PART-II

Detailed Analysis of Xq22.1
INTRODUCTION
I. Introduction

There are certain sensitive regions in the X-chromosome carrying highly important genes namely Xq28, Xp22 and Xq22.1 which are considered as disease prone region. The size of the genomic region is just 4,200,000 bp. The genome Xq22.1 contains 46 genes, of which 40 genes protein coding genes. Out of 40 genes more than 60% of the genes are involved in various genetic disorders. For example the genome region Xq22.1 involved in numerous genetic diseases like X-linked mental retardation, Fabry disease, Mohr-Tranebjaerg syndrome, cancer etc, to add Genes involved in growth hormone regulation and bipolar disorder also being localized in Xq22.1. The pathogenesis of gangliomas, associated rare malignant progression and premature ovarian failure are located in Xq22.1.

The disease causative gene density is found maximum in Xq22.1 in human X-chromosome. The expression of common chromosomal fragile sites on human chromosomes has been proposed to be a cytogenetic expression of gene activity. Two common fragile sites expression on the human X-chromosome were observed which included the fragile site in Xq22.1. The observations made in Part I in the 13th contig of X-chromosome on which Xq22.1 is located with the maximum number of SNPs, ESTs (due to alternate splicing) could probably be a disease causing region and therefore, it was found to be of greater interest for further analysis. This has directed characterizing the genomic region of Xq22.1 by identifying 2 novel genes apart from designing drug target for autophagy.

The study of the tissue specificity of cancer genes in human X-chromosome majorly expressed in sex organs gave a clue that the cancer has been evolved only in mammals.
The study of the tissue specificity of cancer genes in human X-chromosome majorly expressed in sex organs gave a clue that the cancer has been evolved only in mammals and hence of great importance in understanding the mechanism of cancer. Almost more than 50% cancer genes are expressed in sex organs and brains. This clearly gives clue that these genes are only at mammalian level. Further mammalian-specific expansion of more than 20 rapidly-evolving genes on human chromosome Xq22.1 led to the generation of chimeric genes. Evolution events occurred within and between genes from three separate protein families Brain-Expressed X-linked [BEX], WWbp5-like X-linked [WEX] and G Protein-coupled receptor associated sorting protein [GASP], which often are expressed in mammalian brains and associated with receptor mediated signaling and apoptosis are in Xq22.1.

Hence, it is strongly presumed that the detailed analysis of this region may lead to identification or prediction of some of the novel genes and pseudogenes. This further confirmed that Xq22.1 region in Chromosome X is a major area of interest for further research to explore the secrets of nature.

2.1.1 Characterization and statistical analysis of genes on human X-chromosome, Xq22.1

In human X-Chromosome, Xq22.1 is a genomic region involved in numerous genetic diseases. Xq22.1 plays a major role in chromosomal fragile site expression. The expression of common chromosomal fragile sites on human chromosomes has been proposed to be a cytogenetic expression of gene activity. Two common fragile sites expression on the human X-chromosome were observed in females (Austin 1991). There are 20 rapidly-evolving
mammalian genes on human chromosome Xq22.1 in various protein families Brain-Expressed X-linked [BEX], WWbp5-like X-linked [WEX] and G Protein-coupled receptor associated sorting protein [GASP] (Fohn et al., 2001; Winter and Ponting 2005). These genes were identified as novel human homeobox gene ESXR1 with that of mouse Esx1 homeobox gene. ESXR1 has been localized to human Xq22.1- Xq22.3, and their expressions are restricted to placenta and testis. The findings that there are similarities between ESXR1 and Esx1, yet differences between their encoded products are consistent with the idea that placental genes evolve rapidly between mammalian species. (Seki et al., 2000). A discovery of new member of the RAB family identified through a public database search, which revealed that the gene is divided into three exons and spans approximately 7.2kb of the genome in the region Xq22.1-q22.3 in the DNA of X-chromosome. Characterization and statistical analysis is performed on protein coding genes in human genomic region, Xq22.1. would be helpful for their complete identification.

The analysis of protein coding genes were done using a combination of bioinformatics and experimental verification from Human Protein Reference Database (HPRD) on molecular weight, molecular class, molecular function, biological process, localization, expression, isoforms, proteolytic cleavage and Single Nucleotide Polymorphisms (SNPs) of 40 genes as been reported in this study(www.hprd.org). The numbers of genes are shortlisted and their proteomic analyses were done.

Autophagy is the process by which endogenous proteins and damaged organelles are destroyed intracellular. Autophagy is postulated to be essential for cell homeostatic and cell remodeling during differentiation, metamorphosis, non-apoptotic cell death and aging. Reduced levels of autophagy have been described in some malignant tumors and a role of
autophagy in controlling the unregulated cell growth linked to cancer. APG4A in Xq22.1 found to play major role in Autophagy. A detailed study on APG4A was carried out in the present investigation.

2.1.2 Drug target for APG4A

Eukaryotic cells primarily use two distinct mechanisms for large-scale degradation, the proteosome and autophagy; but only autophagy has the capacity to degrade entire organelles. The three types of autophagy are macroautophagy, microautophagy, and chaperone-mediated autophagy (Ohsumi, 2001, kljinsky and Emr 2000, Kim and Klionsky 2000, Salvador et al., 2000, Cuervo et al., 2000, Dunn 1994). Macroautophagy (hereafter called autophagy) plays an important physiological role in human health. In autophagy, a double or multi membrane-bound structure, called the autophagosome or autophagic vacuole, is formed de novo to sequester cytoplasm. Then, the vacuole membrane fuses with the lysosome to deliver the contents into the organelle lumen, where the contents are degraded and the resulting macromolecules recycled.
In the process of autophagy, formation of autophagosome has been poorly characterized so far. In mammalian cells, several groups proposed the potential source of the membrane of autophagosome as Endoplasmic Reticulum (ER), post-Golgi membrane, or yet uncharacterized organelle, phagophore. Thus, the origin of autophagosome is still controversial. Moreover, it is unknown how the double membrane structure of autophagosome is constructed. This confusing situation may be due to lack of a specific marker for autophagosomal membrane. Even in yeast cells, a specific marker protein localized on autophagosome has not been identified, although many Apg proteins have been characterized so far. Since autophagosome is a transient structure in autophagic
process, a marker molecule is prerequisite to trace the formation process. (Takayoshi et al., 1999)

Figure 65- Autophagy targeting pathway

One of the vacuolar resident hydrolases called aminopeptidase I (API), is synthesized as an inactive precursor in the cytosol, and then transported to the vacuole. This system is called the cytoplasm-to-vacuole-targeting (Cvt) pathway. It is a constitutive and biosynthetic pathway, whereas macroautophagy is a starvation-induced, degradative pathway as shown in the figure-65. Electron microscopic analyses have revealed that the Cvt pathway is similar to macroautophagy. Mutants that are defective in API maturation (cvt mutants) have been isolated. Some cvt mutants overlap with autophagy-defective (apg) and autophagy (aut) mutants, and all the original apg mutants are defective in the Cvt
pathway. Some Cvt proteins are specifically required for the Cvt pathway but not for autophagy, whereas the apg17 mutant has a specific defect in autophagy but not in the Cvt pathway. Because both pathways mostly share machinery, analysis of either pathway contributes towards understanding the other. So far, a Cvt-like system has not been found in higher eukaryotes. (Safran et al., 2002)

Two ubiquitin-like systems are required for autophagy.  

i) The ubiquitin system.  

The autophagy-defective 12 (Apg12) system. The carboxy-terminal glycine of Apg12 is activated by Apg7, an E1-like enzyme of the ubiquitin (Ub) system. Subsequently, Apg12 is transferred to Apg10, an E2-like conjugating enzyme. Finally, Apg12 forms a conjugate with Apg5 through an isopeptide bond. The Apg12–Apg5 conjugate forms a large protein complex with Apg16. Apg12 is synthesized as an active precursor and immediately forms a conjugate with Apg5. No deconjugation of Apg12–Apg5 has been observed.  

iii) The Apg8 system. Nascent Apg8 is first processed to a glycine-exposed form by a protease, Apg4. Apg8 is also activated by Apg7 (an E1) then transferred to Apg3 (an E2). Finally, Apg8 forms a conjugate with phosphatidylethanolamine (PE). Apg8–PE is deconjugated by Apg4. (Nature Reviews Molecular Cell Biology 2001).

The function of APG4 is determined by Cysteine protease required for autophagy, which cleaves the C-terminal part of MAP1LC3, GABARAPL2 or GABARAP, allowing the liberation of form I. A subpopulation of form I is subsequently converted to a smaller form (form II). Form II, with a revealed C-terminal glycine, is considered to be the phosphatidylethanolamine (PE)-conjugated form, and has the capacity for the binding to autophagosomes. Preferred substrate is GABARAPL2 followed by MAP1LC3A and
GABARAP. The enzyme regulation is inhibited by N-ethyl maleimide and the probable subcellular location is cytoplasm. There is a need of suitable ligand to induce the autophagy process to control the cancer. Hence the present study on APG4A has been selected for identifying a suitable inducer type ligand.

In the present extensive analysis has been carried out on the gene responsible for autophagy, a process of intracellular bulk degradation in which cytoplasmic components including organelles are sequestered within double-membrane vesicles that deliver the contents to the lysosome/vacuole for degradation. Reduced levels of autophagy have been described in some malignant tumors. It is observed that the APG4A gene which is recently sequenced had a wide role in suppressing various cancers. The work done includes identifying the genomic localization, tissue specificity, functions, and gene expression levels in diseased and normal cell, relation with various cancers, understanding its role in various pathways, protein structure prediction, active site prediction and screening of suitable ligand for the protein.

Exactly how Autophagy suppress tumor development is less clear. Autophagy is a membrane trafficking process leading to lysosomal degradation of cytoplasmic components, a process that is highly conserved from yeast to humans (klionsky and Emr 2000, mizushima et al., 2002). Autophagy might serve a surveillance function at the subcellular level by selectively removing damaged cytoplasmic components.
2.1.3 Identifying Novel gene and verifying experimentally using RT-PCR

The comprehensive analysis of X-chromosome (PART I) led to the discovery of an important region which may be considered as a disease causing region containing a probable gene in it. This region was found to be Xq22.1. Xq22.1 was analyzed in detail to locate a probable gene carrying segment on it which may be implicated in various disorders related to X-chromosome. The sequence of the human plant, animal kingdoms and eukaryotic genome sequencing projects are underway (http://igweb.integratedgenomics.com/GOLD/). However, there is still a real need for accurate and fast tools to find genes. Finding genes in a genomic sequence is far from being a fiddling problem (Claverie et al., 1997). The widely used and recognized approach for genome annotation consists of employing, first, homology methods, also called ‘extrinsic methods’ and, second, gene prediction methods or ‘intrinsic methods’ (Borodovsky et al.,1994; Fickett, 1996 ; Rouze, P et al., 1999; Cho. and Walbot2001 ). Indeed, it seems that only approximately half of the genes can be found by homology to other known genes or proteins. In order to determine the 50% of remaining genes, the only solution is to turn to predictive methods and to elaborate fast, accurate and reliable gene finders (Fickett 1996). Hence a new insilico approach has been developed to identify the Novel gene.

2.1.4 Novel Algorithm for identifying processed and non processed pseudogenes in Xq22.1

Mammalian genomes, such as human and mouse, contain large number of gene like sequences called ‘pseudogenes’. These pseudogenes are inheritable, non functional,
gene homologies that are generally disabled at transcriptional level (Mighell et al., 2000, Balakirev and Ayala 2003). In most cases pseudogenes cannot produce transcripts due to lack of functional promoters, very rarely, some pseudogenes have retained or acquired a functional promoter so they may be transcribed, but these transcripts are not translated due to lack of translational or splicing signal sequences. As a result of their non functionality, pseudogenes are generally released from selective pressure and often accumulate frame shifts, in frame stop codons or interspersed repeats in the original protein coding sequences. Consequently, pseudogenes can be identified operationally through finding regions of homology that have this non gene like features.

Depending on the mechanism by which they were generated, majority of the mammalian pseudogenes can be divided into duplicated pseudogenes and retro-transposed pseudogenes. Duplicated pseudogenes arose from tandem duplication or unequal crossing over, thus they often have retained the original exon intron structures of parental genes, though sometimes incompletely. Retro transposition pseudogenes arise due to the reverse transcription of mRNA transcript followed by integration into the genome (Maestre et al., 1995), Esnault et al., 2000). Therefore, retro transposed pseudogenes are often considered as a special type of retrotransposons, just like long interspersed nuclear elements (LINEs) and short interspersed elements (SINEs) in the mammalian genome (Deininger et al., 2003).

Traditionally, pseudogenes were often discovered as byproducts of gene sequencing or PCR experiments. It is only after the whole genome sequencing projects that large number of pseudogenes were identified and annotated. Using a homology based
approach, Zhang and colleagues (2003) identified more than 8000 retro transposed pseudogenes and more than 3000 duplicated Pseudogenes in the human genome draft. Ohshima and colleagues (2003) used basically the same approach in their survey except that they used an older release of the human genome.

