9.
Figure 3.9  PIP output of SNAP29 gene highlighting conservation of exons and deletion in chimpanzee. The PIP output shows conservation of all exons between human, macaca, mouse, cow, frog and zebrafish indicated by red boxes. The missing segment in between the two chimpanzee copies represents the deletion in chimpanzee resulting to two copies.

In comparison to macaca, the intron 2 in human has more AluY and AluSg SINEs insertions. Intron 3 and exon 3 are similar in human and macaca. Thus the species comparison shows that intron 1, intron 2, exon 3 and intron 3 is actually deleted in chimpanzee resulting in the two copies.

Similarly human Pex26 gene has only one copy of the 7 exons, while chimpanzees have two copies, one with 2 exons and the other with 5 exons. Intron 2 in humans is the connecting link between two transcripts in chimpanzee that results in the one transcript in human as shown in figure 3.10. This new intron 2 in human is formed of parts of coding regions of Pex26 with a 142 bp insert of FLAM_C (SINE/Alu) insert.
Figure 3.10 PIP output highlighting unique SINE insertion in human Pex26 gene. Pipmaker output shows the unique intron2 only present in human that indicates the fusion of the two copies of Pex26 gene in chimpanzee.

Comparison of PEX26 gene with other species as shown in figure 3.11, shows two copies in mouse.

Figure 3.11 PIP output of Pex26 gene highlighting conservation of exons and deletion in chimpanzee. The PIP output shows conservation of all exons between human, and mousegene1 but the last exon 7 is missing in macaca, mousegene2 and cow. The exons conserved are indicated by red boxes. The missing segment in between the two chimpanzee copies represents the deletion in chimpanzee resulting to two copies.

One copy of the mouse gene has all exons conserved similar to the single human gene ortholog. The second copy of the mouse gene has the last exon 7 missing and has a similar gene organization as
seen in macaca and cow. Intron 2 in macaca is different from intron 2 in human and has a AluJo (133 bp) insertion

This comparison between humans, chimpanzees and other species not only shows how the single transcripts in human evolved, but also helps in indicating the susceptibility of the human genome to a variety of different diseases. This can be explained, as an Alu sequence originally has nine potential 5’donor sites and fourteen 3’ acceptor sites. The plus strand of the Alu element is known to have only four of the potential splice sites and the remaining is on minus strand. In this case, there may be a possibility of converting an intron into exon containing the Alu repeat, if the Alu element is inserted into the intron in an orientation opposite to the transcriptional direction of the gene in which it resides. The alteration thus caused by Alu insertions into exons may lead to human diseases by disrupting the genes involved in metabolism, signaling, and transport that are known to have a high number of Alu repeats clustered within them. (Mitchell et al., 1991; Kreahling et al., 2004; Dagan et al., 2004).

3.2 Significance of Nucleotide and Amino Acid Changes

3.2.1 Substitution Preference between First, Second, and Third Codon Position between Human, Chimpanzee, and Baboon

The nucleotide and amino acid changes between humans and other species not only will help in understanding the evolutionary history of the human species, but also will help in finding out the critical changes that differentiates human and other species. Appendix table 2 summarizes the nucleotide positional changes and preferences. It can be seen that 66% of the nucleotide changes observed between humans and chimpanzees...
is in the third position, with only 34% of the changes in the second and first positions. In the case of nucleotide changes between human and baboon, it can be seen that >60% of the changes are in the third position, and 39% of the changes are in the second and third positions. In the comparison between chimpanzee and baboon, 64% of the changes are in the third position and 36% of the changes are in the second and first positions.

Figure 3.12 First, second and third position nucleotide changes. Graph showing the preferences in position of the nucleotide changes between human and chimpanzee and human and baboon.

Overall, the majority of the changes are conserved changes because they mainly reflect substitutions in the third codon position. Figure 3.12 summarizes the three codon position substitutions between human and chimpanzee, and human and baboon. This data clearly shows that substitutions between human and baboon are greater than those observed for human and chimpanzee especially in the first and third positions. This also is supported by the substitution rate calculated using the program eShadow (Ovcharenko et al., 2004) as shown in Appendix table 2, where the average substitution rate (per base) between human and chimpanzee is 1.2% (23 gene set) while
the rate between human and baboon is 2.6% (set of only 11 genes) as shown in figure 3.13.

![Figure 3.13 Substitution rate between human, chimpanzee and baboon.](image)

This is similar to earlier observations seen between human and macaque and human and chimpanzee on chromosome 21 (Magness et al., 2005; Watanabe et al., 2004). All of these results indicate that purifying selection is playing an important role between human and chimpanzee, and that positive selection is playing a major role in human and baboon in comparison to human and chimpanzee (Kimura, 1977; Hughes and Nei, 1988).

### 3.2.2 Transitional Changes versus Transversional Changes
As shown in figure 3.14, most of the substitutions observed when comparing the human, chimpanzee and baboon codons are transitional changes (>70%) compared to transversional changes (<30%), a result similar to earlier studies (Anzali et al., 2003).
In human, \(G\rightarrow A\) and \(C\rightarrow T\) transitional changes are more frequent than \(A\rightarrow G\) and \(T\rightarrow C\) changes. In chimpanzee, \(G\rightarrow A\) and \(C\rightarrow T\) transitional changes are more frequent than \(A\rightarrow G\) and \(T\rightarrow C\) changes and less than that noticed in human. But in baboon \(A\rightarrow G\) and \(C\rightarrow T\) changes are more frequent than \(G\rightarrow A\) and \(T\rightarrow C\) changes. The transversions noticed in all three species show mainly \(G\rightarrow C\) or \(C\rightarrow G\) changes compared to all other transversions. In bovine, it could be observed that the changes also were transitional. Also, dinucleotide changes in codons instead of a single nucleotide change as commonly seen in human, chimpanzee, and baboon.

### 3.2.3 Synonymous(Ks) versus Nonsynonymous(Ka) Changes

To determine the extent of selective pressure or constraints influencing the coding regions, the ratio of \(Ka\), the number of nucleotide substitutions that change amino acids per nucleotide sites resulting to amino acid changes to that of \(Ks\), the
number of substitutions that do not change amino acids per nucleotide sites was calculated using DNASP (Rozas and Rozas 1999). Usually, when a Ka/Ks value is equal to one, the gene is said to be undergoing neutral changes, when a Ka/Ks value is >1, the gene is said to be undergoing positive selection fixing advantageous mutations but when a Ka/Ks value is <1, the gene is likely undergoing purifying selection against deleterious mutations (Zhang et al., 2003). In the comparison between human and chimpanzee, among the 39 genes for which the Ka/Ks ratio was calculated, 35 genes showed Ka/Ks values between 0.0-1.0 and only four genes between 1.0-3.0. Comparison between human and baboon in a set of 14 genes, (the genes compared between human and chimpanzee and human and baboon had genes both inclusive and exclusive of each sets) 13 genes showed Ka/Ks values between 0.0-1.0, only one gene had Ka/Ks value greater than 1 as shown in Appendix table 2. Thus, there is a possibility that the four genes in the set between human and chimpanzee, CECR6 (Ka/Ks=1.02), SLC25A18 (Ka/Ks=1.3), Kiaa0819-zf1 (Ka/Ks=1.05) and zf2 (Ka/Ks=1.14) and one gene in human and baboon comparison CECR6 (Ka/Ks=3.5) showing high Ka/Ks values of 1 or more may have mutations that are beneficial and were maintained by positive selection during evolution. In the overall comparison, among the 39 genes compared between human and chimpanzee, the average Ka/Ks ratio was 0.44. Between human and baboon 14 genes showed Ka/Ks=0.48 and between chimpanzee and baboon among 10 genes the average Ka/Ks=0.40. These results indicate that overall, purifying selection plays a major role in sequence conservation, as it prevents mutations from accumulating in functionally important regions where the
ratio $\text{Ka/Ks}<1$ (Wildman et al., 2003) and also indicate the strong functional constraints in these species (Pagani et al, 2005).

### 3.2.4 Codon Usage Preference between Human, Chimpanzee, and Baboon

According to the genome hypothesis (Grantham 1980; Grantham et al., 1980, 1981), every organism is known to have its own codon usage preference that is selective and has nonrandom usage of synonymous codons for encoding different amino acids but is similar between genes in the same organism. Usually, there is a preference between codons ending with G or C versus A or T. In the comparison of the codon usage between human, chimpanzee and baboon, shown in Appendix table 3, there is an overall preference for codons ending with G or C with correspondingly high relative synonymous codon usage (RSCU) values in the range of 1-4 versus codons ending in A or T ending with RSCU values ranging between 0-2, as in the absence of any codon usage bias, the relative synonymous codon usage (RSCU) value would be 1 (Sharp et al, 1986; Martin-Galiano et al., 2004). In amino acids with 6 codons leucine, serine, arginine, codon usage is more biased towards G or C (82%) over A or T (18%) ending codons as seen in figure 3.15.
Figure 3.15 Codon usage bias in amino acids with six codons. Amino acids with six codons showing preference for G/C ending codons.

As shown in figure 3.16 (a), in amino acids with four codons such as proline, the codon bias between G or C (56%) versus A or T (44%) is almost equal between humans, chimpanzees, and baboons. However, in the case of threonine as shown in figure 3.16 (b), the baboon has more bias towards G or C when compared to humans and chimps.

Figure 3.16 Preference for G/C or A/T ending codons. Codon for Proline (a) showing almost equal preference for G/C or A/T ending codons while Threonine (b) shows preference for G/C ending codons.
Glycine, valine, and alanine also are biased towards G or C ending codons. In the category of amino acids with two codons ending in either C or T, human, chimpanzee, and baboon show more bias towards codons ending in C over T, while in the category of amino acids with two codons ending in either G or A, humans, chimpanzees, baboons, and bovines show more bias towards codons ending in G over A. Similar results have been seen in drosophila with preference to G or C ending codons (Comeron et al., 1998) in contrast to Pseudomonous aeruginosa that has preference to A or T ending codons (Sau et al., 2005). These differences reflect the high codon usage bias in highly expressed genes that facilitates efficient translation (Wagner, 2000). Earlier studies have shown that preference for optimal codons over non-optimal codons boosts the translation mechanism as usage of less optimal codons correspond to less abundant tRNA, resulting to missense and nonsense errors during translation. Thus, there is a pressure for positive selection of optimal codons over non-optimal codons as it improves translational efficiency in an organism. In this comparison, the codon is more biased towards G or C ending codon which generally code for the most abundant tRNAs (Moriyama and Powell, 1997; Duret and Mouchiroud, 1999) and implies that the codon usage has been selectively conserved between the closely related species.
3.2.5 Substitutions, Insertions and Deletions: Hydrophobic vs Hydrophilic Amino Acid Changes

In the comparison between the human and chimpanzee, and baboon DNA sequenced, selective pressure plays an important role for the majority of the amino acid substitutions, insertions and deletions. The identification of insertions and deletions seen in human, chimpanzee and baboon in table 3.7 and Appendix table 4 was based on six possibilities as shown in figure 3.17.

![Figure 3.17 Phylogenetic tree representation to interpret indels.](image)

Figure 3.17 Phylogenetic tree representation to interpret indels. The figure represents the different explanations for insertions or deletions as recorded in table 3.8. (a) shows conservation in all species. (b) shows conservation in human, chimpanzee and an outgroup species other than baboon but deleted in baboon. (c) shows insertion unique only to human. (d) shows insertion in chimpanzee and conserved in human. (e) shows conservation in human and baboon and an outgroup species but deleted in chimpanzee. (f) shows deletion in human and conservation in chimpanzee, baboon and outgroup species.

Of the six possibilities shown in figure 3.17, the scenarios labeled a and c seem to occur more often than the others. Thus, there are more insertions than deletions in the human genome than in chimpanzee and baboon as well as differences in the sequences conserved in codon usage in the three sequenced chimpanzee libraries Clint (ch251),
Donald (rp43) and Gon (ptb) in comparison to humans although this evolutionary drift varies depending upon the human genes studied.