In addition to just relying on existence of truncation or frame disruptions to ascertain the non functionality of the pseudogenes, Torrents et al., (2003) developed a neutrality test by computing the ratio of synonymous to non synonymous substitution in a DNA sequence change the amino acid and this ratio is often used to test whether a sequence is under selective constraints (Hurst 2002). These researches reported more than 20,000 potential human pseudogenes. Even though the reported numbers differ, the results from the three groups are actually consistent with each other.

525 pseudogenes or pseudogene fragments on chromosome 22 were identified by integrating several sources of pseudo gene annotations (Harrison et al., 2005). 2000 ribosomal protein (RP) pseudogenes (Zhang et al., 2002). 120 mitochondrial ribosomal protein (MRP) pseudogenes and 49 cytochrome c (cyc) are identified (Zhang and Gerstein 2003).

In addition to human and mouse number of pseudogenes were also identified in genomes of other eukaryotes including C. elegans (Harrison et al., 2001), budding yeast (Harrison et al., 2002), puffer fish (Dasilva et al., 2002), Drosophila melanogaster (Harrison et al., 2003) and vertebrate pseudogenes (Mighell et al., 2000). Some prokaryotic genomes had many pseudogenes (Ogata et al., 2001, Cole et al., 2001, Homma et al., 2002). Generally Pseudogenes are less in prokaryotes because their genomes are more compact and have
higher DNA deletion rates (Lawrence et al., 2001). In mouse it was reported to have 14,000 putative pseudogenes (Waterston et al., 2002). A more recent study says that about 5000 retro transposed pseudo gene in mouse (Zhang et al., 2004) was found based on human pseudogenes.

Pseudogenes provide snapshots of the ancient genes that were active millions of years ago. They can be analyzed to infer the evolutionary history of particular genes and gene families. Since pseudogenes are free to accumulate mutations, they are also very valuable in studying nucleotide substitution, insertion and deletions (Graur et al., 1989). For example, by comparing the sequence of human cytochrome c (cyc) pseudogenes with the functional cyc gene from human and mouse, it becomes obvious that accelerated evolution had occurred in the primate lineage leading to human (Zhang and Gerstein 2003). In other case it is found that the orthologs of a human keratin pseudo gene in the chimpanzee and gorilla are still functional (Winter et al., 2001).

The original definition of pseudogenes implies that they are transcriptional silent, over the years they have been many reported cases where a pseudo gene can indeed be transcribed. For example it was found that a tumor suppressor gene, PTEN has a transcribed retro transposed pseudo gene that has more transcripts than the parental functional gene (Fujii et al., 1999). There are more than 8000 processed Pseudogenes plus 4000 duplicated Pseudogenes in the latest Golden Path human draft genome (Birney et al., 2001).
REVIEW OF LITERATURE
II. Review of Literature

Complete literature survey on Xq22.1 was done till date to understand the research that is going on in this genomic region. Since it counts more than 1000 literatures, related and very few with significant relevance to the work, that is focused are cited below.

2.3.1 Characterization and statistical analysis of genes on human X-chromosome, Xq22.1

The genome region Xq22.1 involved in numerous genetic diseases like X-linked mental retardation, Fabry disease, Mohr-Tranebjaerg syndrome, cancer etc. Genes involved in growth hormone regulation and bipolar disorder also being localized in Xq22.1. The position of Xq22.1 is Chr X 98,200,001-102,400,000 bp. The size of the genome is 4,200,000. The genome Xq22.1 contains 46 genes, of which 40 genes protein coding genes. Pandita A et al., 2007 gave insight on the pathogenesis of gangliomas and associated rare malignant progression. The gene which is located on chromosome Xq22.1 is involved in premature ovarian failure. (Merhi et al., 2006) revealed that the premature ovarian failure is due to Xp duplication and Xq deletion. (Clarke et al., 2001) described an unreported form of congenital fiber-type disproportion that follows an X-linked inheritance pattern and demonstrated linkage to two regions of the X-chromosome, Xp22.13 to Xp11.4 and Xq13.1 to Xq22.1.

Chen CP identified antenatal diagnosis of de novo partial trisomy associated with Xq (Xq22.1→qter) and terminal Xp deletion accompanying sonographic detection of uterine
growth restriction. Frints et al., (2003) and Raynaud et al., (1998) identified a novel NXF gene cluster on Xq22.1, a good candidate gene X-linked Mental Retardation (XLMR) and classified into syndromal and nonsyndromal. Their observation confirmed the suspicion of a gene involved in growth hormone regulation being localized in Xq22.1.

Scientist identified that Fabry disease is a lysosomal disease caused as the result of the gradual accumulation of Glycosphingolipid in the alpha-galactosidase, an enzyme in Xq22-1 (Caggana et al., 1997; Knol et al., 1999; Derouiche et al., 2000; Arias Martinez et al., 2004). Eng CM et al., (1997) found an X-linked inborn error of glycosphingolipid catabolism which revealed that most mutations in the alpha-galactosidase A were private and that different substitutions of the same codon resulted in markedly different disease phenotypes. Though several mutations of the alpha galactosidase a gene have been spotted, to date no correlation has been established between the genotype and the phenotype. Studies of additional Fabry families will provide information on the nature and frequency of the mutations causing this disease as well as potential insights into the structure/function relationships of this lysosomal hydrolase.

Proteolytic enzymes through their ability to catalyze irreversible hydrolytic reactions play crucial roles in the development and maintenance of all living organisms. Proteases were initially characterized as nonspecific degradative enzymes associated with protein catabolism, but recent studies have demonstrated that they influence a wide range of cellular functions by processing multiple bioactive molecules. These essential processes initiated, regulated, or terminated by proteases include DNA replication, cell cycle progression, cell proliferation, differentiation and migration,
morphogenesis and tissue remodeling, and angiogenesis and apoptosis. An additional process in which proteolytic enzymes have also been recently implicated is autophagy.

It will be of great interest to examine the possibility that autophagins may play specific roles in tumor genesis in a similar way to that reported for other cysteine proteases, such as Unp, HAUSP, Tre-2/USP6, Dub-1, BAP1, and ubiquitin C-terminal hydrolase 1, associated with protein modification pathways that are related to those mediated by autophagins in autophagy and whose unregulated expression or activity has been linked to cancer (Liang et al., 1999).

Autophagy-related cysteine endopeptidase A (APG4A), Ras related protein Rab-40A (RAB40A), Translocase of inner mitochondrial membrane 8 homolog A (T1MM8A) located in Xq22.1 are reported as possibly implicated in cancer. Autophagy is the process by which endogenous proteins and damaged organelles are destroyed intracellularly. Autophagy is postulated to be essential for cell homeostasis and cell remodeling during differentiation, metamorphosis, Non-apoptotic cell death and aging. Reduced levels of autophagy have been described in some malignant tumors and a role for autophagy in controlling the unregulated cell growth linked to cancer. Hence the detailed analysis is accentuated in a 4.2 Mb region of the human X-chromosome, Xq22.1 and in detail on APG4A.

2.2.2 Drug target for APG4A

The knowledge of the molecular mechanisms underlying autophagy has considerably improved after the isolation and characterization of autophagy-defective mutants in the yeast Saccharomyces cerevisiae (Tsukada and Ohsumi. 1993, Thumm et
These mutants were derived from screening for starvation-sensitive yeast strains (apg mutants) or for strains defective in the degradation of specific cytosolic proteins (aut mutants). These mutants partially overlap with those isolated in genetic screens for yeast strains defective in the cytoplasm to vacuole-targeting pathway (cvt mutants), a process that shares significant morphological and mechanistic similarities with autophagy. A series of elegant studies directed to the functional characterization of these autophagy mutants has revealed that two ubiquitin-like conjugation systems are required for yeast autophagy. The first one is initiated by Apg12, a modifier protein whose C-terminal Gly residue forms a covalent isopeptide bond with a Lys residue from Apg5. This conjugation process involves an activating E11-like enzyme called Apg7 and a conjugating E2-like enzyme named Apg10. The second ubiquitin-like system requires the participation of Apg8/Aut7 synthesized as a precursor protein, which is cleaved after a Gly residue by Apg4/Aut2, a recently described cysteine proteinase (Lang et al., 1998, Kim et al., 2000). This Gly-terminal residue from the modifier Apg8/Aut7 is also activated by Apg7, but then the modifier protein is transferred to Apg3 and finally conjugated with membrane-bound phosphatidylethanolamine (PE) through an amide bond (Mizushima et al., 1998). The complex Apg8-PE is also deconjugated by the protease Apg4/Aut7, leading to the release of Apg8/Aut7 from membranes. These modification systems are essential components of the membrane rearrangement dynamics taking place during the formation of autophagosomes and execution of autophagy. Recent studies (Kabeya et al., 2000, Liang et al., 1999) have shown that these ubiquitin-like conjugation systems associated with autophagy in yeast are conserved in higher eukaryotes. In fact, proteins structurally and functionally related with
the diverse yeast Apg/Aut proteins have been described in mammalian cells, and their roles in the process of autophagy have been elucidated in some cases. However, to date, very little is known regarding the putative mammalian homologues of Apg4/Aut2, the yeast cysteine proteinase essential for the proteolytic activation, and subsequent lipidation and delipidation processes of Apg8/Aut7 (Mizushima et al., 1998, Kabeya et al., 2000).

Various studies have also shown that the process of autophagy may be of great relevance in cancer. Thus, the finding that the tumor suppressor beclin 1 (Apg6) is an inducer of autophagy has demonstrated that components of the autophagy machinery may play a fundamental role in the control of the unregulated cell growth associated with tumor development (Xue et al., 1999). Autophagy is also linked with type II (non-apoptotic) programmed cell death and may contribute to death in cells in which caspase activity is blocked (Canzian et al., 1998). These findings together with the multiple observations indicating that expression and activity of many proteolytic enzymes are profoundly deregulated in cancer suggest that specific alterations in autophagin-mediated pathways may also be linked to tumor development. The results obtained in these experiments indicate that these proteases are over expressed in some cancer cells, whereas they appear to be completely absent in other tumor cells. It is also worthwhile mentioning that the regions containing the autophagin genes are frequently altered in several human tumors (Frederick et al., 1998, Wheeler et al., 2006, Doolittle 1986). The novel gene APG4A of Homo sapiens was studied extensively via Insilco. It is believed that with a suitable target for inducing autophagy through APG4A, would be a real breakthrough in cancer therapy.
Figure 66-Subcellular surveillance model

Figure 67-Cellular Surveillance
Autophagy selectively removes damaged cytoplasmic components, in particular, damaged mitochondria, which could otherwise increase mutation rates and promote tumor initiation and progression. Cellular cancerous cells would otherwise cause tumor initiation and progression. In addition, by the orderly elimination of damaged cells, autophagy inhibits necrosis, which could promote progression by inducing angiogenesis and metastasis.

2.2.3 Identifying Novel gene and verifying experimentally using RT-PCR

Comparative genomics is an exciting new field of biological research in which the genome sequences of different species - human, mouse and a wide variety of other organisms from yeast to chimpanzees - are compared. By comparing the finished reference sequence of the human genome with genomes of other organisms, researchers can identify regions of similarity and difference. This information can help scientists better understand the structure and function of human genes and thereby develop new strategies to combat human disease. Comparative genomics also provides a powerful tool for studying evolutionary changes among organisms, helping to identify genes that are conserved among species, as well as genes that give each organism its unique characteristics. It is found that two-thirds of human genes known to be involved in cancer have counterparts in the fruit fly. More surprisingly, when scientists inserted a human gene associated with early-onset Parkinson's disease into fruit flies, they displayed symptoms similar to those seen in humans with the disorder raising the possibility the tiny insects could serve as a new model for testing therapies aimed at Parkinson's. More recently, a comparative
Gene finding typically refers to the area of computational biology that is concerned with algorithmically identifying stretches of sequence, usually genomic DNA, that are biologically functional (Pattersn et al., 2002). This especially includes protein-coding genes, but may also include other functional elements such as RNA genes and regulatory regions. Gene finding is one of the first and most important steps in understanding the genome of a species once it has been sequenced. In its earliest days, "gene finding" was based on painstaking experimentation on living cells and organisms. Statistical analysis of the rates of homologous recombination of several different genes could determine their order on a certain chromosome, and information from many such experiments could be combined to create a genetic map specifying the rough location of known genes relative to each other (Hiller et al., 2006). Today, with comprehensive genome sequence and powerful computational resources at the disposal of the research community, gene finding has been redefined as a largely computational problem. Determining that a sequence is functional should be distinguished from determining the function of the gene or its product. The latter still demands in vivo experimentation through gene knockout and other assays, although frontiers of bioinformatics research are making it increasingly possible to predict the function of a gene based on its sequence alone (Saeys et al., 2007).

As the entire genomes of many different species are sequenced, a promising direction in current research on gene finding is a comparative genomics approach. This is
based on the principle that the forces of natural selection cause genes and other functional elements undergo mutation at a slower rate than the rest of the genome, since mutations in functional elements are more likely to negatively impact the organism than mutations elsewhere (Marashi et al., 2006). Genes can thus be detected by comparing the genomes of related species to detect this evolutionary pressure for conservation. This approach was first applied to the mouse and human genomes, using programs such as SLAM, SGP and Twinscan. Comparative gene finding can also be used to project high quality annotations from one genome to another. Notable examples include Projector, Gene Wise and Gene Mapper. Such techniques now play a central role in the annotation of all genomes. (Runic 2006). Many gene prediction programs are currently publicly available. Most of them are referenced in the Web site maintained by W. Li (http://linkage.rockefeller.edu/wli/gene/). Several reviews have also been written on this topic, among which the most recent are (Claverie 1997, Guigo 1997, Haussler 1998 and Burge and Karlin 1998).

2.2.4 Novel Algorithm for identifying processed and non processed pseudogenes in Xq22.1

Pseudogenes are ‘dead’ copies of genes, which were created by genomic duplication or retrotansposons. The latter types are often referred to as ‘processed pseudogenes’ as they are processed by the LINE1 retrotansposon machinery, i.e. reverse-transcribed from a functional mRNA transcript and integrated into the nuclear genome. In general, the processed pseudogenes are ‘dead-on-arrival’ because they lack promoter sequences and cannot be transcribed (Zhang and Gerstein 2003). The presence of numerous processed pseudogenes derived from the W family of endogenous
retroviruses in the human genome. These pseudogenes are structurally collinear with the retroviral mRNA [RNA, Messenger] followed by a poly (A) tail (Pavlíček et al., 2002).

Pseudogenes are important resources in evolutionary and comparative genomics because they provide molecular records of the ancient genes that existed in the genome millions of years ago. In both mouse and human genomes, similar types of genes give rise to many processed pseudogenes. These tend to be housekeeping genes, which are highly expressed in the germ line (Zhang et al., 2004).

Pseudogenes are defunct relatives of known genes that have lost their protein-coding ability or are otherwise no longer expressed in the cell. The general structural characteristics of these processed pseudogenes include the complete lack of intervening sequences found in the functional counterparts, a poly A tract at the 3' end, and direct repeats flanking the pseudogene sequence. The processed pseudogenes that are direct copies of the corresponding mRNA resulted from RNA polymerase II transcription of the functional counterpart. Pseudogenes that are not a direct copy of the corresponding mRNA may have resulted from RNA polymerase III transcription (Vanin 1985). Pseudogenes may have some gene-like features (such as Promoters, CpG islands, and splice sites), they are nonetheless considered nonfunctional, due to their lack of protein-coding ability resulting from various genetic disablements (stop codons, frame shifts, or a lack of transcription) or their inability to function as an RNA (such as with rRNA pseudogenes). Thus the term, coined in 1977 by Jacq et al., is composed of the prefix pseudo, which means false, and the root gene, which is the central unit of molecular genetics. Pseudogenes are characterized by a combination of homology to a known gene and nonfunctionality. That is, although every pseudogene has a similar DNA sequence to some functional gene, they are
nonetheless unable to produce functional final products (no functionality). (Mighell and Smith, et al., 2000).

In higher eukaryotes, particularly mammals, retrotransposition is a fairly common event that has had a huge impact on the composition of the genome. For example, 30% - 44% of the human genome consists of repetitive elements such as SINEs and LINEs. In the process of retrotransposition, a portion of the mRNA transcript of a gene is spontaneously reverse transcribed back into DNA and inserted into chromosomal DNA. (Jurka 2004; Dewannieux and Heidmann 2005; Esnault, et al., 2003). Once these pseudogenes are inserted back into the genome, they usually contain spliced-out introns and a Poly-A tail, two hallmarks features of cDNAs. However, because they are derived from a mature mRNA product, processed pseudogenes also lack the upstream promoters of normal genes; thus, they are considered "dead on arrival", becoming non-functional pseudogenes immediately upon the retro transposition event. (Graur and Shualii1989) A further characteristic of processed pseudogenes is common truncation of the 5' end relative to the parent sequence, which is a result of the relatively non-processive retro transposition mechanism that creates processed pseudogenes (Pavlicek and Paces 2002).

A copy of a functional gene may arise as a result of a gene duplication event and subsequently acquire mutations that cause it to become nonfunctional. Duplicated pseudogenes usually have all the same characteristics of genes, including an intact exon-intron structure and promoter sequences. The loss of a duplicated gene's functionality usually has little effect on an organism's fitness. (http://www.talkorigins.org/faqs/molgen/). (Edward E. Max. 2003).
In normal gene function (Fig 68), DNA is transcribed into RNA, which is then "processed" by the removal of introns (the non-coding sequences between the gray boxes) and addition of a poly (A) tail. The mature processed RNA is then translated into a chain of amino acids to form a protein. The right panels illustrate the two pathways generating the classical duplicated pseudogene (top) and processed pseudogene (bottom). In the top pathway, DNA duplication generates two copies of the entire gene (upper right box), but mutations in one copy (represented by the "x"s) render it a pseudogene. In the other pathway a processed RNA transcript of a gene can become reverse transcribed into a cDNA copy (lower right box) that inserts back into cell's DNA at a random position in the genome, usually—as shown here—in the spacer DNA between genes (white boxes in the
Figure), Disabled genes or unitary pseudogenes. Various mutations can stop a gene from being successfully transcribed or translated, and a gene may become nonfunctional or deactivated if such a mutation becomes fixed in the population (Nishikimi et al., 1992). Processed pseudogenes often pose a problem for gene prediction programs, often being misidentified as real genes or exons'. It has been proposed that identification of processed pseudogenes can help improve the accuracy of gene prediction methods. (Van Baren and Brent 2006). It has also been shown that the parent sequences that give rise to processed pseudogenes lose their coding potential faster than those giving rise to non-processed pseudogenes (Zheng et al., 2007). In 2003, Hirotsune et al. identified a retro transposed pseudogene whose transcript purportedly plays a trans-regulatory role in the expression of its homologous gene, Makorin1, and suggested this as a general model under which pseudogenes may play an important in biological role. The first pseudogene was reported in 1977 (Jacq et al., 1977). Pseudogenes are DNA sequences that resemble functional genes but are generally thought to have no purpose (Max and Edwards, 2003). Processed genes are found on different chromosomes from their functional counterparts. They lack introns and certain regulator genes, often terminate in adenine series, and are flanked by direct repeats (which are associated with movable genetic elements). They may be complete or incomplete copies of genes or mixtures of several genes. They are believed to have occurred through a 3-step process: Copying DNA into RNA, editing the introns to make mRNA, and then turning the code in the mRNA back into DNA through a reverse transcription process. This process is thought to have created the "L1 family of pseudogenes". Unprocessed pseudogenes are usually found in clusters of similar functional sequences on the same chromosome. Their expression is usually prevented by a
"misplaced" stop codon or codons. There may be other changes from the "original" as the result of deletions, insertions, and point mutations. Many of these are believed to have arisen by gene duplication, which produced an extra copy of the gene. The extra copy could then accumulate mutations without harming the organism since it would still have a completely functional original copy. (Gibson 1994).

(http://www.detectingdesign.com/pseudogenes.html).
MATERIALS AND METHODS
III. Materials and methods

Detailed analysis of Xq22.1 using bioinformatics approach shopped around various Bioinformatics database and tools. For each and every analysis a new protocol was developed to get a clear answer to the questions being asked in various parameters. The various integrated protocols for novel gene identification, designing and algorithm developed to find the pseudogenes have been the form of flow chart.

2.3.1 Characterization and statistical analysis of genes on human X-chromosome, Xq22.1

This work was mainly focused on protein coding genes present in Xq22.1 (40 genes). The analysis of protein coding genes in Xq22.1 was based on following parameters which is shown in the figure-69

Proteomic Analysis

- Molecular Weight
- Molecular class
- Molecular Function
- Biological Processes
- Disease association
- Localization
Expression

Post Translational Modification (Proteolytic cleavage)

The above parameters are extracted from Human Protein reference Database (HPRD)

Further analysis of APG4A

- Isoform analysis
- SNP analysis and validations

This data is taken from db SNP.

Figure 69-Proteomics and SNP analysis of genes present in Xq22.1
2.3.2 Drug target for APG4A

Sequence retrieval:

The amino acid sequence of the APG4A of Homo Sapiens was retrieved from the Swiss-Prot database (entry code Q8WYN0). In this sequence, the point mutation occurs at Cysteine 77 residue.

Functional Prediction and pairwise alignment:

Basic Local Alignment Search Tool (BLAST) is used to infer the function of an uncharacterized biological query sequence by detecting similarity with sequences of known functions. Comparisons are made in a pair wise fashion. Each comparison is given a score reflecting the degree of similarity between the query and the sequence being compared. The higher the score, greater is the degree of similarity. The similarity is measured and shown by aligning two sequences. Alignments can be global or local (algorithm specific). A global alignment is an optimal alignment that includes all characters from each sequence, whereas a local alignment is an optimal alignment that includes only the most similar local region or regions. Generally, the BLAST takes the entire query sequence and compares it against database sequences. It is believed that if the sequence identity between the two aligned sequences (query and database sequences) falls below 20-25%, the function of the query sequence can no longer be reliably inferred (Doolittle 1986). Thus ATG4B was considered for comparison since it is the only well annotated sequence with experimentally determined three-dimensional structure.
Identification of families and multiple sequence alignment using Pfam:

Pfam is a database of protein families (Bateman et al., 2000). The query sequence was submitted to the Pfam database in order to find its families. Cysteine protease family22 members were recognized by retrieving the cross reference of swiss-prot database. The multiple sequence alignment was carried out against the same family members to identify the conserved amino acids using the clustalX program, with default parameters (Jeanmougin et al., 1998). In Multiple sequence Alignment (MSA) the sequence of the members were aligned optimally by bringing the similar or identical characters into the same column of the alignment.

Domain and its nomenclature:

Domain refers to a fragment of a polypeptide chain that can fold into a three-dimensional structure, irrespective of the presence of other fragments of the chain. The Conserved Domain Search (CD-Search), a web based tool for detection of structural and functional domains in protein sequences, was utilized to identify domains, if any, in the query sequence (Bauer and Bryant 2004). The identified domains were then designated according to their pore forming residues.

Tertiary structure prediction:

In homology modeling, a three dimensional (3D) structural model for a protein of known structure is elucidated. Modeling is desirable when the structure of a protein is not elucidated by either X-ray crystallography or NMR spectroscopy. The model facilitates to study the protein function even up to atom level. Comparative modeling method consists of five sequential steps (Greer 1990). The first step is to search for
proteins with known 3D structures that are related to the target sequence using PSI-BLAST. The second step is to pickup those structures that could be used as templates. The third step is to align their sequences with the target sequence. An accurate alignment can be calculated when the target-template sequence identity is above 40%. If the target-template sequence identity is lower than 40%, the alignment generally has gaps and needs manual intervention to minimize the number of misaligned residues. The fourth step is to build the model for the target sequence given its alignment with the template structures. The last step is to evaluate the model using a variety of criteria, if necessary template selection, alignment and model building can be repeated until a satisfactory model is obtained.

For construction of 3D model of APG4A protein of Homo sapiens, the query sequence was submitted to BLAST-P. The comparative model finds a suitable template for the query sequence. The theoretical model was constructed using the program MODELLER8V1 (Fiser and Sali 2003). MODELLER8V1 is computer program that models 3D structure of proteins and their assemblies by satisfying spatial restraints. To run the MODELLER8V1 program three input files namely, (i) atom file, (ii) alignment file and (iii) a script file are required. The atom file is a PDB structure file; the sequence alignment file consists of aligned sequence between target and template, and the script file contains all the commands to execute the program. The aligned sequence of the APG4A of Homo sapiens and APG4B was adjusted manually to improve the stereochemistry of the model using Swiss PDB viewer (Guex and Peitsch 1997). The methodology is shown in figure-70
Drug screening of APG4A

Model validation:

The constructed model of the APG4A protein was evaluated for its stereochemical properties using Ramachandran plot (Ramachandran et al. 1963). The SAVS (Structure Analysis and Validation Server) available online was used to check the favorable and unfavorable properties of the model structure and the model was verified for errors by using PROSA (Sippl 1993) before it was deposited in protein data bank (PDB).
Active site and pockets:

PASS (Putative Active Site with Spheres) is a simple computational tool that uses geometry to characterize regions of buried volume in protein and to identify positions likely to represent binding sites based upon the size, shape and burial extent of these volumes (Brady Jr and Stouten PF 2000). The CASTp is an online server, which is used to predict pockets in proteins. The pocket is defined as an empty space in the protein and is not accessible to any solvent (water molecules) (Andrew et al., 2003). The active sites and pockets in the APG4A protein were predicted using the PASS and CASTp server respectively.

Virtual screening:

Virtual screening is a theoretical method for the prediction of binding affinities between small ligands and receptor protein. Three-dimensional model of the APG4A was used as the receptor protein.