(a)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Human</th>
<th>Chimp</th>
<th>Baboon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q8N7E1/XKRY</td>
<td>aa 197A→206S; aa 423Y→S459 unique in human.<strong>insertions</strong></td>
<td>Absent in chimp</td>
<td></td>
</tr>
<tr>
<td>ENSG0000013860/GRB2</td>
<td>absent in human</td>
<td>11→146F unique in chimpanzee <strong>insertion</strong></td>
<td>absent in baboon</td>
</tr>
<tr>
<td></td>
<td><strong>Deleted</strong> in human</td>
<td><strong>deleted</strong> in chimp</td>
<td>314F→322W unique in baboon</td>
</tr>
<tr>
<td></td>
<td>736T→748P <strong>insertion</strong> in human</td>
<td>absent in chimp</td>
<td>absent in baboon</td>
</tr>
<tr>
<td></td>
<td>conserved in human</td>
<td>conserved in chimp</td>
<td>aa254C→270V highly diverged</td>
</tr>
<tr>
<td>IL17R</td>
<td>aa 184D→199S; aa252IPA254; aa310PEPI313aa <strong>insertions in human</strong></td>
<td>Absent in chimp</td>
<td>Absent in baboon</td>
</tr>
<tr>
<td>CECR6</td>
<td>aa1M→200Y; aa212G→254G; aa276L→281A conserved in human and baboon</td>
<td><strong>Deleted</strong> in chimp</td>
<td>aa1M→200Y; aa212G→254G; aa276L→281A conserved in human and baboon</td>
</tr>
<tr>
<td>CECR5</td>
<td>aa1M→42Q (unique to human)</td>
<td>Conserved with baboon with two aa changes in this region(VH→RA)</td>
<td>Conserved with chimpanzee with two aa changes in this region (RA→VH)</td>
</tr>
<tr>
<td>Genes</td>
<td>Human</td>
<td>Chimp</td>
<td>Baboon</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td>Deleted in human</td>
<td>aa252P→278V conserved in chimpanzee and baboon</td>
<td>aa252P→278V conserved in chimpanzee and baboon</td>
</tr>
<tr>
<td></td>
<td>absent in human</td>
<td>deleted in chimp</td>
<td>aa279S→aa390Q only in baboon</td>
</tr>
<tr>
<td>CECR1</td>
<td>Absent in human</td>
<td>aa651L→736A (unique to chimpanzee only) insertion</td>
<td>absent in baboon</td>
</tr>
<tr>
<td></td>
<td>Deleted in human</td>
<td>737A→835P conserved in chimpanzee and baboon</td>
<td>737A→835P conserved in baboon and chimp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>V421→I499 unique in chimp</td>
<td></td>
</tr>
<tr>
<td>CECR2</td>
<td>Absent in human</td>
<td>732V→M759 insertion in human</td>
<td>1596V→1660D unique to chimpanzee ptb clone, absent in rp43 (Donald)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Absent in chimp</td>
<td></td>
</tr>
<tr>
<td>SLC25A18</td>
<td>Aa209G→264R (diverged from chimpanzee insertion, substitution)</td>
<td>1M→56R; 124G→179G (specific to ptb chimp)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>aa209G→220A (specific to human-insertion)</td>
<td>221G→264G(unique in chimp)</td>
<td></td>
</tr>
<tr>
<td>BCL2L13</td>
<td>1M→40Q specific to human and missing in ptb chimp(similarity starts at 41G with ptb)</td>
<td>201M→234L; 521→556T (specific to ptb clone)</td>
<td></td>
</tr>
<tr>
<td>BID</td>
<td></td>
<td>573R→592D (unique in chimp)</td>
<td></td>
</tr>
<tr>
<td>Genes</td>
<td>Human</td>
<td>Chimp</td>
<td>Baboon</td>
</tr>
<tr>
<td>----------------</td>
<td>--------------------------------------------</td>
<td>--------------------------------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>O94909(pro5)</td>
<td></td>
<td>aa 1M→163I; 966A→1089P (unique in chimp)</td>
<td></td>
</tr>
<tr>
<td>kiaa0819</td>
<td></td>
<td>aa 598 SPPPPPPPGA617 (in human)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>aa598IIIIIIIIIIII617 (in chimp)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1138G→1245A (diverged in human by insertion, substitution)</td>
<td>1138G→1245A (diverged in chimpanzee by insertion, deletion, substitution)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1246M→1256S (unique in human-insertions)</td>
<td></td>
</tr>
<tr>
<td>Q7RTP6/MICAL3</td>
<td></td>
<td>aa739D→758E; aa1066 G→1077I specific to human not in chimpanzee (insertion)</td>
<td>aa317V→386L; aa820D→827T; aa854T→944L unique to chimpanzee not in human. (insertion)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>aa 1080→1145 poor conservation between human and chimpanzee showing great divergence</td>
<td></td>
</tr>
<tr>
<td>Q7Z2D7/pex26</td>
<td></td>
<td>aa 157IILYS161 unique insertion in human and Clint</td>
<td>Absent in chimpanzee Donald. Absent in baboon</td>
</tr>
<tr>
<td></td>
<td></td>
<td>aa 9M→100T similar between human and baboon with two hydrophobic and one hydrophilic change</td>
<td>aa 9M→100T similar between human and baboon with two hydrophobic and one hydrophilic change</td>
</tr>
<tr>
<td>DGCR6L</td>
<td></td>
<td>Absent in human</td>
<td>Deleted in chimp (present in only baboon)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>aa 91L→272A (conserved in human and baboon)</td>
<td>aa 91L→272A (conserved in human and baboon)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Deleted in human</td>
<td>Deleted in chimp</td>
</tr>
<tr>
<td>PRODHIL</td>
<td></td>
<td>aa 451A→480C (specific to Donald chimp-insertion)</td>
<td>aa 451A→450Q (specific to baboon)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>aa 574M→664A (deleted in Donald chimp)</td>
<td></td>
</tr>
<tr>
<td>Genes</td>
<td>Human</td>
<td>Chimp</td>
<td>Baboon</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td>O949099(pro5) kiaa0819</td>
<td>aa 1M→163I ;966A→1089P (unique in chimp)</td>
<td>aa598IIIIIIIIIIII617 (in chimp)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>aa 598 SPPP PPPPPPGE617 (in human)</td>
<td>1138G→1245A (diverged in human by insertion, substitution)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1246M→1256S (unique in human-insertions)</td>
<td></td>
</tr>
<tr>
<td>Q7RTP6/MICAL3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>aa739D→758E; aa1066 G→1077I specific to human not in chimpanzee (insertion)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>aa 1080→1145 poor conservation between human and chimpanzee showing great divergence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q7Z2D7/pex26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>aa 157ILLYS161 unique insertion in human and Clint</td>
<td>Absent in chimpanzee Donald.</td>
<td>Absent in baboon</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q7Z2D7/pex26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGCR6L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRODHL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
As summarized in table 3.8 below and Appendix table 4, the majority of the amino acid substitutions seen in humans, chimpanzees and baboons were changes between the same types of amino acids. The majority of the changes being from hydrophilic to hydrophilic amino acids in comparison to changes from hydrophobic to hydrophobic. However, the minor amino acid substitutions in humans and chimpanzees were changes from hydrophilic to hydrophobic amino acids.
<table>
<thead>
<tr>
<th>Gene</th>
<th>human (philic) /hydrophobic (phobic)</th>
<th>chimp (philic) /hydrophobic (phobic)</th>
<th>baboon (philic) /hydrophobic (phobic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q8N7E1/XKRY</td>
<td>philic→philic</td>
<td>philic→philic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>ENSG00000138860/GRB2</td>
<td>phobic→philic</td>
<td>phobic→philic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>IL17R</td>
<td>philic→philic</td>
<td>phobic→philic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>CECR6</td>
<td>phobic→phobic</td>
<td>phobic→phobic</td>
<td>phobic→phobic</td>
</tr>
<tr>
<td>CECR5</td>
<td>phobic→philic</td>
<td>phobic→philic</td>
<td>philic→phobic</td>
</tr>
<tr>
<td>CECR1</td>
<td>philic→philic</td>
<td>philic→philic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>CECR2</td>
<td>philic→phobic</td>
<td>philic→phobic</td>
<td>phobic→philic</td>
</tr>
<tr>
<td>SLC25A18</td>
<td>phobic &amp; philic</td>
<td>phobic→phobic ;</td>
<td>philic→philic</td>
</tr>
<tr>
<td>BCL2L13</td>
<td>phobic→phobic;</td>
<td>phobic→phobic;</td>
<td>philic→philic</td>
</tr>
<tr>
<td>O94909(pro5) kiaa0819</td>
<td>philic→philic</td>
<td>phobic→phobic;</td>
<td>philic→philic</td>
</tr>
<tr>
<td>Q7RT6/MICAL3</td>
<td>philic→philic;</td>
<td>philic→philic;</td>
<td>philic→philic</td>
</tr>
<tr>
<td>Q7Z2D7/pex26</td>
<td>philic→philic</td>
<td>phobic→philic;</td>
<td>philic→phobic</td>
</tr>
<tr>
<td>Dger6(more of dger6l)</td>
<td>philic→philic</td>
<td>phobic→philic;</td>
<td>philic→philic</td>
</tr>
<tr>
<td>PRODH(more prodh)</td>
<td>phobic→phobic</td>
<td>phobic→phobic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>DGCR2</td>
<td>phobic→phobic</td>
<td>phobic→phobic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>STK22B</td>
<td>philic→philic</td>
<td>philic→philic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>SLC25A1</td>
<td>philic→philic</td>
<td>philic→philic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>MRPL40</td>
<td>philic→philic</td>
<td>philic→philic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>CDC435L</td>
<td>phobic→phobic</td>
<td>phobic→phobic</td>
<td>phobic→phobic</td>
</tr>
<tr>
<td>Q99648/CDCREL-1</td>
<td>philic→philic;</td>
<td>philic→philic;</td>
<td>philic→philic</td>
</tr>
<tr>
<td>BHB1L</td>
<td>phobic→phobic</td>
<td>phobic→phobic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>TXNRD2</td>
<td>philic→philic</td>
<td>philic→philic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>COMT</td>
<td>phobic→phobic</td>
<td>phobic→phobic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>ARVCF</td>
<td>phobic→phobic</td>
<td>phobic→philic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>Q8IZ69/HTF9C</td>
<td>philic→philic</td>
<td>philic→philic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>ZDHHC8</td>
<td>phobic→phobic</td>
<td>phobic→phobic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>ZNF74</td>
<td>philic→philic</td>
<td>philic→philic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>SCARF2</td>
<td>philic→philic</td>
<td>philic→philic</td>
<td>philic→philic</td>
</tr>
<tr>
<td>PCQAP</td>
<td>philic→phobic</td>
<td>philic→phobic</td>
<td>phobic→philic</td>
</tr>
<tr>
<td>SERPIND1</td>
<td>phobic→phobic</td>
<td>phobic→phobic</td>
<td>phobic→phobic</td>
</tr>
<tr>
<td>SNAP29</td>
<td>philic→phobic</td>
<td>phobic→phobic</td>
<td>phobic→philic</td>
</tr>
</tbody>
</table>

Table 3.8 The table shows the major trends in amino acid substitutions (hydrophilic or hydrophobic) in different genes of human, chimpanzee and baboon.
In baboons, the minor amino acid substitutions were changes from hydrophobic to hydrophilic amino acids. In bovine, a trend similar to baboon is observed in Appendix table 4 showing that the major substitutions are changes from hydrophilic to hydrophilic and the minor substitutions are changes from hydrophobic to hydrophilic amino acids. Thus the majority of the changes are conserved changes between similar types of amino acids and mainly in the hydrophilic type of amino acids which indicates that since these residues are usually exposed, they are more susceptible to evolve at a faster rate in comparison to the hydrophobic amino acids which are usually buried (Yap and Speed, 2005, although the effect of these amino acid substitutions on the biochemical properties of the encoded proteins is unknown.