As there is no heteroatom available in template protein (2cy7), the template sequence against the PDB database was blasted and the hits like 1H0H and 2FNO were found, which show considerable similarity particularly in the conserved domain region. Fortunately it is found that ligand ion “Fe” in 1H0H show very optimal interaction with cysteine residue. The active site of our target protein was also Cysteine residue. The fes4 was sketched in JME editor in NCI database (Ihlenfeldt et al. 2002) and found NSC667220 as the suitable hit for this protein.

Automated docking:

Generation of Receptor-Ligand complex:
The calculations of binding energy using Auto Dock, polar hydrogen's were added to the receptor with the PROTONATE utility from AMBER. AMBER united atom force field charges were assigned, and solvating parameters were added using the ADDSOL utility. The 3D affinity grid fields were created using the auxiliary program Auto Grid. The residue Cys 77 was chosen as the grid center. In this stage, the protein was embedded in the 3D grid and a probe atom was placed at each grid point. The affinity and electrostatic potential grid was calculated for each type of atom in the ligand molecule. The number of grid points in x, y, z-axis was 60 x 60 x 60 with grid points separated by 0.375 Å resolution. The grid parameter file was shown below.

Ligands that had a peptide-like N- or C-terminal end were assigned a charge. Hydrogen atoms were added to fill all empty valences, and Kollman united-atom charges were also assigned to the ligands. Rotatable dihedral in the ligands were assigned using the program AutoTors and were allowed to rotate freely. The nonpolar hydrogens were removed and the partial charges from these were added to the carbon that held the hydrogen. The atom type for the aromatic carbons was reassigned to be handled by the aromatic carbon grid map. These preparations were done for the ligand using the AutoTors module.

The initial population during docking was random ligand conformations in random orientations and at random translations. Each docking experiment was derived from 100 different runs that were set to terminate after a maximum of 2500,000 energy evaluations or 27,000 generations, yielding 10 docked conformations. A pseudo-Solis and Wets local search was then used to minimize energy of the population. The probability that docking solution in the population would undergo a local search was set to 0.06 and the
constraint was set to a maximum of 300 iterations per search. The maximum numbers of success or failures before changing the size of local search space were both set to 4. The starting conformations of the ligand were set to random positions. Translations were set to have a maximum limit of 2 Å/step and the orientation step size for the angular component and the torsions had a maximum limit at 50 degrees/step. At the end of a docking job with multiple runs, the program performed cluster analysis. The solutions with ligand all-atom RMSDs within 1.0 Å of each other were clustered together and ranked by the lowest energy representative. The lowest-energy solution of the lowest ligand all-atom RMSD cluster was accepted as the calculated binding energy.

In this work extensive analysis was done on gene responsible for autophagy, a process of intracellular bulk degradation in which cytoplasmic components including organelles are sequestered within double-membrane vesicles that deliver the contents to the lysosome/vacuole for degradation. Reduced levels of autophagy have been described in some malignant tumors. It is found that APG4A which is recently sequenced has a wide role in suppressing various cancers. This work includes identifying the genomic localization, tissue specificity, functions, gene expression level in diseased and normal cell, relation with various cancers, understanding its role in various pathways, protein structure prediction, active site prediction and screening of suitable ligand for the protein. The probable active site was identified as Cys 77 and suitable ligand was identified as 1, 2-dicarboxyethanethiolate, which showed considerable interaction with Cysteine residue. Various analogs of the original ligand were also selected and docked with the receptor. The binding energy of those analogs also increased the confidence of using those ligands as
candidate for this particular protein. Further extensive research work on this protein will leads to a new milestone in Autophagy research which is as shown in the figure-71.

![Figure 71-Model validation of APG4A gene](image)

### 2.3.3 Identifying Novel gene and verifying experimentally using RT-PCR.

Comparative analysis:

Repeat masked genomic segments were aligned with chimp genome using BLAT and mouse, rat and fugu genomes using BLASTN tool provided by NCBI. Homologous stretches between the genomes were then aligned at the protein level after six-frame translation using TBLASTX. Conserved regions of high protein-coding potential were distinguished from conserved non-coding regions based on continuity and identity of
the alignment. Protein stretches greater than 20 amino acids and possessing an identity of > 75% identities were considered further. These thresholds were not as rigid when evidence of transcription for the region was observed. Adjacent regions with coding potential were then concatenated to identify the presence of any conserved domains as well as to identify similar protein sequences in the databases. SMART and Pfam were used for domain analysis. Steps up to the TBLASTX were largely automated by a tool developed in-house, but the selection of putative protein coding regions was purely manual. The same criteria were followed for identification of novel coding exons as for the known protein coding genes. All the novel predictions were tested by PCR analysis of pooled cDNAs as described below.

Gene prediction program:

Xq22.1 region was broken into segments. The segments representing already known genes were eliminated from analysis and the remaining genomic segments were then selectively screened using various gene prediction tools such as MCG genes, Geneid, Genescan and SGP. The exons with no known genes and with probability of containing only one gene were selected for further analysis. Corresponding to these genomic segments, exon sequences were found out from the NCBI website and divided into 9 subsets. Blast program was used for comparing human exon sequence with mouse genome. These subsets of exon sequences were subjected to Blast analysis for comparison with the mouse genome under the category of reference sequence protein, non-reference sequence protein and build protein. Protein accession number, its reading frame, the name of the protein and percentage homology with mouse genome could be obtained using Blast as shown in the figure-72.
Bioinformatics approach for functional annotation of the predicted novel gene.

The results were confirmed with TBLastX at protein level and followed by Blast2. This enabled to find out the region of sequence similarity between query sequences of 2 different genomes, which yielded clues about structure and function of novel sequences. The results obtained using Blast 2 was further verified with another gene prediction tool called Augustus. It is a more recent tool and helps to confirm the Blast results and also to
find the frame in which the gene exists. It was found that the gene was in the negative frame/strand. The localization was predicted to be nucleus by using PSORT. Domain analysis of the predicted protein was done by smart. Electronic northern blot was done from gene Atlas. Homology with the structure of the predicted protein enabled the selection of the template using CPH model server 2.

Structure was elucidated and quality factor of the structure was checked using verify #D, PROSO. The probability for localization in nucleus was obtained from Xrcc6 human protein which consists of 608 amino acids and was found to be on chromosome 22q11-q13. SMART predicted the functional domain as Ku78 which was found to be a dimer bound to DNA for repair mechanism. Inter actin network and structure elucidation revealed that it was similar to XRcc6 mouse protein. The gene was validated by experimental method by designing appropriate primers for the predicted gene.

Using bioinformatics tools and Novel Insilco approaches, a novel gene could be predicted and also, the amino acid sequence of the protein thought to be encoded by it could be retrieved from Augustus. This information was further utilized to model the protein even before the results could be obtained using the biotechnological techniques. The results obtained were subjected for the experimental validation which reassured that the predicted protein was a nuclear protein figure-73
Functional annotation of protein

Gene and protein sequence was predicted by using Augustus

Highlighting the ORF sequence

Labelled start and stop codons

Localization of gene using PSORT

Domain analysis using SMART

Electronic northern blot analysis

template structure was selected using CPH model server 2

Elucidating protein-protein interaction

Structure of the gene was predicted

Validation of novel gene using PROCHECK, WHATCHECK, ERRAT, VERIFY3D, prove

Homology modeling

Overall quality factor of model structure was found out

Figure 73-Functional annotation of the novel protein identified
RT PCR analysis:

Verifying the results by RT-PCR:

Bioinformatics tools gave the amino acid sequence as well as the homology model for the predicted protein. This was further supported by homology with Xrcc6 human protein that suggested a nuclear localization for predicted protein with a role in DNA repair functions. All these results were finally confirmed in the laboratory using RT-PCR and other protocols. Illumination of the gel with UV light showed that the predicted protein expressed only when the cells were irradiated and DNA was damaged. The steps followed were

Lymphocyte isolation:

Lymphocytes were isolated from peripheral blood obtained from young (23-25 years), non-smoking men by centrifugation in a density gradient of Gradisol L (15 min, 280g, 4°C). The viability of the cells was measured by trypan blue exclusion assay and was found to be about 99%. The final concentration of the lymphocytes was adjusted to 1-3x10⁵ cells/ml by adding RPMI 1640 to the single cell suspension.

Irradiation:

In the in vitro study, the lymphocytes in culture were irradiated with 2 Gy of gamma rays which was generated using a 60Co source (Theratron Eldorado 6) at a dose rate of 0.5 Gy/min as determined by calibration using an air ionization chamber (Capinter PR-06C) connected to an electrometer (Capintex 192X)

RNA extraction and reverse transcriptase polymerase chain reaction (RT-PCR):
Total RNA was extracted and isolated from the cells (unirradiated- control and
irradiated – experimental) using TRI REAGENT LS (Sigma) according to the instruction of
the supplier. The RNA pellet was dissolved in diethyl pyrocarbonate (DEPC) treated water
and the concentration and purity of the samples were assessed spectrophotometrically at
260 and 280nm. One microgram of RNA was reverse transcribed using the PowerScript
Reverse Transcriptase (Clonetech) according to the Clontech Laboratories Protocol, using a
combination of oligo (dT) 15 primers and random hexamers. The number of PCR cycles
was calibrated for each transcript to target the exponential phase of the reaction. The PCR
reaction were carried out using 1U of Taq DNA polymerase (Geneid) as follows: 5 min
denaturation at 95°C, followed by the predetermined number of cycles of 30s at 95°C, 40s
at 55°C, 1 min at 72°C. At final extension was carried out for 5 min at 72°C. Primers were
(Gene Bank No.) : F CAGGGACTCCTCTGGGTACAR, AND CGTCAGATTGTGCTGGAGAA. PCR
products were separated on 1.5% agarose gel in TAE buffer and stained with ethidium
bromide. The image was captured under UV light and the intensities of the bands were
analyzed and quantified by gel documentation system (Lark Innovative) which is shown in
the figure-74.
2.3.4 Novel Algorithm for identifying processed and non processed pseudogenes in Xq22.1

From NCBI web page the genomic region of Xq22.1 [98M - 102M] were downloaded. Each one is split into fragments of 100kb each using Perl programming as shown in the figure-75. Since Pseudogenes share some of the common characteristics of LINES and SINEs, which include completely lack of introns, the presence of small flanking direct repeats, the splited fragments are masked with the help of repeat masker using (http://www.girinst.org/Censor_Server Data_Entry_Forms.html) (Jurka et al., 2005). The masked sequence is subjected to BLASTN (Altschul et al., 1990) against NR
database. The assignment of function to a gene found more accurate and is often implied from that of a gene with a homologous sequence at protein level in genome annotation. Hence Using BLASTX the closest human protein homolog was shortlisted. Then each sequence is realigned with homologous protein using the TFASTY program (with default parameter) of the FASTA package (W. R. Pearson et al., 1998).

The matched regions are potential gene homolog and are termed pseudogenes and those without a match were discarded. In these analyses, genes from annotated genomes and protein databases have first been clustered into paralog families and then used to survey whole genomes for copies or homolog. To categories these Pseudogenes, search of 1,000 nucleotide region on upstream of the genome for candidate AATAAA polyadenylation signals and existence of a polyadenine tail of at least 15 - 20 nucleotides were done, in addition to 70 percent identity with the gene on TFASTY Program (Pearson and Lipman, 1988). If the conditions were satisfied then it is termed as processed pseudogenes) and the rest were termed as non processed pseudo genes and these non processed pseudogenes were further confirmed with gene prediction tool using Genome Scan (Burge, C. and Karlin, S. 1997, InterPro Nicola). Mulder et al., 2007) Motif assignments and consequently GO class assignments. The results were negative or smaller for the non-processed pseudogenes in comparison with the gene totals and those for processed pseudogenes as shown in the figure-76.
Figure 75-Splitting of Xq22.1 genomic region

DOWNLOAD Xq22.1 GENOMIC SEQUENCE FROM NCBI

EACH OF THESE GENOMIC SEQUENCE WAS SPLIT INTO FRAGMENTS USING PERL PROGRAMMING

SPLITED FRAGMENTS ARE MASKED BY REPEAT MASKER

THE MASKED SEQUENCE WAS SUBJECTED TO BLASTN AGAINST NR DATABASE

MATCHED REGIONS ARE POTENTIAL GENES (PSEUDGENES)

EACH SEQUENCE RE-ALIGNED WITH HOMOLOGOUS PROTEIN USING THE TFASTY

USING BLASTX THE CLOSEST HUMAN PROTEIN HOMOLOG SHORTLISTED

HOMOLOGOUS SEQUENCE OF THE GENE ANNOTATED

THOSE WITHOUT MATCH WERE DISCARDED

ANNOTATED GENES OF PROTEIN DATABASE CLUSTERED INTO PARALOG FAMILIES

PARALOG FAMILIES USED TO SURVEY WHOLE GENOMES FOR COPIES OR HOMOLOG

PSEUDGENES ARE CATAGORISED BY

NON-PROCESSED PSEUDGENES CONFIRMED BY GENOMESCAN, INTERPRO

IF THE CONDITION SATISFIED, IT IS CALLED PROCESSED PSEUDGENES

70 PERCENT IDENTITY WITH THE GENE ON TFASTY

SEARCHING NUCLEOTIDE REGION ON UPSTREAM OF GENOME FOR POLYADENYLATION SIGNAL AND POLY A TAIL

RESULTS WERE NEGATIVE OR SMALLER FOR NON-PROCESSED PSEUDGENES IN COMPARISON WITH PROCESSED PSEUDGENES

Figure 76-Novel algorithm for identifying processed and non-processed (classical) pseudogenes
RESULTS
IV. Results

The topic covered in this chapter includes detailed proteomic analysis of Xq22.1 and alternate splicing events of all the protein coding genes of this region. An investigation on APG4A of Xq22.1 from gene to drug discovery, identification of novel protein coding genes and novel coding exons in Xq22.1 by manual curation as well as gene prediction tools using comparative genomics approach has been carried out and the results are shown in the four of graphs and figure.