3.3 Gene Expression Studies in Zebrafish

Gene expression studies of human chromosome 22 (CES-DGCR) gene orthologs in zebrafish sheds light on when and where the genes are expressed as an initial step in investigating the function of their encoded proteins. Although there have been several northern blot, RT-PCR, and microarray studies on selected genes in the ~4.5 Mbp region of human chromosome 22 (Footz et al., 2001; Strausberg et al., 2000, 2002; Fantom et al., 2005; Nagase et al., 1999, 2000; Kawai J et al., 2001; Thierry-Mieg et al., www.aceview.org; Lash et al., 2000; Matsumoto et al., 2003; Golmuntz et al., 1996; Heisterkamp et al., 1995; Galili et al., 1997; Funke et al., 1998; Maynard et al., 2003; Chieffo et al., 1997; Kochilas et al., 2003; Sirokin et al., 1997; Shiohama et al., 2003; Shmueli et al., 2003; Ishii et al., 2002; Lund et al., 2000; Berti et al., 2001; Schuler et al., 1997; Pontius et al., 2003; Wong and Cantley, 1994; Nakagawa et al.,
1996; de Jong et al., 1995; Guris et al., 2001), detailed expression information often was unavailable for most of the genes in this region. Therefore, in the present studies, the expression profiles of the genes, whose probe size and the location of the probes is listed in the Appendix table 5 and schematically placed on human chromosome 22 in figure 3.18, and figure 3.29 were investigated.

3.3.1 CES Region Gene Expression Studies

![Figure 3.18 CES region of human chromosome 22.](image)

This region represents the genes in the CES region of human chromosome 22 and the highlighted genes were studied for their expression pattern in zebrafish model system. The second gene, that is labeled NOVEL, has been assigned an Ensembl gene number of ENSG00000138860.

3.3.1.1 ENSG00000138860

ENSG00000138860 located at position 15,818,177-15,863,443 on human chromosome 22 between genes KCNMB3L and IL17R and encodes for a 639 aa protein with 63% amino acid identity to the 360 aa zebrafish chromosome 1 ortholog. The gene has a pleckstrin-like domain but has no known description of function or expression pattern. Our whole mount *in situ* studies showed expression throughout the body at basal levels and more prominent expression in the tectum, midbrain, eye,
hindbrain, and spinal chord neuron in the early stages of development in zebrafish. Decrease in expression is seen in the brain by 72hpf, with expression mainly in the retina and branchial arches.

Figure 3.19 Expression pattern of the ENSG00000138860 gene. The whole mount expression pattern of the gene ENSG00000138860 was observed throughout the body. At 22-24hpf (A) and (B) and 35hpf (C) the expression is more concentrated in the eye, tectum, midbrain and hindbrain. At 48pf (D and D-1) the expression mainly is in the brain, pectoral fin and spinal chord neuron. At 72hpf E (top view) and F (side view) the expression mainly is observed in the brain and also in the retina and branchial arches.

3.3.1.2 Interleukin 17 receptor (IL17R)
IL17R, a gene that is located at the position 15,940,412 -15,965,941 in human chromosome 22, has been shown by Northern blot analysis to be widely expressed (Footz et al., 2001) and cDNA libraries of early embryo (embryonic day 0.5-4.5-15.5,18.5, postnatal day 1.0,5.0,15.0) of mouse showed expression mainly in brain (Strausberg et al., 2002). This gene has 40% amino acid identity to the zebrafish ortholog with higher conservation in the receptor domain. The expression patterns observed in zebrafish also indicate extensive expression throughout the body, similar to that noted by others above, but in the present studies, expression seems mainly concentrated in the brain region during early stages.

**Figure 3.20 Expression pattern of the IL17R gene.** The whole mount expression pattern of gene IL17R was observed mainly in the brain, tectum, midbrain and hindbrain and the eyes, at 19-24hpf stages shown in A, B and C expression patterns, with basal levels of expression in other parts of the body. At 35hpf (D) the expression mainly is restricted to the brain and proctodeum. At 42hpf (E) and (E-1) show expression in the brain and proctodeum. At 48hpf (F) the expression mainly is restricted the brain and otic vesicle. At 60hpf (G and G-1) and 72hpf (H) the expression mainly is observed in the branchial arches with no expression in the brain.

The major expression occurs in the brain, tectum, midbrain and hindbrain and the eyes, as noticed at 19-24hpf stages, with basal levels of expression in other parts of the body. At stages 35 and 42hpf, the expression mainly is restricted to the brain and proctodeum. At 48hpf the expression mainly is
restricted in the brain and otic vesicle. During later stages (60 and 72hpf) the expression mainly is observed in the branchial arches with no expression in the brain. Since the earlier studies showed that il17r are known to be cytokine receptor belonging to the type 1 membrane protein and involved in signaling (Aggarwal et al., 2002; Strausberg R,L et al., 2002), the expression in the brain during early stages and the ubiquitous distribution may indicate that this gene is essential as an immune response regulator both in the embryos and adults (Yao et al., 1997).

**3.3.1.3 Cat Eye Syndrome Chromosome Region, Candidate 5 (CECR5)**

**CECR5**, a gene present at 15,992,955-16,020,731 in human chromosome 22, is the ortholog of a zebrafish gene located on chromosome 23. This gene encodes for a 423 aa with 55% amino acid identity to the 399 aa ortholog in zebrafish. Previous northern blot analyses were done only on adult tissues and showed that expression was present in all tissues studied. cDNA libraries constructed from postnatal mouse embryos show positive expression in brain libraries with additional expression in the retina and kidney (Footz et al., 2001; Strausberg R,L et al., 2002; Fantom et al., 2005). The present whole mount *in situ* hybridization studies in zebrafish embryos indicated expression throughout the body and major concentration in the brain. At early stages expression is throughout the body with higher expression in the brain, liver, pectoral fin and pronephric duct. But at later stages, the expression mainly is in the retina and branchial arches with reduction in expression in the brain. The level of expression in the brain is thus at its peak at early embryonic stages and decreases during later stages. CECR5 has a signal peptidase 1 S26A, domain that has been implicated in serine peptidase activity (Dalbey and Von Heijne, 1993) a function that must be required in the observed tissues at the different developmental expression times.
Figure 3.21 Expression pattern of the CECR5 gene. The whole mount expression pattern of the gene CECR5 was observed throughout the body and major concentration is in the brain. At (24hpf) A expression is throughout the body, B (35hpf), C (35hpf – dorsal view) also throughout the body with higher expression in the brain, liver, pectoral fin and pronephric duct. At 42hpf D and E (dorsal view) the expression is throughout the body with higher expression in the brain, pectoral fins and pronephric duct. At 48hpf (F) the expression mainly is in the brain, pectoral fins and pronephric duct. At 60hpf (G) the expression mainly is in the brain and pectoral fins. At 72hpf (H) the expression mainly is in the retina and branchial arches with reduction in expression in the brain. The level of expression in the brain is at its peak at early embryonic stages as seen in A-1(24hpf), B-1(35hpf), D-1(42hpf), a slight reduction can be noticed at F-1(48hpf) and reduced expression in later stages at G-1(60hpf) and H-1(72hpf) is evident.

### 3.3.1.4 Cat Eye Syndrome Chromosome Region, Candidate 2 (CECR2)

**CECR2**, present at position 16,331,181-16,408,399 in chromosome 22, has a 60% amino acid identity with its zebrafish ortholog. Northern blot analysis showed that this gene is expressed in all tissues and EST sequences from a cDNA library (kiaa1740) and mouse cDNA (embryonic day 13.5,14.5,16.5,17.5) showed elevated expression in the brain (Nagase *et al*., 2000; Kawai *J et al*., 2001) as well as cDNA in muscles, retina and lungs (Thierry-Mieg *et al*., www.aceview.org). This gene has an AT hook, a DNA
binding motif as a transcription factor, two bromodomains that are a class of regulatory proteins that mediates protein interactions necessary for transcription activation, and a homology to guanylate binding protein-1, associated with a transcription factor involved in neurulation and chromatin remodeling (Banting et al., 2004).

The present whole mount in situ studies confirmed the earlier reported elevated expression in brain but extend this observation now to include the otic vesicle, notochord, and pectoral fins. The expression pattern is seen throughout the body at 19 and 22 hpf stages. The expression is present all over the body also at stage 31hpf but higher levels of expression are seen mainly in the brain and notochord. At stages, 35, 42 and 48 hpf, the expression is all over the body with higher concentration in the brain, pectoral fin and otic vesicle. At 72hpf, the expression mainly is concentrated in the otic vesicle, retina and basal expression is seen in the brain and branchial arches.

Figure 3.22 Expression pattern of the CECR2 gene. The whole mount expression pattern of gene CECR2 in (A) 19hpf and (B) 22hpf are present throughout the body. At 31hpf (C) the expression is also present throughout the body with higher levels of expression mainly in the brain and notochord. At 35hpf
(D) and (E) (dorsal view) and 42hpf (G) the expression is throughout the body with higher concentration in the brain, pectoral fin and otic vesicle. At the 48hpf (H) the expression mainly is concentrated in the brain, otic vesicle and the pectoral fin. At the 72hpf (I) the expression mainly is concentrated in the otic vesicle, retina and light expression in the branchial arches and brain.

3.3.1.5 BCL2-like 13 apoptosis facilitator (BCL2L13)

**BCL2L13** is a gene located at the position 16,486,236-16,587,937 with 47% identity at the amino acid level to the zebrafish ortholog. Earlier Northern blot studies have shown foetal expression to be mainly in brain followed by lung, liver, and kidney, and in all tissues in the adults (Footz *et al.*, 2001).

**Figure 3.23 Expression pattern of the BCL2L13 gene.** At 19hpf (A), 22hpf (B) and 24hpf (C) the expression mainly is in the cerebral vein. At (D) 35hpf and 42hpf (E-dorsal view) the expression is more concentrated in the midbrain, hindbrain and otic vesicle. D-1, D-2 and D-3 (35hpf-42hpf) show the...
expression at higher magnification and shows the details of restricted expression in the midbrain and hindbrain region and otic vesicle. A similar expression is also shown at 48hpf (F). No expression is seen at the stage 72hpf (G).

SAGE expression studies showed expression in brain, spinal chord, and muscles (Strausberg R, L et al., 2000, 2002; Lash et al., 2000). This gene is known to encode for a apoptosis inhibitor domain belonging to the Bcl2 family (Kataoka et al., 2001). Whole mount in situ studies showed expression mainly in the forebrain and midbrain compared with the highest expression at 42hpf, and with no expression at 72hpf. In the early stages, 19-24hpf embryos showed basal level expression throughout the anterior part of the embryo with higher expression in the cerebral vein.

3.3.1.6 O94909/Kiaa0891

O94909/Kiaa0891, a chromosome 22 gene (11 exons) located at the position 16,644,972-16,689,189, has two orthologs in zebrafish, one being zf1, present at 307280-316461bp of ctg14067 and the other zf2, present at 107344-119287bp of ctg11065, with 4 and 7 exons respectively. Similarly two copies of this single copy human gene also are observed in chimpanzee.
Figure 3.24  zF1 and zF2 gene comparison to human Kiaa0819 gene. The PIP output clearly shows the conservation of specific exons in zF1 and zF2 in comparison to human. The blue boxes show the exons from which the probes were made to test the expression pattern for each of the copies of kiaa0819 human gene. zF1 probe was specifically made from exon1 unique to zF1 and human and zF2 probe was made specifically from exon11 unique to zF2 and human.

There is a unique intron 9 in human made entirely of Alu repeats that joins these two transcripts in human. To investigate this novel gene, a specific zF1 probe was made from the zebrafish exon1 that matches exon1 of the human gene and a specific zF2 probe was made from zebrafish exon4 that matched exon 11 of the human gene as shown in figure 3.24 above.

3.3.1.6.1 Expression of zF1
Figure 3.25 Expression pattern of the zf1 gene. The whole mount expression pattern of gene zf1 at 24hpf (A), 48hpf(B), 72hpf (C), 120hpf(D) shows the expression in the ear. E(48hpf) shows the three cristae and the two maculae in the otic placode.

The whole mount in situ hybridization with zf1 sense and antisense ssDNA probes indicated that zf1 the ortholog of human gene O94909 (kiaa0819) shows expression in the otic placode showing expression in the distinct sensory patches-three cristae (associated with a semicircular canal) and two maculae (associated with an otolith) in 48hpf and starts expression by 24hpf and continues until 120hpf as seen in figure 3.25. No expression was observed using the sense probe indicating an antisense, opposite orientation RNA was not presence.

3.3.1.6.2 Expression of zf2
Figure 3.26 Expression pattern of the zf2 gene. The whole mount expression pattern of zf2 in (A) at 12hpf shows expression throughout with higher expression in the brain as also seen at 24hpf (B) and 35hpf (C). During 48hpf (D), the expression is restricted to the otic vesicle and pectoral fins and a enlarged view of the 48hpf expression can be seen in D-1. At 72hpf (E) the expression mainly is in the otic vesicle.

The expression of zf2 mainly is concentrated in the brain and pectoral fin, and in the later stages was mainly in the otic vesicle, and branchial arches with nonoverlapping pattern in comparison to zf1, and no expression was seen during 24hpf in the otic vesicle as shown above in figure 3.26.

3.3.1.6.3 Expression of zf1 in (acerebellar) ace mutant embryo
Figure 3.27 Expression pattern of the zf1 gene in an ace mutant. The whole mount expression pattern of zf1 gene in an ace mutant embryo showed no expression at all in the otic placode as the placode was missing cristae and maculae but in (A) as in case of certain mutants there is one otolith instead of the regular two and the expression can be seen at that particular spot. The light blue is the background in reference to controls. B and C shows abnormal otic palcode which is small and constricted. C-1 is an enlarged view of C showing constricted otic placode.

The expression of zf1 antisense probe in the ace mutant embryo (Brand et al., 1996) as seen in figure 3.27 showed reduced expression in the otic placode when compared to the wild type counterpart that had expression in the otolith, but no expression in the mutant where both the otolith and cristae were absent. The mutants that retained one otolith showed expression of zf1. The phenotype of this mutant as described by Brand et al., 1996 is lack of midbrain-hindbrain boundary, smaller otocyst with one otolith or no otolith and semicircular canal formation is affected and of reduced size.
It should be noted that since expression of the zf1 gene was observed in the otic placode, it is likely involved in ear formation. Interestingly, since one of the abnormalities of DiGeorge Syndrome is ear malformations, the deletion of this gene may be responsible for this observed ear abnormality. Early embryological experiments suggests that in a developing ear, the ear pattern formation depends on continual interaction of the otic vesicle with the hindbrain. Of the genes involved in ear development and function that have been cloned, a whole range of molecular identities is demonstrated from signaling molecules to proteins involved in transcription elongation (Whitfeild et al., 2002; Solomon et al., 2004). Further support for this correlation is obtained from the analysis of the ace mutant. The expression is highly reduced as ace mutants have the smaller ear and missing otolith (Brand et al., 1996). This indicates that the expression of the gene O94909 (kiaa0819) is specific to the sensory patches in the ear and has a similar function to Fgf8, can be dependent to Fgf8 for its induction, or it is the transcription factor it induces that are expressed by otic placode precursor.

The expression study of the zf2 gene shows that it is expressed in the early stages mainly in the brain as seen in figure 3.26 but in the later stages expression is seen in the otic vesicle, pectoral fin, and hindbrain. The expression differences between zf1 and zf2 show that on the basis of differences in expression patterns of zf1 and zf2 orthologs of human chromosome 22 genes, both zebrafish genes apparently play independent and nonredundant roles. It is well established that the teleost genome underwent a duplication after its divergence from the mammalian lineage (Jaillon et al., 2004). This duplication resulted in 2 or more paralogous copies of genes in the
zebrafish genome, each of which is orthologous to a mammalian gene. In some instances these appear to have been maintained through divergence of function so that different sub-functions of an ancestral gene are partitioned between its duplicate descendants. This division of sub-functions may be manifest by different expression patterns (Chiang et al., 2001; Whitfield, 2002).

<table>
<thead>
<tr>
<th></th>
<th>Transcription factor binding Sites (1000bp upstream of the first exon)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUMAN</td>
<td>NIT2(2), GATA-1(3), ADR1(4), IK-2, HSF(8), CAP(2), SRY, STRE(2), CF2-II, AP-4, MZF1</td>
</tr>
<tr>
<td>ZF1</td>
<td>GATA-1(2), NIT2(2), CAP(5), HSF(15), SRY(3), IK-2, Skn-1, MATa1, c-Ets-, ADR1</td>
</tr>
<tr>
<td>ZF2</td>
<td>CREB, GATA-1(2), ADR1(4), HSF(8), CAP, GCR1</td>
</tr>
</tbody>
</table>

Table 3.9 TRANSFAC Web Based software used to identify transcription factor binding sites showing conservation of regulatory elements between human and zf1 and human and zf2.

An analysis of the regulatory elements in the upstream sequences of the human, zf1 and zf2 genes as shown in the table 3.9 above indicate that there are conserved potential binding sites of transcription factors between human and zf1 that include GATA-1 (2), NIT2(2), CAP(5), HSF(15), SRY(3), IK-2, ADR1 domains and between human and zf2 that include GATA-1(2), ADR1(4), HSF(8), and CAP domains. These upstream domains observed for zf1 and zf2 also are observed in the single upstream region of the human gene. These results indicate that the expression of the gene in human would be the cumulative expression of zf1 and zf2 that may indicate duplication from a single gene. The unique domains and transcription binding sites indicate the
divergence. The loss of regulatory modules in each of the duplicates has allowed the survival of the two genes as seen in the case of hoxb1 gene (Prince and Pickett, 2002). Therefore, in zebrafish both genes may have evolved after gene duplication in the fish lineage and present differential transcriptional control and non redundant function, showing functional diversification after the fish-specific genome duplication.

To prove this further, a detailed study must be performed to investigate whether either of these unique and specific expression domains were due to “neofunctionalization” (duplicated genes developing new function and new expression pattern), “sunfunctionalization” (duplicated genes developing different function or different expression pattern) (Lynch and Force, 2000; Ohta, 2003) or were present in the ancient gene before duplication. Further experiments in mouse embryos will help clarify this observation, as there is a lack of expression information from the recent unduplicated tetrapod outgroup.

3.3.1.7 Peroxisome biogenesis factor 26 (Q7Z2D7/PEX26)

**Q7Z2D7/PEX26**, located at the position 16,935,240-16,948,970, has a 50% amino acid identity with the zebrafish chromosome 3 ortholog. This is an integral peroxisome protein that interacts with PEX1 and PER6 complex via a direct interaction with pex6. This gene has a AT-DNA binding domain (Reeves and Nissen, 1990). Thus its main function is protein binding. Northern blot analysis showed expression in high levels in brain and also expressed in liver, kidney, and skeletal muscles (Matsumoto et al., 2003).
Figure 3.28 Expression pattern of the Pex26 gene. The whole mount expression pattern of Pex26 gene shows expression all throughout the body during early stages at 22-25hpf (A and B) but at higher level in the brain. At 35hpf (C) the expression mainly is in the blood vasculature and brain. At 48hpf (D) and at 72hpf (E) the expression mainly is in the vasculature and liver with lower level in the brain.

cDNA library EST sequences show that this gene is expressed mainly in brain (mouse embryo E12.5-15.5), kidney, pancreas, and liver (postnatal embryos), an observation confirmed by the CGAP (Cancer Genome Anatomy Project) est studies (Strausberg R.L et al., 2000, 2002). Our whole mount in situ hybridization studies showed similar expression throughout the body during early stages at 22-25hpf but at higher level in the brain. The expression mainly is in the blood vasculature and brain at 35hpf and the expression mainly is in the vasculature and liver with lower level in the brain during 48 and 72hpf.

3.3.2 DGCR Region Specific Gene Expressions
Figure 3.29 DGCR region of human chromosome 22. This region represents the genes in the DGCR region of human chromosome 22. The red boxes represent genes for which expression studies were performed. Green boxes represent genes with no single PCR products or PCR products of a significantly larger or smaller than predicted size. Yellow boxes represent genes with small exons (63-200 bases) where unique primers could not be picked successfully. Black boxes represent genes with no zebrafish orthologs. Blue boxes represent genes with only single exons in zebrafish and likely represented processed pseudogenes with the functional gene elsewhere in the genome. The NOVEL genes have been assigned Ensemble numbers ENSG00000185643, ENSG00000182364 and ENSG00000183099, respectively.

3.3.2.1 ENSG00000185643

Gene ENSG00000185643, located at position 17,394,067-17,396,638 on human chromosome 22 between genes prodh and dgcr2, encodes for a 256 amino acid protein with 48% amino acid identity to the 309 amino acid zebrafish chromosome 21 ortholog. The gene has a carbonic anhydrase domain but has no previously studied expression pattern.
Figure 3.30 Expression pattern of the ENSG00000185643 gene. (A) At 24hpf increased expression in parts of brain in comparison to the other parts also showing expression like somites/myotomes, pronephric duct and proctodeum. (B) and C (yolk sac removed) at 28hpf show expression in the brain, myotomes and pronephric duct. (D) at stage 31hpf shows more expression in the branchial arches, intestine, pectoral fin and hindbrain. (E) at 42hpf and (F) at 48hpf show increased expression in the hindbrain, branchial arches, pectoral fins and at 48hpf no expression is noticed in the pronephric duct while the same can be seen in 42hpf. (G) at 72hpf the expression is restricted to branchial arches and pectoral fin.

Our present whole mount in situ studies showed expression in midbrain, hindbrain, otic placode, pectoral fins, and myotomes at early stages of development in zebrafish. At 24hpf an increased expression is seen in parts of brain in comparison to the other parts also showing expression like in somites/myotomes, pronephric duct and proctodeum. At 28hpf expression is seen in the brain, myotomes and pronephric duct and at 31hpf expression is higher in the branchial arches, intestine, pectoral fin, and hindbrain. At stages 42hpf to 48hpf an increased expression is seen in the hindbrain, branchial arches, pectoral fins but at 48hpf no expression is noticed in the pronephric duct while the same can be seen in 42hpf. It is seen that at 72hpf the
expression is restricted to branchial arches and pectoral fin. This expression pattern indicates that this gene likely is important in early developing brain due to its higher expression in this region.

3.3.2.2 Solute carrier family 25 (mitochondrial carrier; citrate transporter), member 1) (SLC25A1)

SLC25A1 gene located at position 17,537,646-17,540,814, is an integral membrane protein that likely is a transmembrane transporter. The gene encodes for a 311 aa protein that has 74% amino acid identity to the zebrafish 304 aa ortholog. This gene has a mitochondrial carrier protein, substrate carrier, and an adenine nucleotide translocator 1 domain. Northern blot analysis showed expression in various adult tissues but not in brain but high expression levels in fetal brain, lung, liver, and kidney (Golmuntz et al., 1996; Heisterkamp et al., 1995). RT–PCR on mouse tissues extracts showed expression in brain, limb buds, lungs, liver (Galili et al., 1997; FANTOM consortium, 2005). Our whole mount in situ studies showed high expression of this gene in the midbrain and hindbrain, frontal region, and in the pectoral fins during stages 24, 48, and 72 hpf even though basal levels of expression was noticed throughout the body.
Figure 3.31 Expression pattern of the SLC25A1 gene. The whole mount expression of the gene SLC25A1 shows expression mainly in the brain. At 19-20hpf (A) the expression mainly is in the brain showing specific higher expression in different parts of the brain like forebrain, tectum, midbrain and hindbrain. At 31-35hpf (B and C) the expression mainly is in the brain, tectum, hindbrain and cranial ganglion. At 42-48hpf (D and E) the expression mainly is seen in the nose, tectum, hindbrain and cranial ganglion. At 72hpf (F) the expression mainly is in the hindbrain, retina and pectoral fin.