Functional annotations of the predicted protein using bioinformatics approach to facilitate the experimental validation were carried out and finally confirm the presence of these novel genes. The results are also shown which are related in the identification of the Pseudo genes on Xq22.1

2.4.1 Characterization and statistical analysis of genes on human X-chromosome, Xq22.1

The analysis shows that GPRASP1 (G Protein-coupled receptor associated sorting protein 1) has a highest molecular weight (156845 Da) by screening a cDNA library for cDNAs with the potential to encode large proteins, a full-length cDNA encoding KIAA0443 called GASP. The predicted protein contains 1,395 amino acids. RT-PCR analysis revealed ubiquitous expression of KIAA0443 with highest levels in brain, kidney, and ovary, whereas SMPX (Small muscular protein) has the lowest molecular weight protein-encoding gene in
Xq22.1 (9560 Da). The gene consists of five exons (> or =172, 57, 84, 148, > or =422 bp) and four introns (3639, 10410, 6052, 31134 bp) comprising together 52.1 kb and is preferentially and abundantly expressed in heart and skeletal muscle. Statistical analysis is

![Figure 77-Molecular weight analysis](image)

**Figure 77-Molecular weight analysis**

Series 1: Xq22.1 genes

Molecular class:

In Xq22.1, molecular classes of most of the genes are unknown (37%). Presently, the transcription regulatory protein (TCEAL1, TCEAL2) is a major molecular class (12%) identified in Xq22.1. (Pillutla et al., 1999) revealed that TCEAL1 is the gene
which consists of three exons and two introns that code for p21/SIIR and their expression was lowest in hematopoietic cells of both normal and tumor origin. It was mapped to human chromosome Xq22.1 by fluorescence in situ hybridization. 9% of the molecular class identified as RNA binding protein (CSTF2, NXF2) and 8% of the molecular class identified as GTPase (RAB9B, RAB40A). Transcription factor (TAF7L, ESX1L) and cysteine protease (APG4A) are the molecular class identified as 7%. Transport / Cargo protein (GPRASP1, SYTL4) is one of the molecular classes as 5%. Ribosomal subunit (RPL36A), Adaptor molecule (NGFRAP1), Enzyme: Methyltransferase (CXorf34), Integral membrane protein (TMEM35), Enzyme: Translocase, these are the molecular classes' shows similar percentage (3%) in human genome Xq22.1 as shown in the figure-78.

Figure 78- Percentage of Molecular classes of gene in Xq22.1
Gene localization analysis

In Xq22.1, most of the gene localization is unknown (37%). 29% genes are localized in nucleus (TCEAL1, ARMCX5, TCEAL2, CSTF2, ESX1L, NXF2 and SMPX). 18% are localized in cytoplasm (APG4A, RPL36A, NGFRAP1 and MID2). Some genes (TIMM8A) are localized in mitochondria (3%). 3% genes (SYTL4) are localized in secretory granule. In some genes (TAF7L) primary localization is in nucleus and alternative localization is in the cytoplasm (3%) as shown in the figure 79.

Figure 79-Analysis of Xq22.1 Gene localization
Biological Process analysis

Most of the biological process in human genome Xq22.1 about 37% is unknown. 32% of the biological process is nucleic acid metabolism (TAF7L, TCEAL1, TCEAL2, CSTF2, CXorf34, ESX1L and NXF2). Protein metabolism (APG4A, RPL36A) and Signal transduction (NGFRAP1, RAB9B and RAB40A) is the biological processes observed in genome Xq22.1 as 10%. About 5% of biological process is observed as transport (GPRASP1, SYTL4). Metabolic pathways (TIMM8A) and cell growth factor maintenance (M1D2) are the biological processes observed as least (3%) in Xq22.1 which are shown in the figure-80.

![Biological Process Analysis](image)

**Figure 80-Biological process in Human Genome Xq22.1**

Expression Analysis

Figure-81, in the gene expression analysis maximum gene expression about 30% is observed in Testis (TAF7L, FSHTRH1, APG4A, TCEAL3, NGFRAP1 and
Further, about 9% genes are expressed in Heart (APG4A, TCEAL1, TIMM8A and SMPX) and Prostate (APG4A, TCEAL1 and MID2). 8% gene expression is observed in Spleen (FSHPRH1, APG4A and TCEAL1) and skeletal muscle (APG4A, TCEAL1, TIMM8A and SMPX). Brain (FSHPRH1, TCEAL1, TIMM8A and SMPX) and Ovary (FSHPRH1, TCEAL1, NGFRAP1 and MID2) show similar gene expression of about 7%. 5% gene expression is observed in most parts of the body. The list includes Lung (FSHTRH1, TCEAL1), Placenta (TCEAL1, ESX1L), Liver (TCEAL1, TIMM8A) and Kidney (TCEAL1, TIMM8A). In Fetus (APG4A, MID2), Colon (TCEAL1) and Thymus (TCEAL1) the gene expression is about 4%. About 2% genes are expressed as Ubiquitous (RAB9B). Finally the least expression of gene i.e. about 1% is observed in B cell (CSTF2) and small intestine (MID2).

Figure 81- Percentage of Xq22.1 Gene expression Analysis in Organs
ISOFORM ANALYSIS IN Xq22.1:

Statistical analysis of the genes present on Xq22.1 indicated that 8% isoform genes were present in this region is depicted in figure 82. This percentage included an Autophagy-related cysteine endopeptidase-2 protein (APG4A). APG4A has three isoforms denoted as Isoform a, Isoform b and Isoform c and the mapping shown in the figure 83.

Figure 82: Isoform analysis in Xq22.
Figure 83-GENETIC MAPPING by CLUSTALW Results

PROTEOLYTIC CLEAVAGE:

The attention is now turned to a different mechanism of enzyme regulation. Many enzymes acquire full enzymatic activity as they spontaneously fold into their characteristic 3D forms. In contrast, other enzymes are synthesized as inactive precursors that are subsequently activated by cleavage of one or a few specific peptide bonds. The inactive precursor is called a ‘zymogen’ (or a proenzyme). An energy source (ATP) is not needed for cleavage. Therefore, in contrast to the reversible regulation by phosphorylation, even a
protein located outside cells can be activated by this means. Proteolytic enzymes through their ability to catalyze irreversible hydrolytic reactions play crucial roles in the development and maintenance of all living organisms. Proteases were initially characterized as nonspecific degradative enzymes associated with protein catabolism, but recent studies have demonstrated that they influence a wide range of cellular functions by processing multiple bioactive molecules. These essential processes initiated, regulated, or terminated by proteases include DNA replication, cell cycle progression, cell proliferation, differentiation and migration, morphogenesis and tissue remodeling, and angiogenesis and apoptosis. An additional process in which proteolytic enzymes have also been recently implicated is autophagy.

Proteolytic cleavage in APG4A:

Autophagy is a biological process involved in the intracellular destruction of endogenous proteins and removal of damaged organelles and has been suggested to be essential for cell homeostasis as well as for cell remodeling during differentiation, metamorphosis, non-apoptotic cell death, and ageing.
Through analysis, it was shown that in Xq22.1 genome, 8% of APG4A isoform genes were formed by proteolytic cleavage. This was caused by a proteolytic enzyme called HsAp4gA. Therefore, 8% proteolytic cleavage occurs in the Xq22.1 of X-chromosome genome. This was also confirmed by the statistical studies of proteolytic cleavage in Xq22.1. GATE-16 (Golgi-associated ATPase enhancer of 60kDa) is an essential component of intra-Golgi transport and post-mitotic Golgi reassembly. The COOH terminus of GATE-16 undergoes post-translational cleavage by a human cysteine protease HsAp4gA, which exposes the glycine 116 residues.

Single nucleotide polymorphisms (SNPs):

Single nucleotide polymorphisms or SNPs (pronounced "snips") are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. For example a SNP might change the DNA sequence AAGGCTAA to
ATGGCTAA. For a variation to be considered an SNP, it must occur in at least 1% of the population. SNPs, which make up about 90% of all human genetic variations, occurs every 100 to 300 bases along the 3-billion-base human genome. Two of every three SNPs involve the replacement of cytosine (C) by thymine (T). SNPs can occur in both coding (gene) and noncoding regions of the genome. Many SNPs have no effect on cell function, but scientists believe others could predispose people to disease or influence their response to a drug. Although more than 99% of human DNA sequences are the same across the population, variations in DNA sequence can have a major impact on how humans respond to a disease; environmental insults such as bacteria, viruses, toxins, and chemicals; drugs and other therapies. This makes SNPs of great value for biomedical research and for developing pharmaceutical products or medical diagnostics. SNPs are also evolutionarily stable, not changing much from generation to generation, making them easier to follow in population studies. Scientists believe that SNP maps help them to identify the multiple genes associated with complex diseases such as cancer, diabetes, vascular disease, and some forms of mental illness.

SNPs do not cause disease, but they can help determine the likelihood that someone might develop a particular disease. One of the genes associated with Autophagy-related cysteine endopeptidase 2 or APG4A in Xq22.1 is a good example of how SNPs affect disease development. This gene contains 65 SNPs. A change of one amino acid in the APG4A protein alters its structure and function enough to make disease development more likely. SNPs are not absolute indicators of disease development.

SNPs Databases:
The single nucleotide polymorphism database (dbSNP) has been designed to support submissions and research into a broad range of biological problems. These include physical mapping, functional analysis and pharmacogenomics, association studies and evolutionary studies.

![PROTEIN SNP DATA ANALYSIS FOR Xq22.1 IN HUMAN](image)

**Figure 85-SNP Analysis**

In Xq22.1, it was discovered that APG4A had much more single nucleotide polymorphisms (SNPs) than any other gene. This collection of polymorphisms includes single-base nucleotide substitutions (also known as single nucleotide polymorphism (or) SNPs), small-scale multi-base deletion or insertion (also called deletion insertion polymorphisms (or) DIPs) and retrotransposable element insertion and microsatellite repeat variations (also called short tandem repeats (or) STRs).
Analysis of the nature of SNPs in APG4A revealed that more than 20% of the single nucleotide polymorphisms were caused by the three different alleles C/T, G/T and A/G. But all other alleles were less than 8% and some of the APG4A alleles were insertion deletion polymorphisms, deletions represented by '-' in allele string (in del) type as shown in the figure-86

Figure 86-APG4A gene analysis
Functional Analysis:

Variations that occur in functional regions of genes or in conserved non-coding regions of genes can cause significant changes in the complement of transcribed sequences, leading to changes in protein expression, which may affect the aspects of the phenotype such as metabolism or cell signaling. The possible functional implications of DNA sequence variations in dbSNP were noted in terms of how variation could alter mRNA transcripts as shown in the figure-87.

![FUNCTION ANALYSIS](image)

**Figure 87-Analysis of SNP positions**

In APG4A gene SNPs', a variant may have more than one functional role. It was indicated in the statistical analysis that 52% of functional regions had variants in transcripts, but not in the coding region intervals. 41% of functional regions showed variation in introns, but not in the first two or last two bases of introns. VALIDATION ANALYSIS: SNPs assays validated directly by the submitter through the validation section showed the type of evidence used to confirm.
Validation status of APG4A gene variant may be found by more than one method. 15% validation status of APG4A was found by 2hit-2allele. All alleles have been observed in 2+ chromosomes 9% by genotype and 8% by frequency and by cluster. Maximum validation status in APG4A was found to be unknown (or) no validation could be reported for this ref SNP as shown in the figure-88

VALIDATION STATUS ANALYSIS

![Graph showing validation status analysis]

Figure 88-Validation analysis

2.5.2 Drug target for APG4A

Genomic location Mapping of APG4A gene in human genome is based on Ensemble database (Clamp et al., 2003). It is found that this gene belongs to X-
chromosome, which is prone for disease linkage. Chromosome Entrez Gene cytogenetic band: Xq22.1-q22.3 Ensemble cytogenetic band: Xq22.3

Gene in genomic location:

Bands according to Ensembl, locations according to Gene loc (Rosen et al., 2003).