In the early stages, at 19-20hpf the expression mainly is in the brain showing higher expression in different parts of the brain like forebrain, tectum, midbrain and hindbrain. Expression during stages 31-35hpf mainly is in the brain, tectum, hindbrain and cranial ganglion and at 42-48hpf, the expression mainly is seen in the nose, tectum, hindbrain and cranial ganglion. During the later stages, as seen in 72hpf, the expression mainly is in the hindbrain, retina and pectoral fin.
3.3.2.3 Mitochondrial ribosomal protein L40 (MRPL40)

MRPL40, located at position 17,793,979-17,798,148 in human chromosome 22, has a bipartite nuclear localization signal domain and encodes for a 206 amino acid residue protein that has 51% amino acid identity to its 200 amino acid zebrafish ortholog. This gene has a bipartite nuclear localization signal domain. Earlier studies in E10 mouse using a RT-PCR assay with a probe size of 497bp showed expression in the brain and branchial arches in embryos and in later stages showed expression in brain, heart, lung, kidney, liver (Funke et al., 1998; FANTOM consortium, 2005). Earlier whole mount in situ hybridization in mouse showed high expression in brain and the first and second pharyngeal arches (Funke et al., 1998; Maynard et al., 2003) and our present whole mount in situ studies confirm expression in the brain, pectoral fins and branchial arches at 24, 48, and 72hpf. Although the function of this gene is unknown, it is clear that this gene is essential in the embryonic brain and branchial arches development and may be a critical gene resulting in a portion of the velocardiofacial syndrome (VCFS) phenotype.
Figure 3.32 Expression pattern of the MRPL40 gene. The whole mount expression pattern of the gene MRPL40 shows expression throughout the body as seen at 12hpf (A), 25hpf (B) and 32hpf C and D (dorsal view) with major expression in the brain, pectoral fin and branchial arches. At 42hpf E and F (dorsal view) and 48hpf G, the expression mainly is in the hindbrain and branchial arches and also pectoral fins. At 72hpf, the expression is seen only in the hindbrain and pectoral fin.

3.3.2.4 T box-1 (TBX1)

TBX1, a T-box transcription factor, is located at position 18,118,780-18,145,099 on human chromosome 22. This gene encodes for a 398 amino acid protein that has a sequence identity of 76.2% with the 460 amino acid zebrafish ortholog. This gene is a transcription factor and has a T box DNA binding domain involved in heart
formation (Flageman et al, 2005). Northern blot analysis showed Tbx1 expression in mouse during early embryogenesis is in the pharyngeal arches, pouches, and otic vesicle and later during development in the vertebral column and tooth bud (Chieffo et al., 1997). RT-PCR studies in mouse (E8.5-12.5) show expression in otic placode, otocyst, head, mandibular, and maxillary process (Raft et al., 2004).

**Figure 3.33 Expression pattern of the TBX1 gene.** The whole mount *in situ* expression of the gene TBX1 is throughout the body but concentrated mainly in the brain. At 12hpf (A), 18hpf (B), 31hpf (C and C-1) the expression is observed in the brain, but at 35hpf (D) the expression also can be seen in the mouth and heart. At 42hpf (E and E-1) the expression is seen in the brain, heart and pectoral fin. At 48hpf (F and F-1) the expression is in the mouth and branchial arches while at 72hpf (G and G-1) the expression mainly is in the mouth, retina and branchial arches, heart and proctodeum.

PCR array expression studies showed foetal and postnatal brain expression of Tbx1 (Maynard et al., 2003). Other *in situ* studies have shown expression in the heart, pharyngeal arches, and otic vesicle (Kochilas et al., 2003). In our present whole mount
in situ studies expression is seen throughout the body but concentrated mainly in the brain in the early stages but at later stages expression is also seen in the heart. At stages 12, 18, 31hpf the expression is observed mainly in the brain and at 35, 42 and 48 hpf the expression is seen in the mouth, heart, brain and pectoral fin. At 48hpf the expression is also seen in the branchial arches while at 72hpf the expression mainly is in the mouth, retina, branchial arches, heart and proctodeum.

3.3.2.5 Armadillo repeat gene deleted in velocardiofacial syndrome (ARVCF)

ARVCF, a gene located in position 18,331,975-18,378,863 on human chromosome 22 and has a 50% identity with zebrafish ortholog. This gene has a bipartite nuclear localization signal domain as well as Armadillo and HEAT elements that have been implicated in signaling and intracellular transport. Earlier studies showed that this gene is ubiquitously present in foetal and adult tissues (Sirotkin et al., 1997), as well as present in brain, heart, branchial arches, and forelimb buds (Maynard et al., 2003). SAGE analysis (CGAP- Strausberg R, L et al., 2000; Lash et al., 2000) showed expression in the brain and kidney.
Figure 3.34 Expression pattern of the ARVCF gene. The whole mount expression of the gene ARVCF mainly is present all over the body as seen at 12hpf (A), 24hpf (B), 31hpf (C) and 35hpf (D and E) and also specifically in the somites in 35hpf (D and E). At 42hpf (F) expression is seen in the pectoral fin and retina with a reduction in expression in the brain, slight increase in expression in brain is noticed at 48hpf (G), and again reduction in expression is noticed in 60hpf (H) and 72hpf (I).

Our present whole mount in situ studies show expression all over the body as seen at 12, 24, 31, and 35hpf and also specifically in the somites at 35hpf. Expression at 42hpf is seen mainly in the pectoral fin and retina. The brain expression shows a reduction in expression at 42hpf, slight increase at 48hpf, and again reduction is noticed at later stages at 60 and 72hpf.
3.3.2.6 DiGeorge syndrome critical region gene 8 (DGCR8)

**DGCR8**, located at the position 18,442,360-18,473,950 and encodes a 777 amino acid protein with 50% identity to the zebrafish ortholog on chromosome 5. DGCR8 has both a double stranded RNA binding and WW40 domain. RNA *in situ* studies show expression in the forebrain, heart, branchial arches and arterial system (Shiohama *et al*., 2003). RT-PCR studies indicated expression in the brain, heart, forelimb bud and branchial arches in the embryonic stages in mouse but in adult stages expression was observed in the brain, heart, spinal chord, liver, kidney, and spleen (Maynard *et al*., 2003; FANTOM Consortium, 2005). Our present whole mount *in situ* zebrafish embryo hybridization studies are consistent with these earlier studies as it showed expression in hindbrain, branchial arches, heart, and pectoral fins and no expression in the heart in later stages. At stages 12 and 22 hpf the expression is throughout the body with slightly higher expression in the eye, midbrain and hindbrain. At stages 31, 42 and 48hpf the expression mainly is in the heart, branchial arches and otic vesicle. But in the later stages at 60 and 72hpf the expression is no longer in the heart but in the branchial arches, otic vesicle and in the pectoral fins.
Figure 3.35 Expression pattern of the DGCR8 gene. The whole mount expression pattern of the gene DGCR8 shows expression mainly in the brain and heart during early stages of development but no expression in the heart in later stages. At 12hpf (A) and 22hpf (B) the expression is all over the body with slightly higher expression in the eye, midbrain and hindbrain. At 31hpf (C), 42hpf (D) and 48hpf (E) the expression mainly is in the heart, branchial arches and otic vesicle. But at 60hpf (F) and 72hpf (G) the expression is no longer in the heart but more in the branchial arches and also in the otic vesicle and in the pectoral fins.
3.3.2.7 HpaII tiny fragments locus 9C (Q8IZ69 /HTF9C/ NM_022727)

**Q8IZ69/HTF9C/NM_022727** gene located at the position 18,473,952-18,479,258 in human chromosome 22 has 56% amino acid identity with the zebrafish ortholog. The present whole mount *in situ* studies showed expression mainly in the somites as shown in figure 3.30 and 3.31, with 7 above the yolk cell, 10 above the yolk extension, and 13 posterior to the anus which will later gives rise to myotomes and dermatome.

**Figure 3.36 Expression pattern of the HTF9C gene.** The whole mount expression pattern of gene HTF9C showed expression mainly in the somites and developing myotomes. At 19hpf (A) and 22hpf (B), the expression is very low in the somites. At 25hpf (C) the expression shows slight increase in the somites. The peak of expression in the somites and myotomes mainly is at 48hpf as seen in D (side view), E (dorsal view), E-1 and E-2. At 72hpf (F), there is no expression seen in the myotomes and slight expression is seen in the retina.
The expression of this gene could be seen at 20-24hpf, higher expression at 48hpf and no expression at 72hpf. Numerous studies have shown that many transcription factors and signaling factors are involved in the formation of somites (van Eeden et al., 1996; Heather et al., 2000). This gene has protein domains similar to the SAM (S-adenosylmethionine-dependent methyltransferase), Uracil5-methyltransferase, and RNA binding domains. The nucleotide binding domain of the methyltransferase makes up the major domain that contributes to DNA methylation, that effects transcription factor binding and is an important epigenetic factor (Martin et al., 1999).

Figure 3.37  Expression in the somites and developing myotomes. Detailed and enlarged view of expression in the somites and developing myotomes are shown as marked by arrows.
Since DNA methyltransferases mRNA are present in differentiating somites, this gene likely is involved in similar processes in early developmental stages and depletes at later stages in zebrafish development.

3.3.2.8 Zinc finger, DHHC-type containing 8 (ZDHHC8)

**ZDHHC8**, located at the position 18,493,918-18,508,512 in human chromosome 22, has a putative transmembrane palmitoyltransferase zinc finger DHHC domain containing protein 8 precursor. This gene encodes for a 765 amino acid protein that has 56% amino acid identity to its 716 amino acid zebrafish ortholog. RT-PCR studies showed that this gene in mouse (E 10.5) is present in the branchial arch, heart, forebrain, forelimb, frontonasal region, and arterial system (Maynard et al., 2003).
Figure 3.38 Expression pattern of the ZDHHC8 gene. The expression pattern of the gene ZDHHC8 shows expression in the otic vesicle and heart showing expression in the ventricle and atrium. At 12hpf (A) and 22hpf (B) the expression of this gene is not seen. But at stages from 35-48hpf as seen in C, D, E the expression mainly is at a higher level in the heart and the otic vesicle. The enlarged picture of the heart (E-1) shows expression in the pericardial cavity, ventricle and atrium. At 72hpf (F) the expression mainly is in the otic vesicle relative to the basal expression in the brain.

SAGE and CGAP studies showed expression in the brain, spinal chord, and heart (Strausberg R,L et al., 2000; Lash et al., 2000) while other cDNA studies showed expression in the brain (Nagase et al., 1999). Our present zebrafish whole mount in situ hybridization studies confirm expression in the brain, ear, distal part of tail, branchial arches, and heart showing expression in the ventricle and atrium. At stages 12 and 22 hpf the expression of this gene is not seen. But at stages from 35-48hpf the expression mainly is at a higher level in the heart and the otic vesicle. At 72hpf, a basal expression is seen in the brain, with higher expression in the otic vesicle.
3.3. 2. 9 ENSG00000182364

Gene ENSG00000182364, located at position 18,758,069-18,802,360 on human chromosome 22, encodes for a 256 amino acid protein with 33% amino acid identity to zebrafish chromosome 2 ortholog.

Figure 3.39 Expression pattern of the ENSG00000182364 gene. (A) At 24hpf the expression is throughout the body with slightly higher and defined expression in the brain mainly eyes, tectum, hindbrain and myotomes. (B) and (C) at 31-35hpf shows concentration of expression mainly in the brain followed by branchial arches, pectoral fins, liver (primodium) and gut. (D) and (E) at 42-48hpf shows increased expression in brain and branchial arches and also expression in pectoral fins and liver. (F) At stage 72hpf, a basal level expression is noticed in the brain, with reduced expression in branchial arches, liver and gut.

The gene is expressed in all normal human tissues based on Genenote analysis (Shmueli et al., 2003) but with expression mainly in the brain. Our present whole mount in situ studies indicate that at 24hpf the expression is throughout the body with slightly higher and defined expression in the brain mainly eyes, tectum, hindbrain and myotomes. At 31-35hpf the expression mainly is in the brain followed by branchial
arches, pectoral fins, liver (primodium) and gut. At 42-48hpf there is increased expression in brain and branchial arches and also expression in pectoral fins and liver. At 72hpf, a basal level expression is noticed in the brain, with reduced expression in branchial arches, liver and gut.

3.3.2.10 ENSG00000183099

Gene ENSG00000183099, located at position 18,935,803-18,940,755 on human chromosome 22, encodes for a 256 amino acid protein with 61% amino acid identity to the 489 amino acid zebrafish chromosome 16 ortholog and encodes a gamma-glutamyl transpeptidase domain. Whole genome expression profiles in normal human tissues based on Genenote analysis (Shmueli et al., 2003) showed expression in all tissues with notable expression in the kidney, prostate, liver, lung, and pancreas. Our present whole mount in situ studies show expression throughout the embryo at 24hpf mainly in the brain region and also the myotomes and proctodeum but more defined expression at 31hpf in the pronephric duct, liver, branchial arches, brain, somites, myotomes. At 42hpf expression is seen in the brain (more in the hindbrain), branchial arches, liver and pectoral fins and at 48hpf there is an increase in expression in comparison to 42hpf. A gradual decrease in expression at 72hpf in the zebrafish embryos. This expression pattern indicates that this gene is important in early likely is involved in the formation of the kidney, lungs and liver.
Figure 3.40 Expression pattern of the ENSG00000183099 gene. At 24hpf (A), the expression is noticed mainly in the brain region and also the myotomes and proctodeum. At 31hpf (B) localized and increased expression is observed in the brain and pronephric duct. C (dorsal view) and D (side view) at 42hpf shows expression in brain (more in the hindbrain), branchial arches, liver and pectoral fins. At 48hpf E and E-1 show increase in expression in the hindbrain, branchial arches, pectoral fin, liver and gut in comparison to 42hpf. At 72hpf (F), expression is observed in blood vessel, pectoral fin and liver.

3.3.2.11 Scavenger receptor class F, member 2 (SCARF2)

SCARF2, located at position 19,103,429-19,116,700 in human chromosome 22, is known to have both a protein binding and a receptor activity. It contains EGF-like, Laminin type EGF-like, and ATP/GTP binding site motifs as well as A (P loop) and Dopamine D4 receptor domains that are known to have interactions with growth factors. This gene encodes for a 868 amino acid protein with 52% amino acid identity with the 854 amino acid zebrafish ortholog. RT-PCR in mouse indicated expression in the eyes, lungs, and urogenital organs (Lund et al., 2000; FANTOM Consortium 2005).
Northern blot analysis showed high expression in heart, and expression in the lung, ovary, and placenta (Ishii et al., 2002).

**Figure 3.41 Expression pattern of the SCARF2 gene.** The whole mount expression pattern shows expression mainly in the brain and in the later stages in the pharynx. At 12hpf (A) the expression is at its basal level and throughout the body. At 24hpf (B) the expression mainly is in the midbrain, tectum. At 32hpf (C) the expression mainly is in the retina, tectum, telencephalon, midbrain and hindbrain. At 35hpf (D) the expression is in the telencephalon, tectum, midbrain and hindbrain. At 48hpf (E) the expression is more in the anterior half of the body especially in the nose, midbrain, hindbrain, tectum, branchial arches, otic vesicle and pectoral fins. At 72hpf (F) and (ventral view, F-1) the expression mainly is in the hindbrain and pharynx.

Our present whole mount *in situ* studies also show high expression mainly in the brain and in the later stages in the pharynx. At 12hpf a basal level of expression is seen all over the body. At 24hpf the expression mainly is in the midbrain, tectum and at 32hpf the expression mainly is in the retina, tectum, telencephalon, midbrain and hindbrain. At 35hpf the expression is also in the telencephalon, tectum, midbrain and hindbrain. But at 48hpf the expression is more in the anterior half of the body.
especially in the nose, midbrain, hindbrain, tectum, branchial arches, otic vesicle and pectoral fins. By the stage of 72hpf the expression mainly is restricted in the hindbrain and pharynx.

3.3.2.12 NM_032775 (ENSG00000185214 /Q96B68/NP_116164.2/KLHL22 (Kelch-like 22))

NM_032775, located at position 19,120,360-19,174,676 in human chromosome 22 encodes for a 634 amino acid protein that is 61% identical to the 585 amino acid zebrafish ortholog and contains kelch-like and BTB /POZ (BR-C, ttk and bab) (Pox virus and Zinc finger) domains. Kelch-like domains seem to occur in genes expressed in both skeletal muscle and heart (Wu et al., 2004) although this particular gene has no known previous expression pattern. Our present whole mount in situ hybridization studies showed basal expression all over the body mainly in the epidermis and photoreceptors in the eye in the early stages of development.
Figure 3.42 Expression of the NM_032775/ENSG00000185214 gene. The whole mount expression of the ENSG00000185214 gene shows expression all over the body mainly in the epidermis especially in the eye in the early stages of development. At 22hpf (A) and 32pf (B) the expression is all over the body and the epidermal layer. At 48hpf (C) and 60hpf (D) the expression mainly is in the photoreceptor cells of the eye (mainly rods) and is more concentrated at 48hpf than at 60hpf. At 72hpf (E) the expression in the photoreceptor cells are very less compared to the earlier stages.

At 22 and 32 hpf the expression is observed throughout the body and the epidermal layer. At 48 and 60 hpf expression also is seen in the photoreceptors in the eyes, mainly in the rods. Experiments have shown that rods accumulate faster as a ventral patch initially and rods outside of this region accumulate by slowly radiating unevenly from the center of the retina (Raymond et al., 1995).
A similar distribution is observed in the expression pattern as shown in figure 3.43. Thus, as proposed for Drosophila (Cagan and Zipursky, 1992), this process likely requires cell-cell communication tempting the speculation that this NM_032775/ENSG00000185214 gene may function as a signaling molecule or be involved as one of the signal transduction components in zebrafish and may be even humans. At later stages (72hpf) the expression in the photoreceptors and epidermis is very less compared to the earlier stages.

Figure 3.43 Enlarged view of the eye photoreceptor cells. The enlarged view of the eye clearly showing expression in the photoreceptor cells (mainly rods) of the retina and no expression in the lens.
3.3.2.13 Positive cofactor 2, multiprotein complex) glutamine/Q-rich-associated protein (PCQAP)

PCQAP, is located at position 19,186,443-19,266,458 in human chromosome 22. This gene encodes for a 788 amino acid protein that has a 60% identity to the 802 amino acid protein encoded by the zebrafish chromosome 21 ortholog. This gene has three significant domains, namely a proline rich domain, an extensin like protein, and a bipartite nuclear localization signal domain.

![Figure 3.44](image)

**Figure 3.44 Expression pattern of the PCQAP gene.** The whole mount expression pattern of the gene PCQAP shows expression mainly throughout the body, but especially in the venous system, gut and pectoral fin. At 12hpf (A) and 16hpf (B) there is basal expression throughout the body but a relatively higher expression is found in the eye and anterior region of the embryo. At 22hpf (C) the expression mainly is in the eyes, midbrain and hindbrain. At 32hpf (D), 48hpf (E) and 60hpf (F) the expression mainly is in the eyes, tectum, venous system, gut and pectoral fin. At 72hpf (G) the expression mainly is in the borders of the lens, vasculature and pectoral fin.

Although mouse *in situ* hybridization experiments have shown almost ubiquitous expression, high expression levels were observed in mouse embryonic frontonasal region, limbs, and pharyngeal arches (Berti *et al.*, 2001). Unigene and
SAGE data also showed high expression in the spinal chord and brain (Schuler et al., 1997; Pontius et al., 2003; Strausberg R, L et al., 2000). Our present whole mount in situ hybridization studies show expression pattern mainly throughout the body, but especially in the venous system, gut and pectoral fin. At 12hpf-16hpf there is basal expression throughout the body but a relatively higher expression is found in the eye and anterior region of the embryo. The expression mainly is in the eyes, midbrain and hindbrain at 22hpf. At stages 32, 48 and 60 hpf, the expression mainly is in the eyes, tectum, venous system, gut and pectoral fin. In the later stages at 72hpf, the expression mainly is in the borders of the lens, vasculature and pectoral fin.

3.3.2.14 Phosphatidylinositol 4-kinase, catalytic, alpha polypeptide (PIK4CA)

PIK4CA, a member of the phosphatidylinositol 4-kinase family, is located at position 19,386,545-19,517,555 in human chromosome 22. This gene encodes for a 2044 amino acid protein that has a 70% identity to the 1427 amino acid zebrafish ortholog. This protein encodes for a basic leucine zipper transcription factor domain, a phosphatidylinositol 3- and 4-kinase catalytic domain and a bipartite nuclear localization domain.
Figure 3.45 Expression pattern of the PIK4CA gene. The whole mount expression pattern of PIK4CA gene shows expression throughout the body with higher levels of expression in the brain. At 16hpf(A) and 18hpf(B) shows a higher expression in the presumptive brain region compared to rest of the body. At 22hpf(C) the expression mainly is in the brain and also in the proctodeum and myotomes. At 32hpf(D) the expression in mainly concentrated in the brain, pectoral fins and also in the notochord and proctodeum. At 42-48hpf(E and F) the expression mainly is in the brain, notochord, and proctodeum. The expression is reduced with respect to earlier stages. At 60-72hpf(G), the expression in the brain is still reduced and expression can also be seen in the otic vesicle, pericardium and notochord.

Earlier Northern blot studies have reported high expression in the brain and moderate expression in the placenta (Wong and Cantley, 1994). Whole mount in situ studies in rodents showed high expression in fetal brains compared to adult brains (Nakagawa et al., 1996). Our present whole mount in situ studies in zebrafish confirms the expression in the brain and indicates basal levels of expression in the pectoral fins. At stages 16hpf-18hpf shows a higher expression in the presumptive brain region compared to rest of the body. At 22hpf the expression mainly is in the brain and also in the proctodeum and myotomes. At 32hpf the expression in mainly concentrated in the brain, pectoral fins and also in the notochord and proctodeum. At 42-48hpf the expression mainly is in the brain, notochord, and proctodeum and in comparison to
earlier stages, the expression is reduced. At 60-72hpf, expression can also be seen in the otic vesicle, pericardium and notochord in addition to the reduced expression in the brain. Thus this gene therefore likely is important in brain development, and it may be related to various reported brain disorders (Saito et al., 2003).

3.3.2.15 v-crk sarcoma virus CT10 oncogene homolog (avian)-like gene (CRKL)

CRKL gene, located at position 19,596,268-19,632,588 on human chromosome 22, encodes for a 303 amino acid protein with 82.1% amino acid identity to the 305 amino acid zebrafish chromosome 21 ortholog. Each gene consists of three exons and it contains SH2-SH3-SH3 domains similar to those observed in tyrosine kinases (de Jong, 1995; Uemura, 1997). RT-PCR studies in mouse showed high expression in the embryonic brain (mouse E10-E18 stages) compared to expression in muscles, kidney, liver, while in adults high levels of expression were observed in the brain and in several other tissues (Maynard et al., 2003; de Jong et al., 1995). A null mutation of this gene in mice caused craniofacial and cardiac defects (Guris et al., 2001).
Our present whole mount *in situ* hybridization studies showed expression all over the body during early stages but during late stage is restricted to the otic vesicle and pectoral fin. At stages 22-24hpf the expression is all over the body especially in the brain, notochord and choroid neural hinge and tail bud.

![Figure 3.46 Expression for pattern of the CRKL gene.](image)

The whole mount expression for CRKL shows expression throughout the body during early stages but during late stages is restricted to the otic vesicle and pectoral fin. (A) At 22-24hpf the expression is throughout especially in the brain, notochord and choroid neural hinge and tail bud. At 32hpf B (whole side view), C (dorsal view) and B-1 (side enlarged view) the expression also is throughout the body but higher in the midbrain, cerebellum, diencephalon, telencephalon, hindbrain, tectum, branchial arches and pectoral fin bud. At 35hpf D (dorsal view), D-1 (enlarged view of the tail) showing the expression in the notochord. At 42hpf (G) the expression mainly is in the hindbrain and vasculature. At 48hpf (F) the expression mainly is in the otic vesicle and pectoral fin. At 72hpf (G and G-1 side view) the expression mainly is in the otic vesicle, pectoral fin and pharynx.

At 32hpf, the expression is also all over the body but higher in midbrain, cerebellum, diencephalon, telencephalon, hindbrain, tectum, branchial arches and pectoral fin bud. At 35hpf the expression is in the notochord. At 42hpf the expression
mainly is in the hindbrain and vasculature. At 48hpf the expression mainly is in the otic vesicle and pectoral fin. At 72hpf the expression mainly is in the otic vesicle, pectoral fin and pharynx.