**Figure 89-Genomic location of APG4Agene in X-chromosome**

Start: 107,141,079 bp from pter

End: 107,204,045 bp from pter

Size: 62,966 bases

Orientation: plus strand

Tissue Specificity:

Using Electronic Northern Blot, the tissue specificity of APG4A gene was tabulated below using Gene Card Database (Rebhan et al., 1997) and a pie chart drawn based on the analysis from the table-27 and the graphical representation in figure-90.
Figure 90-Normalised expression of cancer genes

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Normalized Expression (%)</th>
<th>Cluster Clones: Tissue Clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spleen</td>
<td>21.96</td>
<td>3:15835</td>
</tr>
<tr>
<td>Blood</td>
<td>11.72</td>
<td>5:49438</td>
</tr>
<tr>
<td>Uterus</td>
<td>7.38</td>
<td>9:141322</td>
</tr>
<tr>
<td>Lymph node</td>
<td>6.76</td>
<td>7:120024</td>
</tr>
<tr>
<td>Bone</td>
<td>6.44</td>
<td>3:53985</td>
</tr>
<tr>
<td>Unclassified</td>
<td>4.77</td>
<td>11:267127</td>
</tr>
<tr>
<td>Testis</td>
<td>3.98</td>
<td>4:116513</td>
</tr>
<tr>
<td>Placenta</td>
<td>3.83</td>
<td>5:151135</td>
</tr>
<tr>
<td>Muscle</td>
<td>3.53</td>
<td>3:98454</td>
</tr>
<tr>
<td>Cervix</td>
<td>3.43</td>
<td>1:33757</td>
</tr>
</tbody>
</table>

Table 26-Expression of APG4A gene in various tissues

GENE EXPRESSION: The graphical representation of expression for APG4A gene on various organs of the body shown below for both in normal and during cancer in figure-91.
EST EXPRESSION:

Similarly the EST’s expression for the APG4A gene in Normal and Cancer tissues shown in figure-92.

Figure 91-Level of APG4A gene expression in normal and cancer cells

Figure 92-Level of APG4A EST expression in normal and cancer cells
Receptor model:

The computationally modeled 3D structure of Autophagy causing gene APG4A of Homo sapiens with PDB ID: 2FUY was selected as the molecular target for drug design. The Swiss-prot sequence and PDB 3D structure was given in figure -93. Validation of the structure are shown in figure-94, 95, 96

Modeling of target sequence:

![Modeled proteins APG4A (PDB id: 2FUY Author's publication in PDB)](image)

**Figure 93-Modeled proteins APG4A (PDB id: 2FUY Author's publication in PDB)**

**PROTEIN MODEL VALIDATION:**

Protein model validation is used using Ramachandran’s plot as shown in the figure-. All the residues are in allowed region. The disallowed region residues were energy minimized using Gromos in swisspdb viewer.
Figure 94-Protein model validation

Validation using procheck summary is given in the figure-95
--- PROCHECK SUMMARY ---

/usr/httpd/html/Services/SAVS/jobs/3118752/apdb.pdb 2.0 398 residues

Ramachandran plot: 89.4% core 9.1% allow 1.1% gener 0.3% disall

All Ramachandrans: 11 labelled residues (out of 396)
Chil-chi2 plots: 6 labelled residues (out of 253)

Main-chain params: 6 better 0 inside 0 worse
Side-chain params: 5 better 0 inside 0 worse

Residue properties: Max. deviation: 4.0 Bad contacts: 7
Bond len/angle: 6.7 Morris et al class: 1 1 2

G-factors
Dihedrals: -0.03 Covalent: -0.22 Overall: -0.10
H/c bond lengths: 96.9% within limits 1.1% highlighted
H/c bond angles: 92.3% within limits 7.7% highlighted
Planar groups: 99.3% within limits 0.7% highlighted

May be worth investigating further. * Worth investigating further.

--- Figure 95-Procheck Summary APG4A ---

Validation using PROSA PLOT is given in the figure-96

PROSA PLOT – 2FUY (TARGET-YELLOW) Vs 2CY7 (TEMPLATE – RED):

--- Figure 96-PROSA plot for APG4A ---
Active sites and binding pockets:

   The active site residue Cysteine 77 of APG4A was confirmed as potential one from swissprot, the putative active sites with spheres (PASS) around the active site residue are shown below. Binding sites and active sites of proteins are associated with structural pockets and cavities. A cavity (or void) is an interior empty space that is not accessible to the solvent probe. It has no mouth openings to the outside bulk solution shown in the figure-97.

Active Site prediction results:

   Figure 97-PASS active site in the conserved domain region

   CASTP provides identification and measurements of surface accessible pockets as well as interior inaccessible cavities, for proteins and other molecules. It measures analytically the area and volume of each pocket and cavity, both in solvent accessible surface (SA, Richard's surface) and molecular surface (MS, Connolly's surface). It also
measures the number of mouth openings, area of the openings, and circumference of mouth lips, in both SA and MS surfaces for each pocket shown in figure-98.

Figure 98-Available pockets for APG4A as predicted by the CASTp

Ligand generation: Diethylaminomethanedithioate, the suitable lead molecule for Autophagy gene APG4A was selected from the seed structure fes4 when searched against Pubchem database, which has also shown considerable action against cancer. The docking parameter file used for PLA2 is given figure -99.
Figure 99-LIGPLOT of ligand's interactions with protein 1H0H

The predicted molecular descriptor from NCI database for this particular ligand. Simultaneously various analogs of this ligand are also selected as shown in the figure 100 and its annotation is given in the table-28.

Screened ligand and its properties:

<table>
<thead>
<tr>
<th>Assay ID(AID)</th>
<th>Assay Name</th>
<th>Activity Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>155</td>
<td>NCI Yeast Anticancer Drug Screen. Data for the rad59 strain</td>
<td>Active</td>
</tr>
<tr>
<td>157</td>
<td>NCI Yeast Anticancer Drug Screen. Data for the mec2-1 strain</td>
<td>Active</td>
</tr>
<tr>
<td>161</td>
<td>NCI Yeast Anticancer Drug Screen. Data for the apt1 mph1 strain</td>
<td>Active</td>
</tr>
<tr>
<td>165</td>
<td>NCI Yeast Anticancer Drug Screen. Data for the cdk2 rad14 strain</td>
<td>Active</td>
</tr>
<tr>
<td>167</td>
<td>NCI Yeast Anticancer Drug Screen. Data for the bol3 strain</td>
<td>Active</td>
</tr>
<tr>
<td>175</td>
<td>NCI Yeast Anticancer Drug Screen. Data for the msh1 rad18 strain</td>
<td>Active</td>
</tr>
</tbody>
</table>

Figure 100-Ligand with NCI anticancer drug screen data
IUPAC Name: 1, 2-dicarboxyethanethiolate; diethylaminomethanedithioate; iron (+3) cation

Synonym – NSC667220

Canonical SMILES: CCN(C(C(=S)-))[S-].CCN(C(C(=S)-)[S-].C(C(C(=O)O)[S-]-[S-]).C(C(=0)O).[Fe+3]

<table>
<thead>
<tr>
<th>S. No</th>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Molecular Weight</td>
<td>501.534g/mol</td>
</tr>
<tr>
<td>2</td>
<td>Molecular Formula</td>
<td>C14H25FeN204S5</td>
</tr>
<tr>
<td>3</td>
<td>Hydrogen Bond Donor Count</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Hydrogen Bond Acceptor Count</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>Rota table Bond Count</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 27-NCI Predicted property for ligand

Molecular Docking:

The ligand was docked with receptor protein using Autodock3 and from the resulting confirmations; the best confirmation is selected based on the binding energy values. The result was shown in the figure-101.

Docking conformation for ligand NSC667220 result using Autodock 3:

BINDING ENERGY: -4.72kcal/mol

Figure 101-2fuy Vs 1, 2-dicarboxyethanethiolate complex
The various analogs of the ligand are also docked with the same receptor protein and binding energy values are considered for research as shown in the figure-102 and the analogs table is given in the table-29.

Docking conformation for NSC667220 analogs result using Autodock 3:

![Docking conformation for different analogs near active site](image)

**Figure 102-Docked confirmations for different analogs near active site**

**BINDING ENERGY VALUES FOR ANALOG COMPOUNDS:**

<table>
<thead>
<tr>
<th>ID</th>
<th>STRUCTURE</th>
<th>BINDING ENERGY(Kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSC634040</td>
<td><img src="image" alt="Structure" /></td>
<td>-5.28</td>
</tr>
<tr>
<td>Compound</td>
<td>Score</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>NSC634041</td>
<td>-5.11</td>
<td></td>
</tr>
<tr>
<td>NSC665970*</td>
<td>-4.81</td>
<td></td>
</tr>
<tr>
<td>NSC665971*</td>
<td>-6.39</td>
<td></td>
</tr>
<tr>
<td>NSC695576</td>
<td>-5.59</td>
<td></td>
</tr>
</tbody>
</table>
The grid parameter that is generated in autodock is shown below in Table 30:

<table>
<thead>
<tr>
<th>receptor 2fuy.pdbqs</th>
<th># macromolecule</th>
</tr>
</thead>
<tbody>
<tr>
<td>gridfld 2fuy.maps.fld</td>
<td># grid_data_file</td>
</tr>
<tr>
<td>npts 60 60 60</td>
<td># num.grid points in xyz</td>
</tr>
<tr>
<td>spacing 0.375</td>
<td># spacing(A)</td>
</tr>
<tr>
<td>gridcenter 41.408 5.058 94.482</td>
<td># xyz-coordinates or auto</td>
</tr>
<tr>
<td>types CNOSH # atom type names</td>
<td># atom type names</td>
</tr>
<tr>
<td>smooth 0.5 # store minimum energy w/in rad(A)</td>
<td># store minimum energy w/in rad(A)</td>
</tr>
<tr>
<td>map 2fuy.C.map</td>
<td># atom-specific affinity map</td>
</tr>
<tr>
<td>nbp_r_eps 4.00 0.0222750 12 6 # C-C lj</td>
<td># C-C lj</td>
</tr>
<tr>
<td>nbp_r_eps 3.75 0.0230026 12 6 # C-N lj</td>
<td># C-N lj</td>
</tr>
<tr>
<td>nbp_r_eps 3.60 0.0257202 12 6 # C-O lj</td>
<td># C-O lj</td>
</tr>
<tr>
<td>nbp_r_eps 4.00 0.0257202 12 6 # C-S lj</td>
<td># C-S lj</td>
</tr>
<tr>
<td>nbp_r_eps 3.00 0.0081378 12 6 # C-H lj</td>
<td># C-H lj</td>
</tr>
<tr>
<td>nbp_r_eps 3.00 0.0081378 12 6 # C-H lj</td>
<td># C-H lj</td>
</tr>
<tr>
<td>nbp_r_eps 3.00 0.0081378 12 6 # C-H lj</td>
<td># C-H lj</td>
</tr>
<tr>
<td>sol_par 12.77 0.6844 # C atomic fragmental volume, solvation parameters</td>
<td># C atomic fragmental volume, solvation parameters</td>
</tr>
<tr>
<td>constant 0.000 # C grid map constant energy</td>
<td># C grid map constant energy</td>
</tr>
<tr>
<td>map 2fuy.N.map</td>
<td># atom-specific affinity map</td>
</tr>
<tr>
<td>nbp_r_eps 3.75 0.0230026 12 6 # N-C lj</td>
<td># N-C lj</td>
</tr>
<tr>
<td>nbp_r_eps 3.50 0.0237600 12 6 # N-N lj</td>
<td># N-N lj</td>
</tr>
<tr>
<td>nbp_r_eps 3.35 0.0265667 12 6 # N-O lj</td>
<td># N-O lj</td>
</tr>
<tr>
<td>nbp_r_eps 3.75 0.0265667 12 6 # N-S lj</td>
<td># N-S lj</td>
</tr>
<tr>
<td>nbp_r_eps 1.90 0.3280000 12 10 # N-H hb</td>
<td># N-H hb</td>
</tr>
<tr>
<td>nbp_r_eps 1.90 0.3280000 12 10 # N-H hb</td>
<td># N-H hb</td>
</tr>
<tr>
<td>nbp_r_eps 1.90 0.3280000 12 10 # N-H hb</td>
<td># N-H hb</td>
</tr>
<tr>
<td>sol_par 0.00 0.0000 # N atomic fragmental volume, solvation parameters</td>
<td># N atomic fragmental volume, solvation parameters</td>
</tr>
<tr>
<td>constant 0.000 # N grid map constant energy</td>
<td># N grid map constant energy</td>
</tr>
<tr>
<td>map 2fuy.O.map</td>
<td># atom-specific affinity map</td>
</tr>
<tr>
<td>nbp_r_eps 3.60 0.0257202 12 6 # O-C lj</td>
<td># O-C lj</td>
</tr>
<tr>
<td>nbp_r_eps 3.35 0.0265667 12 6 # O-N lj</td>
<td># O-N lj</td>
</tr>
<tr>
<td>nbp_r_eps 3.20 0.0297000 12 6 # O-O lj</td>
<td># O-O lj</td>
</tr>
<tr>
<td>nbp_r_eps</td>
<td>3.60 0.0297000 12 6</td>
</tr>
<tr>
<td>nbp_r_eps</td>
<td>1.90 0.3280000 12 10</td>
</tr>
<tr>
<td>nbp_r_eps</td>
<td>1.90 0.3280000 12 10</td>
</tr>
<tr>
<td>nbp_r_eps</td>
<td>1.90 0.3280000 12 10</td>
</tr>
<tr>
<td>sol_par</td>
<td>0.00 0.0000</td>
</tr>
<tr>
<td>constant</td>
<td>0.236</td>
</tr>
<tr>
<td>map 2fuy.S.map</td>
<td># atom-specific affinity map</td>
</tr>
<tr>
<td>nbp_r_eps</td>
<td>4.00 0.0257202 12 6</td>
</tr>
<tr>
<td>nbp_r_eps</td>
<td>3.75 0.0265667 12 6</td>
</tr>
<tr>
<td>nbp_r_eps</td>
<td>3.60 0.0297000 12 6</td>
</tr>
<tr>
<td>nbp_r_eps</td>
<td>4.00 0.0297000 12 6</td>
</tr>
<tr>
<td>nbp_r_eps</td>
<td>2.50 0.0656000 12 10</td>
</tr>
<tr>
<td>nbp_r_eps</td>
<td>2.50 0.0656000 12 10</td>
</tr>
<tr>
<td>nbp_r_eps</td>
<td>2.50 0.0656000 12 10</td>
</tr>
<tr>
<td>sol_par</td>
<td>0.000 0.000</td>
</tr>
<tr>
<td>constant</td>
<td>0.000</td>
</tr>
<tr>
<td>map 2fuy.H.map</td>
<td># atom-specific affinity map</td>
</tr>
<tr>
<td>nbp_r_eps</td>
<td>3.00 0.0081378 12 6</td>
</tr>
<tr>
<td>nbp_r_eps</td>
<td>1.90 0.3280000 12 10</td>
</tr>
</tbody>
</table>

---

| nbp_r_eps | 1.90 0.3280000 12 10 | H-O hb |
| nbp_r_eps | 2.50 0.0656000 12 10 | H-S hb |
| nbp_r_eps | 2.00 0.0029700 12 6 | H-H lj |
| nbp_r_eps | 2.00 0.0029700 12 6 | H-H lj |
| nbp_r_eps | 2.00 0.0029700 12 6 | H-H lj |
| sol_par | 0.00 0.0000 | H atomic fragmental volume, solvation parameters |
| constant | 0.118 | H grid map constant energy |
| map 2fuy.f.map | # atom-specific affinity map |
| nbp_r_eps | 2.65 0.0057469 12 6 | f-C lj |
| nbp_r_eps | 2.40 0.0059400 12 6 | f-N lj |
| nbp_r_eps | 2.25 0.0066379 12 6 | f-O lj |
| nbp_r_eps | 2.65 0.0066379 12 6 | f-S lj |
| nbp_r_eps | 1.65 0.0020938 12 6 | f-H lj |
| nbp_r_eps | 1.65 0.0020938 12 6 | f-H lj |
| nbp_r_eps | 1.65 0.0020938 12 6 | f-H lj |
| sol_par | 0.000 0.000 | f atomic fragmental volume, solvation parameters |
| constant | 0.000 | f grid map constant energy |
| elecmap 2fuy.e.map | # electrostatic potential map |
| dielectric | -0.1146 | <0, distance-dep.diel;>0, constant |

---

outlev 1 # diagnostic output level
seed pid time # seeds for random generator
types CNOSHf # atom type names
fld 2fuy.maps.fld # grid_data_file
map 2fuy.C.map # atom-specific affinity map
map 2fuy.N.map # atom-specific affinity map
map 2fuy.O.map # atom-specific affinity map
map 2fuy.S.map  # atom-specific affinity map
map 2fuy.H.map  # atom-specific affinity map
map 2fuy.f.map  # atom-specific affinity map
map 2fuy.e.map  # electrostatics map
move ligand    # small molecule
about 5.7499 -0.1071 0.0  # small molecule center
tran0 random   # initial coordinates/A or random
quat0 random   # initial quaternion
ndihe 0        # number of active torsions
dihe0 random   # initial dihedrals (relative) or random
tstep 2.0      # translation step/A
qstep 50.0     # quaternion step/deg
dstep 50.0     # torsion step/deg
torsdof 15 0.3113  # torsional degrees of freedom and coefficient

intnbp_r_eps  4.00 0.0222750 12 6  # C-C Ij
intnbp_r_eps  3.75 0.0230026 12 6  # C-N Ij
intnbp_r_eps  3.60 0.0257202 12 6  # C-O Ij
intnbp_r_eps  4.00 0.0257202 12 6  # C-S Ij
intnbp_r_eps  3.00 0.0081378 12 6  # C-H Ij
intnbp_r_eps  2.65 0.0057469 12 6  # C-f Ij
intnbp_r_eps  3.50 0.0237600 12 6  # N-N Ij
intnbp_r_eps  3.35 0.0265667 12 6  # N-O Ij
intnbp_r_eps  3.75 0.0265667 12 6  # N-S Ij
intnbp_r_eps  2.75 0.0084051 12 6  # N-H Ij
intnbp_r_eps  2.40 0.0059400 12 6  # N-f Ij
intnbp_r_eps  3.20 0.0297000 12 6  # O-O Ij
intnbp_r_eps  3.60 0.0297000 12 6  # O-S Ij
intnbp_r_eps  2.60 0.0093852 12 6  # O-H Ij
intnbp_r_eps  2.25 0.0066379 12 6  # O-f Ij
intnbp_r_eps  4.00 0.0297000 12 6  # S-S Ij
intnbp_r_eps  3.00 0.0093852 12 6  # S-H Ij
intnbp_r_eps  2.65 0.0066379 12 6  # S-f Ij
intnbp_r_eps  2.00 0.0029700 12 6  # H-H Ij
intnbp_r_eps  1.65 0.0020938 12 6  # H-f Ij
intnbp_r_eps  1.30 0.0014850 12 6  # f-f Ij

# rmstol 0.5  # cluster_tolerance/A
extng 1000.0  # external grid energy
e0max 0.0 10000  # max initial energy; max number of retries
ga_num_size 150  # number of individuals in population
nga_num_evals 250000  # maximum number of energy evaluations
ga_num_generations 27000  # maximum number of generations

ga_elitism 1  # number of top individuals to survive to next generation
ga_mutation_rate 0.02  # rate of gene mutation
ga_crossover_rate 0.8  # rate of crossover
ga_window_size 10  #
ga_cauchy_alpha 0.0  # Alpha parameter of Cauchy distribution
2.4.3 Identifying Novel gene and verifying experimentally using RT-PCR.

Novel gene-1

Putative Novel Protein coding Regions based on comparative genomic Analysis is done when the NCBI Build 34 with 3 exons. The genomic region from 98941268-98955150, the gene being predicted in negative strand with orthologs in Chimp, mouse, rat and fugu. The tissue source that is taken from the placenta which is represented in table-31

| Novel 1 | 98955150-98954815 | 98949909-98949610 | 98941927-98941268 | 3 | Minus | 7 Transmembrane domains | Yes | Yes | yes | YES | AY589511 NM_212559 | Placenta |

Table 31-Novel Protein coding Regions based on comparative genomic Analysis

| Table 30-Receptor (2fuy)-Grid parameter file |

<table>
<thead>
<tr>
<th>ga_cauchy_beta 1.0</th>
<th># Beta parameter Cauchy distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>set_ga</td>
<td># set the above parameters for GA or LGA</td>
</tr>
<tr>
<td>sw_max_its 300</td>
<td># iterations of Solis and Wets local search</td>
</tr>
<tr>
<td>sw_max_succ 4</td>
<td># consecutive successes before changing rho</td>
</tr>
<tr>
<td>sw_max_fail 4</td>
<td># consecutive failures before changing rho</td>
</tr>
<tr>
<td>sw_rho 1.0</td>
<td># size of local search space to sample</td>
</tr>
<tr>
<td>sw_lb_rho 0.01</td>
<td># lower bound on rho</td>
</tr>
<tr>
<td>ls_search_freq 0.06</td>
<td># probability of performing local search on individual</td>
</tr>
<tr>
<td>set_sw1</td>
<td># set the above Solis and Wets parameters</td>
</tr>
<tr>
<td>ga_run 10</td>
<td># do this many hybrid GA-LS runs</td>
</tr>
<tr>
<td>Analysis</td>
<td># perform a ranked cluster analysis</td>
</tr>
</tbody>
</table>

Table 30-Receptor (2fuy)-Grid parameter file
Experimental Validation:

Novel protein coding genes identified will be experimentally verified by RT-PCR. Total mRNA will be extracted from the tissues and the first strand cDNA synthesis will be carried out using oligo dT primer. The next step will be a normal PCR reaction with gene specific primers designed based on the gene structure predicted by comparative genomic approach. The PCR product will be radiolabeled and used as a probe to fish out the full length cDNA from the cDNA library by colony hybridization. The amplified products by RT-PCR are shown in the figure-41.
Experimental validation of novel genes:

An agarose gel with PCR amplification showing expected bands from predicted coding exons of novel genes as well as novel exons that were deduced based on comparative genomics. Novel 1 encodes a protein similar to Kell blood group protein. Total m-RNA will be extracted from the tissues and the first strand cDNA synthesis will be carried out using oligo dT primer. The next step will be a normal PCR reaction with gene specific primers designed based on the gene structure predicted by comparative genomic approach. The PCR product will be radiolabelled and used as a probe to fish out the full length cDNA from the cDNA library by colony hybridization.

PCR analysis:

CDNAs from heart, kidney, lung, breast, intestine, muscle, brain, liver were procured from Invitrogen (Carlsbad, CA, USA) or OriGene (Rockville, MD, USA), pooled and used as template in a PCR reaction to test the expression of different predicted novel genes or novel exons. The design of primers was such that the amplified product contained two exons in the case of multi-exon genes. The PCR products were eluted from the gel using QiaQuick gel extraction kit (Qiagen)

Novel gene-2

Identification of novel protein coding gene 2 by exon Stitching.

The contig NT_011651.16 was then broken up into segments the sequences representing already known genes was were excised out. The remaining genomic segments
were then selectively screened using various gene prediction tools such as MCG genes, Geneid, Genescan, SGP and genomic sequences which showed promise of finding genes in more than one gene prediction tool was selected for further analysis. The exon sequences in each of these regions were obtained from the NCBI website and divided into nine subsets. These subsets were then subjected to BLAST analysis in many sequential different combinations to find the set of sequences with the maximum similarity with known mouse genes and various other information such as the coincident reference sequence protein, non reference sequence protein, build protein, the protein accession number, reading frame, protein name and the percentage homology with the mouse genome was obtained.

The results were confirmed first TblastX to confirm at protein level. Blast 2 is another tool that helps find regions of sequence similarity which further yield functional and evolutionary clues about structure and function of novel sequences. A diagrammatic representation of the same was also obtained which delineated the percentage identity between the two sequences. Augustus was used to further confirm the BLAST results and to find out the frame in which the gene exists.

- Genomic region of Xq22.1- 98,500,000-99, 100, 00 (Build 36.2)
- Region where known gene is already present- 99020770-99029659
- Probable gene carrying region using various gene prediction tools-98508661-99081497
- Contig no -NT_011651.170
- Strand= Negative
• Cds start- 98861133
• Cds end- 99075687
• Total no of exons- 4
  1) 99075687-99075630
  2) 98989949-98989881
  3) 98862683-98862395
  4) 98862139-98861133

Highlighting the ORF sequence

ATGGTGGAAC CATGGGGACT TGGGGGTGGG GCAGAGGTGG GAACATTTGT ATCAGTTAAG 60
TCAGCTTCGT GGCTCCTGT GGAGCCAGGG CTGAGCCTTG TGACGCGCAC TCGCCAATTA 120
GAGATTGATC AGCCAGCAGT CAAATGCATT CTCCAGTCCT TGCAAGAAGG ATCAGCCTTT 180
TCTGTGCCAG CCTCGATCGC CTGTGTGCTTT GGTCTCTTTT TCTCCCGTGC TTGGATCCGT 240
CCTCGGCGGG GCCGCTCCTGT TGCTGAGACT CGGGTATACTG TTCTGTGAGC CCAGCTCCCT 300
TTAGTCACGT TTGCTGGGCT CTGCTACCGA ATAGTTGGGA TTACCAGAATA GTCCTCCCTTC 360
TCGCCGTGCA GCACAGATGC TGTAAGTGCC ACCCGCGTCC CGGTCAAGTG CCTGTGCCCG 420
AGCTCGCCGC CGCCGTGTGT GCTGCGCTGA TAGTTTCCAG TTGTTCTTTTC TTGGTGATGG 480
TGATATAGAT GTACAGTACT ACATGGGACT AATATATAGG AGTTTCCAGT TGTTCTTTTC 540
GGGTGATGGA GTATATAGTG TACAGTACTA CATGGCACTA ATATATAGGG GCTATGGTTG 600
AATCTCAGAG CGAAGGTGAG TTGACTCTTT TGACATGAG CATCCAGTGT ATCCAAAGTG 660

245
CGAGTATTC AGAAGAGGAG CTGAAATGCA CAGGCTCTTC CACCAGGTGG AATCTTGCCT 1920
GCAGCATCTC TCCTGCCATC AAAGGTGCTG CCAAAGTGGC AAAGATCCTT TAAAGACTGA 1980
TGGCACTTCT AAAGGAGGGG TTTCTCGTGG TTGGCACTAC AGCCTCAGGC ATTTGTACCA 2040
CACTACTGGA GGTTCCAAT TATTCTTATG ATCAGACCCT TGCACTTCTGA GTCAATTGTG 2100
TTCAAAGTAT TTCAATAACA TGTAATCAG CAGCCTCTGT GGTTGGTTCT TTTCTGTTG 2160
CACTCTCCT ACTAGCTTCT TTGATTGCG ATCC 2194

✓ Start codon in green.
✓ Stop codon in red.
✓ Coding in orange.