3.3.3 PHD finger protein 5A (Phf5a)

**PHF5A**, a PHD finger like domain containing protein, is present at position 40,180,222-40,189,191 in human chromosome 22q13.2, and has a 100% identity with the zebrafish chromosome 12 ortholog. The VISTA figure 3.47 and the pipmaker output in figure 3.48 show the high conservation of all four exons in PHF5a gene between human and zebrafish.

**Figure 3.47 VISTA output of Phf5a gene.** VISTA output shows the conservation of all four exons of Phf5a gene in all species compared.
Earlier studies in *C. elegans* (Trappe et al., 2002) and mouse (Trappe et al., 2002) indicate that this gene is essential for cell viability and morphogenetic development. Our whole mount *in situ* studies using a 280bp RNA probe in zebrafish embryos showed expression in the brain at 12hpf, 24hpf, 48hpf, with high levels of expression at 55-72hpf, and slight reduction at 96, 120, and 168hpf as seen in figure 3.49 (A, B, C, D, E, I, J, K). To further understand the importance of this gene in development, whole mount *in situ* hybridization was carried out in a recessive mutant *hi1373* (Golling et al., 2002). This mutant has a proviral insertion in intron 1 of the PHF5A gene, causing the formation of a truncated PHF5A protein. The mutants showed both phenotypic and expression changes in comparison to the wild type. The mutants have a curved body and constricted yolk sac extension as seen in figure 3.49.
(F, G, and H) compared to the wild type embryos and die within 3 to 6 days of age. Expression results in mutants with reference to wild type showed extremely reduced expression at 96, 120, and 168hpf.

Figure 3.49 Expression pattern of gene Phf5a. The whole mount expression pattern of gene Phf5a in both wild type and mutant (HI1373) zebrafish embryos. In the wild type embryos (A) at 12hpf shows expression throughout the neural plate. (B) at 24hpf and (C) at 48hpf, expression can be seen in the brain. (D) at 55hpf and (E) 72hpf shows peak level of expression in brain in comparison to earlier stages. (I) at 96hpf, (J) at 120hpf and (K) at 168hpf still shows high level of expression in brain but slightly reduced in comparison to 55 and 72hpf stages. When compared to the wild type stages I (96hpf), J (120hpf), K (168hpf), the mutants (F) at 96hpf, (G) at 120hpf and (H) at 168hpf showed both phenotypic abnormalities (constricted yolk sac, bend tail) and expression defects showing a drastic decrease in expression in the brain. Very few embryos lived past 168hpf, as the mutation was lethal in mature zebrafish.

These results show that PHF5A plays an important role in the survival of the embryo as the mutants show phenotypic defects as a consequence of reduced expression, ultimately leading to death. This gene likely could be functioning as a transcription factor or chromatin modulating or chromatin mediated transcriptional regulator, since the gene has a highly conserved PHD finger like domain in all the orthologs.

3.4 Summary
Throughout the present expression studies of human chromosome 22 CES-DGCR region orthologs in zebrafish, we observed that the majority of these genes were expressed in the brain in the early embryonic stages, and that most of the genes contain domains related to transcription factors or signaling factors. The genes with well-defined expression data in mouse also showed similar expression in zebrafish, thereby confirming the conservation of their expression during evolution.
4. Conclusions

4.1 Comparative Analysis between Different Species

Comparative analysis of the genes in the 4.5Mb region in human chromosome 22 with orthologous regions of the chimpanzee, baboon, bovine, mouse and zebrafish genomes facilitates our understanding the evolutionary relationships between species that shared a common ancestor at various times during the past 400 million years. These comparisons reveal that the distribution and types of repeats in the human, chimpanzee and baboon genomes favor the expansion of the genome with new exons, introns or UTR regions and an increased vulnerability to genetic disorders. Lineage specific repeat sequences inserted into the human genome include AluYα5 and AluYb9, while AluYg, AluYc3, AluYc5 are observed in chimpanzee, AluYα5 and AluYg occurs in baboon and bucentaur are bovine specific. Greater than 60% of the major repeats in human belong to the AluY family while they represent only 56% in the orthologous regions of chimpanzee. This observation suggests that human chromosome 22 is more susceptible to recombination than chimpanzee chromosome 22 and thus we humans are more prone to a variety of different diseases than chimpanzees. In humans, the frequency of insertions is more than that of deletions but in the case of chimpanzee there were more deletions observed than insertions while baboon was more similar to humans with more insertions than deletions. This likely indicates the influence of genomic expansion and the vulnerability for both beneficial and deleterious insertional changes. Sequence comparisons of genes sequenced from all three chimp libraries.
(Clint, ch251; Donald, rp43; and Gon, ptb) revealed that they differed from each other to the same extent as humans and chimpanzees differ.

The majority of the amino acid substitutions in human, chimpanzee, baboon and bovine were changes from hydrophilic to hydrophilic amino acids with only a few examples of changes from hydrophobic to hydrophobic amino acids. From an evolutionary perspective, these changes show the selective constraint of favoring substitutions between similar classes of amino acids for species that are closer in evolutionary time. In addition, since hydrophilic amino acid residues usually are exposed to the aqueous environment, they often evolve at a faster rate in comparison to the hydrophobic amino acids that usually are buried within the protein three-dimensional structure. Since it also is known that changes in the hydrophilic solvent exposed environment to hydrophobic residues are less destabilizing and changes in the hydrophobic core of the protein are more destabilizing (Schwehm et al., 1998). Similarly, the minor substitutions observed in human and chimpanzee are less destabilizing as they are substitutions from hydrophilic to hydrophobic and the minor substitutions in baboon and bovine that were from hydrophobic to hydrophilic residues are more destabilizing. But the vast majority of the hydrophilic and hydrophobic amino acid substitutions are neutral and are context dependent.

The nucleotide and amino acid changes observed in the codon distribution in the regions compared from human, chimpanzee and baboon showed that >60% of the changes occurred in the third position, while the first and second position changes occurred slightly more often between humans and baboons (39%) than between humans and chimpanzees (34%). However, these differences alone are not sufficient to
describe the spectrum of differences in the primate phenotypes. The substitution rate comparison between human and chimp (1.2%) and between human and baboon (2.6%) and the Ka/Ks ratio (human and chimp (0.44); human and baboon (0.48)) all point to the selective pressure or constraints influencing the coding region that results from overall purifying selection especially between human and chimp in comparison to human and baboon. The few genes that showed Ka/Ks>1 occurred in the CES region that is known for its duplication in CES patients and maintaining these mutations through positive selection may have been beneficial for the evolutionary survival of their specific gene product functions. Codon usage was more biased towards codons ending in G or C (RSCU-relative codon usage value ranging between 1-4) as their tRNA are more abundant and >70% of the nucleotide changes in primates were transitional mononucleotide substitutions, while in the bovine genome the transitional changes often involved dinucleotide substitutions.

Through comparing the genes in the 4.5 MB human chromosome 22 CES-DGCR region with chimpanzee, and evaluating the changes in the repeats, base substitution rates, amino acid codon usage and nucleotide changes, highlight the differences between the two species. Comparison with other species have revealed the evolutionary conservation and changes present in human at the coding level. However, the biological and genetic significance of these differences and the genetic basis for the CES, DGCR, schizophrenia and other genetic diseases related to this region of human chromosome 22 are yet to be determined. The effect of the observed differences in the regulation of genes will be interesting and will help to further our understanding of the genetic basis for these syndromes as recently was observed (Gilad et al., 2006).
Scanning for regions highly diverged and highly conserved (on the basis of SNP’s), determining the critical copy number of the genes present with respect to LCR’s in the entire chimpanzee chromosome 22 with respect to human chromosome 22 also would be enlightening.

4.2 Genes Postulated to be Involved in Cat-Eye Syndrome (CES)

The expression profiling of genes in the cat eye syndrome region using zebrafish model system extends the earlier studies carried out to understand this rare developmental disorder. Both known and predicted genes in this region showed interesting expression patterns, many of which matched the altered phenotype in CES patients.

<table>
<thead>
<tr>
<th>Genes</th>
<th>Human studies /northern blot /RT-PCR</th>
<th>Mouse model studies</th>
<th>Zebrafish model studies by whole mount in situ studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000138860</td>
<td></td>
<td></td>
<td>Expressed in the tectum, midbrain, eye, hindbrain and spinal chord neuron in early stages and shows decreased expression in later stages.</td>
</tr>
<tr>
<td>IL17R</td>
<td>All over the body</td>
<td></td>
<td>Expression majority in the brain and all over the embryo.</td>
</tr>
<tr>
<td>CECR5</td>
<td>All over the body</td>
<td>Expression in brain libraries with notable expression in the retina and kidney.</td>
<td>All over the body especially in the brain, pectoral fin, retina and pronephric duct</td>
</tr>
<tr>
<td>CECR2</td>
<td>EST’s sequence from a mouse cDNA library of brain, muscles, eye and lungs</td>
<td></td>
<td>Expression in the brain, ear and eye.</td>
</tr>
<tr>
<td>BCL2L13</td>
<td>In all adult and foetal tissues tested especially in brain, spinal chord and muscles.</td>
<td></td>
<td>Expression in the brain especially midbrain and hind brain, muscles in the early stages of development and after 48hpf no expression is seen.</td>
</tr>
<tr>
<td>O94909/kiaa0819</td>
<td>Expression mainly in the ear</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pex26</td>
<td>Expression in brain, liver, kidney, and skeletal muscles</td>
<td>EST’s sequence from a mouse cDNA library mainly in brain and kidney, pancreas and liver.</td>
<td>Expression in the brain, notochord, liver, otic placode and pectoral fins</td>
</tr>
</tbody>
</table>
As shown above in table 4.1, the expression patterns observed show the efficacy of using the zebrafish model to study this disease. Since Cat Eye Syndromes results in defects in the eye, ear, nose, head and neck, cardiovascular system, gastrointestinal system, various skeletal defects and urogenital system (McTaggart et al., 1998; Schinzel et al., 1981; McDermid et al., 2002), the genes that correlate to the different phenotypes with respect to their expression patterns in zebrafish are shown in table 4.2.

### Table 4.1

<table>
<thead>
<tr>
<th>Phenotypes affecting different organs in CES syndrome</th>
<th>Genes expressed in different organs based on Zebrafish whole mount in situ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eyes</td>
<td>CECR5, CECR2, ENSG00000138860</td>
</tr>
<tr>
<td>Ear</td>
<td>CECR2, O94909(kiaa0819:zf1 and zf2)</td>
</tr>
<tr>
<td>Brain</td>
<td>ENSG00000138860, il17r, CECR5, CECR2, BCL2L13</td>
</tr>
<tr>
<td>Kidney</td>
<td>CECR5 , PEX26</td>
</tr>
<tr>
<td>Liver</td>
<td>Pex26</td>
</tr>
</tbody>
</table>

### Table 4.2

In the case of the novel gene ENSG00000138860, the presence of the pleckstrin homology domain (Ingley and Hemmings, 1994) that is known to be involved in intracellular signaling, may be sensitive to dosage changes and thus be an interesting candidate gene that alters the phenotype when it is over expressed. The gene kiaa0819 with two copies (zf1 and zf2) in zebrafish also showed interesting expression. The zf1 genes in the early embryonic and late stages showed consistent expression in the ear while zf2 showed basal expression throughout the body with slightly high expression in the brain in the early stages. However, in later developmental stages it showed expression similar to zf1 in the ear with both specific expression in early stages and sub-functionalization in later stages. It also was noted...
that even though the expression in both cases is in the ear, zf1 was expressed specifically in the otolith and maculae but zf2 was expressed in the otic vesicle. The conservation of transcription factors upstream of these genes are unique to each copy but similar to the single human gene transcription factors that may indicate differential regulation. The study of zf1 gene with ace mutant defective in the fgf8 gene with no expression in ear, allows speculation of the relatedness between this gene and its dependency on fgf8 growth factor, studies that could be expanded by experiments in which the zf1 genes could be knock down and rescued with fgf8 and vice versa. It also was noticed that genes CECR2, BCL2L13, ENSG00000138860 have reduction in expression at 72hpf compared to early stages signifying the early embryonic importance of these genes. Based on the location of each of the genes and their expression pattern in early embryonic stages suggests the important role played by the genes in contributing to the anomalies seen in CES disorder and over-expression of each of the gene or gene group mainly to the distal region may be responsible for the sum total of the alterations in the CES phenotype.

4.3 Genes Postulated to be Involved in DiGeorge Syndrome

DGCR/VCFS is caused by deletions of human chromosome 22q11.2 (genes contained in 1.5 Mb or 3 Mb of 22q11) and these deletions have been intensely studied to better understand the etiology of this syndrome that is mainly categorized as causing a variety of phenotypes affecting multiple organs, including craniofacial defects, thymus hypoplasia, cardiac problems, abnormalities in neural crest migration, hearing loss, cleft palate, skeletal, limb, kidney, increased incidence of schizophrenia, growth
delays. Throughout the earlier work, DGCR has been thought of as a single gene defect, that has lead to numerous studies and arguments pointing to one gene or another in the major DGCR region as being responsible for the observed phenotype as observed in genes such as HIRA (Pizzuti, 1999; Roberts et al., 1999; 2002), COMT (Baker et al., 2005), PRODH (Paylor et al., 2001; Gogos et al., 1999), CRKL (Guris et al., 2001), Es2/dgsi (Lindsay et al., 1998), UFD1L (Pizzuti et al., 1997; Yamagishi et al., 2003), and DGCR6 (Liu et al., 2002). Studies using in situ hybridization the mouse model (Funke et al., 2001), northern blot analysis studies of 24 genes (DGCR2 to ZNF74), deletion and complementary duplication rescue studies in mice (Puech et al., 2000), have attempted to account for the observed DGCR/VCF phenotype anomalies. Results have showed that more than one gene is affected during this disease as was observed in the case of UFD1L or shared expression domains as seen in DGSI and GSCL. Recent deletion studies in mice also have shown that Tbx1 is a haplosufficient gene that may independently cause the multiple DGCR/VCF phenotype (Lindsay et al., 2001). It also should be pointed out that there is a strong correlation between the lack of the DGCR postulated genes like ZDHHC8 (Mukai et al., 2004), ARVCF (Chen et al., 2005; Sanders et al, 2005), COMT (Shifman et al., 2004; Handoko et al., 2005), UFD1L (De Luca et al., 2001), PCQAP (De Luca et al., 2003; Sandhu et al., 2004), DGCR6 and PRODH (Liu et al., 2002) and the higher incidences of schizophrenia.

Further refinement in the annotation of the human chromosome 22 has increased the number of genes spanning the critical DGCR region during the course of this dissertation research. The expression of novel genes in addition to known genes not addressed before this study results in the genes listed in table 4.3 and further
establishes the possibility of this disorder being multi-gene rather than single gene related. The idea of DGCR being a multigene syndrome previously was investigated by studying brain gene expression (Maynard et al., 2003).

### Summary of the list of DGCR genes: known (9) and unknown (4) gene expression

<table>
<thead>
<tr>
<th>Gene</th>
<th>Zebrafish whole mount <em>in situ</em> studies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ENSG00000185643</strong></td>
<td>Expression in midbrain, hindbrain, otic placode, pectoral fins, pronephric duct, branchial arches and myotomes at early stages of development in zebrafish</td>
</tr>
<tr>
<td><strong>SLC25A1</strong></td>
<td>Expression of this gene in the midbrain and hindbrain, frontal region, and in the pectoral fins during stages 24, 48 and 72hpf</td>
</tr>
<tr>
<td><strong>MRPL40</strong></td>
<td>Expression in brain and branchial arches at 24, 48, and 72hpf stages</td>
</tr>
<tr>
<td><strong>TBX1</strong></td>
<td>24hpf, 48hpf, and 72hpf zebrafish embryos show expression in the brain, heart, otic placode, swim bladder, and notochord</td>
</tr>
<tr>
<td><strong>ARVCF</strong></td>
<td>Expression in the brain and frontonasal region of the zebrafish embryos</td>
</tr>
<tr>
<td><strong>DGCR8</strong></td>
<td>Expression in hindbrain, branchial arches, heart, and pectoral fins.</td>
</tr>
<tr>
<td><strong>HTF9C</strong></td>
<td>Expression mainly in the somites, 7 above the yolk cell, 10 above the yolk extension, and 13 posterior to the anus which will later gives rise to myotomes and dermatoine. The expression of this gene could be seen at 20-24hpf, higher expression at 48hpf and no expression at 72hpf</td>
</tr>
<tr>
<td><strong>ZDHHC8</strong></td>
<td>Expression in the brain, ear, distal part of tail, branchial arches, and heart</td>
</tr>
<tr>
<td><strong>ENSG00000182364</strong></td>
<td>Expression in brain, branchial arches, thymus, pectoral fin, spinal chord and in additional to these at 24hpf, the expression is also seen in the somites and the myotomes at early stages of development in zebrafish.</td>
</tr>
<tr>
<td><strong>ENSG00000183099</strong></td>
<td>Expression all over the body at 24hpf and more defined expression at 48hpf in the pronephric duct, liver, pharyngeal arches, brain, somites, myotomes and gradual decrease at 72hpf in zebrafish.</td>
</tr>
<tr>
<td><strong>SCARF2</strong></td>
<td>Expression in brain, branchial arches, eyes and all over the body at 24hpf and 48hpf but shows dramatic reduction in 72hpf.</td>
</tr>
<tr>
<td><strong>ENSG00000185214</strong></td>
<td>Basal expression all over the embryo but mainly in the epidermis especially in the eye in the early stages of development.</td>
</tr>
<tr>
<td><strong>PCQAP</strong></td>
<td>Expression in the notochord, pectoral fins and brain</td>
</tr>
<tr>
<td><strong>PIK4CA</strong></td>
<td>Expression in the brain and indicated basal levels of expression in pectoral fins. Since this gene is important in brain development.</td>
</tr>
<tr>
<td><strong>CRKL</strong></td>
<td>Expression all over especially in the brain, notochord, muscles.</td>
</tr>
<tr>
<td><strong>Phf5a</strong></td>
<td>Expression in the brain.</td>
</tr>
</tbody>
</table>

**Table 4.3** List of genes in the DGCR region with their expression pattern in zebrafish.

In the present studies, based on the expression pattern as summarized in table 4.4 above, it is reasonable to conclude that if these genes are deleted the result would be sufficient to produce the broad spectrum of phenotypes observed in DGCR/VCF.
Phenotypes noticed in DGCR with respect to different organs | Genes studied in this dissertation using zebrafish model whole mount in situ hybridization (exon specific ssDNA probe)
---|---
Brain | ENSG00000185643, SLC25A1, MRPL40, TBX1, ARVCF, DGCR8, HTF9C, ZDHHC8, ENSG00000182364, ENSG00000183099, SCARF2, ENSG00000185214, PCQAP, PIK4CA, CRKL, Phf5a
Heart | TBX1, DGCR8, ZDHHC8
Skeleton (muscles, notochord) | ENSG00000185643, TBX1, ENSG00000182364, ENSG00000183099, PCQAP
Limb (pectoral fins) | ENSG00000185643, SLC25A1, DGCR8, ENSG00000182364, PCQAP, PIK4CA,
Thymus | ENSG00000182364
Pharyngeal arches | SCARF2, ENSG00000182364, ENSG00000183099, DGCR8, ZDHHC8
Kidney | ENSG00000183099, CECR5, ENSG00000185643,
Craniofacial (nose, ear, eyes, mouth) | ARVCF, SLC25A1, ENSG00000185214, TBX1, DGCR8, ZDHHC8, SCARF2, ENSG00000185643.

Table 4.4 List of genes grouped based on their expression in relation to the phenotypes noticed in DGCR syndrome.

Information from earlier deletions studies in mice (Df1 (ES2 to UFD1L-1.2Mb), Df2 (Es2 to T10->500Kb), Df3 (ES2 to CDCREL-1->700kb), Df4 (T10 to HIRA), and 550kb (DGCR2 to ARVCF) deletion (Lindsay et al., 2001; Puech et al., 2000; Schinke et al., 2001; Prescott et al., 2005) and the zebrafish in situ expression studies reported in this dissertation, enables further analysis of the effect of deletions in the DGCR/VCF region. The earlier deletions carried out mainly were focused on proving that the TBX1 causes cardiovascular defects (Lindsay et al., 2001; Merscher et al., 2001) entirely based on non-overlapping gene expression of genes near TBX1, in addition to the loss of function studies that were carried out to show the requirement of TBX1 for inner ear morphogenesis (Vitelli et al., 2003; Raft et al., 2003). The embryonic stages used during the experiments were limited due to constraints using the mice model. However our studies using zebrafish expression data from very early stages (one cell stage onwards) paves way to further refine these deletion studies in mice, an experiment that is not feasible in zebrafish because the deletions must be
restricted to only a few of the potential disease-related because of the very short stretches of synteny in zebrafish that are spread over different chromosomes in comparison to mice where they are collinear on chromosome 16.

Taken together, the results of our present studies indicate that the majority of the predicted genes in the DGCR critical region indeed are expressed in regions of the developing embryo that are directly related to the phenotype presented by DGCR patients. Thus, one of the major conclusions from this dissertation research is to support and extend the hypothesis that both CES and DGCR/VCFS are multigene based diseases. In the future, it is likely that additional studies aimed at determining the direct involvement of these gene in the various genetic disorders linked to this region of human chromosome 22 will require further analysis of the pathways and other interactions in which the encoded gene products are involved.
**References**


Isolation and characterization of a novel gene from the DiGeorge chromosomal region that encodes for a mediator subunit. Genomics, 74(3), 320-332.


74. Chieffo, C.; Garvey, N.; Gong, W.; Roe, B.; Zhang, G.; Silver, L.; Emanuel, B. S.; Budarf, M. L. (1997) Isolation and characterization of a gene from the


83. 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. Bio. Chem. 63, 395-438.


zebrafish ferroportin 1 identifies a conserved vertebrate iron exporter. Nature. 

403, 776-781.


151


156


247. Kataoka, T., Holler, N., Micheau, O., Martinon, F., Tinel, A., Hofmann, K.,
        Tschopp, J. (2001) Bcl-rambo, a novel Bcl-2 homologue that induces apoptosis
249. Kehrer-Sawatzki, H., Schreiner, B., Taenzer, S., Platzer, M., Muller, S.,
        Hameister, H. (2002) Molecular characterization of the pericentric inversion that
        causes differences between chimpanzee chromosome 19 and human chromosome
        16, 1019-1026.
        Genetic linkage mapping of zebrafish genes and ESTs. Genome Res. 10, 558-567.
        in a large-scale C. briggsae-C. elegans genomic alignment. Genome Res, 10,
        1115-1125.
253. Khaitovich, P., Muetzel, B., She, X., Lachmann, M., Hellmann, I., Dietzsch, J.,
        Steigele, S., Do, H.-H., Weiss, G., Enard, W., Heissig, F., Arendt, T., Nieselt-
        expression in human and chimpanzee brains. Genome Res, 14, 1462-1473.


region of conserved synteny with the human velocardiofacial syndrome region on chromosome 22q11.2. Genomics, 63(3), 378-383.


complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro. DNA Res. 7,347-355.


genomics and the origins of vertebrate chromosomes. Genome Res. 10-1890-1902.


TBX1 is required for inner ear morphogenesis. Hum. Molec. Genet. 12, 2041- 
2048.

290, 1671.


496. Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, 
Genet. 26, 225-228.

Kuroki, Y., Noguchi, H., BenKahla, A., Lehrach, H., Sudbrak, R., Kube, M., 
Taenzer, S., Galgoczy, P., Platzer, M., Scharfe, M., Nordsiek, G., Bloecker, H., 
Hellmann, I., Khaitovich, P., Paabo, S., Reinhardt, R., Zheng, J.-H., Zhang, L.- 
G.-C., Oota, S., Kitano, T., Kohara, Y., Saitou, N., Park, S.-H., Wang, Y.-S., 