The protein sequence being

```
MDLTTHSVKFWNPVHFEESSNLDDLQQKVRAKETRKWALSRLKLKLKDINIVSIGYSLVQKALK
PPSIKLYRKTNEPVKTFKRFSINTGSLLPSTDTRQSQQIYGSQJIEKEETEELKWFDEPALMLMV
FKPLVMLKEHHYPRPSLFTYPEESPVNGSSSLFLSSALLHKCLEKVIALCRYTPRRNIPPSVALVSQE
EEDDLQKIQVTTPGFLVFLPVADDKRKVPFTEKVMATPEQVDNMKAIQKLCYFTYRSDSFENPV
LQQHFGNLPLELILIYMEEQAVDLILPKVEEMNKRLGSMVNEFKELVYSPDPKNPGEKITYKRCNNE
GSRSRKPRKSIQKRS
```

The localization is being predicted as nucleus from Psort or less probable in cytoplasm could be shuttling protein. The probability got from xrcc6 human protein, the length of the protein being 608 and present in the chromosome 22q11-q13, but the predicted sequence shows homology with mouse genome which gives various possibility whether the predicted protein can be pseudogene or other gene. To proceed we did domain
analysis of the predicted protein and found it has ku78 and its functionality being double strand streak break repair via non homologus end joining as shown in the figure-102.

**Figure 102-Ku 78 Auto antigen**

Electronic northern plot was performed and found that this protein is homologous to thyroid antigen protein and the expressed in lymphocytes and hence for experimental validation the tissues from lymphocytes were taken. The electronic northern blot results are shown in the figure-103.
Figure 103-Electronic northern blot analysis
The interaction network/probable protein-protein interaction showed its interaction with other genes like XRCC4, LIG4, PRKDC, ATM, ORC2L, TERF1, TERT, and G22P1 as shown in the figure-104.

A new gene was predicted using the algorithm and the functional annotation proved its validation to a greater extent and hence it is taken to wetlab for further verification.

Using Augustus the gene prediction tool, we found the gene is in the negative strand, the gene and the protein sequences are displayed.
Homology Modeling and Structure Validation of a Novel Protein

The sequence of the novel protein is

MDLTTHSVKFWNPVHFEESSNLEDLLQQVRAKETRKWALSRLKLKNKDIVCSVGIYSLV
QALKPSSIKLYRTNEPVTKSFKNINTGSLLLPSDTKRSQYGHSQIILEKEETEEL
KWDFEPALMLVFKPLVLMKHEHYPRPSLFTYPEESPVNGSTLSALLIKCLEKVVIAL
CRYTPRNNIPPYLVALVSQEEELDDQKIQVTTPGFQLVFPLFADDKRKVPFTEKVMATPE
QVDNMKAIQKLCFTYRSDSFENPVLQHFGNLPLELILYMLEEQAVDLILPKVEEMNKL
GSMVNEFKELVYSPDYNPEGKITKRKCNNEGSRKPKRSIQKRS

The templates were found using CPH model server 2

Template having the lowest E-Value was chosen to carry out the homology modeling. Then the template i.e. 1JEQ chain: A score was retrieved and subjected to Profile-
Profile alignment along with the sequence of the novel protein. Initially, three templates were suggested by the server i.e. 1JEQ chain: A 1JEY chain: A the results obtained is as follows

Round 0. All hits:

entry: 1JEQ chain: A score: 473 E: 1e-134

entry: 1JEY chain: A score: 466 E: 1e-131

entry: 1T0I chain: A score: 28 E: 8.6

1T0I chain: A.

Round 0. Hits better than threshold: 0.050000:

entry: 1JEQ chain: A score: 473 E: 1e-134

entry: 1JEY chain: A score: 466 E: 1e-131

Score: 657.5 bits

Identity: 85.7 %

Query: 1 VHFEESSNLQKVRAETKWLISRLKLKNKDIVISGIYSLVQKALKPPIKLYR 60

VHFEESS LEDLL+KVKRAETRK ALSRLKLKNKDIVISGIY+LVQKALKPP IKLYR

Temp1: 1 VHFEESSKLEDLLKVRKAEKTRKRALSLKLKLKNKDIVISVIYNLVQKALKPPPIKLYR 60

Query: 61 KTNEPVKTKSRFNTGSLLLPSDTKRSGIYGSQIILEKEETELKWFDEPALMLMVF 120

+TNEPVKTK+R FN +TG LLLPSDTKRSGIYGS QIILEKEETELK FD+P LMLM F
The visualization of the structure (figure-105) was done using the VMD tool. The figure below shows the structure of the novel protein.
Figure 105-Structure of Novel Protein using VMD tools

The modeled structure of the novel protein was also validated using the SAVS (Structural Analysis and Validation Server). It utilizes 5 programs to do the validation:

- PROCHECK
- WHAT_CHECK
- ERRAT
- VERIFY_3D

The following results were obtained from the respective programs:
Procheck:

The Modeled Structure of Novel Protein using PROCHECK program gives RAMACHANDRAN PLOT. This tool gives details about the Ramachandran plot as shown in the figure-106

![Ramachandran Plot for Novel protein-2](image)

**Figure 106-Ramachandran Plot for Novel protein-2**

With the program giving the overall quality factor of the modeled structure as 89.744

Verifying the gene results by RT-PCR:

Bioinformatics tools gave the amino acid sequence as well as the homology model for the predicted protein. This was further supported by homology with Xrcc6 human protein that suggested a nuclear localization for predicted protein with a role in DNA repair functions. All these results were finally confirmed in the laboratory using RT-PCR and other protocols. Illumination of the gel with UV light showed that the predicted
protein expressed only when the cells were irradiated and DNA was damaged. The steps followed were:

Lymphocyte isolation

Lymphocytes were isolated from peripheral blood obtained from young (23-25 years), non-smoking men by centrifugation in a density gradient of Gradisol L (15 min, 280g, 4°C). The viability of the cells was measured by trypan blue exclusion assay and was found to be about 99%. The final concentration of the lymphocytes was adjusted to 1-3x10^5 cells/ml by adding RPMI 1640 to the single cell suspension.

Irradiation:

In the in vitro study, the lymphocytes in culture were irradiated with 2 Gy of gamma rays which was generated using a 60Co source (Theratron Eldorado 6) at a dose rate of 0.5 Gy/ min as determined by calibration using an air ionization chamber (Capinter PR-06C) connected to an electrometer (Capintex 192X). RNA extraction and reverse transcriptase polymerase chain reaction (RT-PCR)

Total RNA was extracted and isolated from the cells (unirradiated- control and irradiated – experimental) using TRI REAGENT LS (Sigma) according to the instruction of the supplier. The RNA pellet was dissolved in diethyl pyrocarbonate (DEPC) treated water and the concentration and purity of the samples were assessed spectrophotometrically at 260 and 280nm. One microgram of RNA was reverse transcribed using the PowerScri0pt Reverse Transcriptase (Clonetech) according to the Clontech Laboratories Protocol, using a combination of oligo (dT) 15 primers and random hexamers. The number of PCR cycles was calibrated for each transcript to target the exponential phase of the reaction. The PCR
reaction were carried out using 1U of Taq DNA polymerase (Geneid) as follows: 5 min denaturation at 95°C, followed by the predetermined number of cycles of 30s at 95°C, 40s at 55°C, 1 min at 72°C. At final extension was carried out for 5 min at 72°C. Primers were (Gene Bank No.) : F CAGGGACTCTCCTGGGTACAR, AND CGTCAGATTGTGCTGGAGAA. PCR products were separated on 1.5% agarose gel in TAE buffer and stained with ethidium bromide. The image was captured under UV light and the intensities of the bands were analyzed and quantified by gel documentation system (Lark Innovative).

![Image of gel electrophoresis](image)

**Figure 107-Experimental validation of novel gene 2**

Lane 1: 1 kb ladder
Lane 2: control (no amplified product)
Lane 3: Irradiated sample

The unknown gene is sequenced (outsourced to Bangalore Genei- Bangalore) and the unknown sequence is shown below
>Unknown sequence (length 2,193)

ATGGTGGAACCATGGGGACTTGGGGGTGGGGCAGACGTGGGAACATTTGT
ATCAGTTAAGTCAGCTTCGTTGCTGCCCTGTGGAGCCAGCCAGCAAGTCAAATGCATTC
TCCAGTCTTGCAAGAAGGATCGACGCTCTTTTCTGTTGGCAAGCTCAGATCGCCT
TGCTTCTTGGTCTCTTTTCTCCCTGCTTTGATCCTGCTCGGCGGCGGCGGCGG
ACCTGGTGGCTGAGACTCGGGGTACCGTTCTGCTGCACCCAGCTCCCTTTAGT
CCGGTTTGGCTTGCTCTGGTACCAAATAGTTGGGATTACCGAAGAGTCCCC
TTCCTCGGCGTGTCAGCACAGATGCTGTGACTGCCACCCGCGTCCCCGTCAA
GTGCTCTGTGCCCGAGCTGCGGCCGCCGTGTGTGCTGCGCTGAGTTTCC
AGTTGTCTCTTCTGCTGATGCTGTGATATAGATGTGCACTACAGATGCCAC
TAAAATATAGGAGTTTCCAGTGTCTTTTCTGCTGGGTGATGCTGTGATATAGATG
TACAGTACTACATGGCACTAATATATAGGGGTATGTTGAATGCAGAG
CGAAGGTCAGTTGACTCTTTTGACATGAGCATCCAGTGTATCCAAAGTGT
GTACATGAGTAAGATCGAATCGAGATCTTTATGATACAGTACA
AAGACAAAAATTCAAGTTAATAAAAAATATATATACGTTTCATAGAGTG
GAAAATCCAGGGTCAAACAATAATTCTAAAGCTTACAGTTAAGGGCGA
AAGACAAAAATTCAGTGAATTTTAAAAATTTACGTCTTACAGGAGTTG
GAAAAATCCAGGTGCAAAAAACAAATTCTAAAGCTTGACCAGTTTAAGGGGCA
GCAGGGACAAAAATGTTTCGAAGACCTGATGGGCCATGGATCTGACTACT
CATTCAGTGAAAGTTCTGGAATCCAGTTCACTTTGAGAATCCAGCAATCTA
GAAGACCTGTGGACAGAAGGTTTCGTTGACAAGGAGACCAGGAAATGGGCAC
TCAGCAGGTTAAAGCTGAAGCTCAACAAAGACATAGTGATCTCTGTGGGC
ATTTACAGTCTCGTCCAGAAGGCTCTCAACAAAGCATTGGATCTCTGTGGGC
CGGAAAACAAATGAACCAGTGAAAACCAAGTCCCGGAAATTTAATATTAA
TACCGGCAGTTTGGCTTCTGCTAGCGATACCAAGCGGTCTCAGATCTATGG
GATGTCATCAGATTATACTGGAGAAAGAGGAAACAGAAGCCTAAAATGG
TTTGATGAACCAGCTTTTGATGCTCATGTTTTTCAAGCCTTTGTAATGTCTG
AAGGACACCATTACCAGGAGGCTTCTTCTTCAGCTACGCTCTGGTACTGAC
GCCGTGAATGGAGTCTTCCACCTGTTCACTGCTCAGGTACTGCTCATATTC
GGACAAGGTGCTAGTATGTCAGGATACACACACCACCGACGGAACATCC
CCCTTTATTTTGCTTGGCTTTGTCACAGGAAGAGGATGGGATGACCAG
AAAATTCAGGTGACTACTCCAGTTTCCAGCTGGTCTTTTTACTCCCCTTGCT
GATGATAAAAAGGAAGGTTGCTCTTTACTGAAAAAGTACATGGCAACTCCAGA
GCAGGTGGACAACATGAAGGCTATCATCAGAAGGTCTGCTTCACATACA
GAAGTGACAGCTTTGAGAACAACCGTGCAGCAGCAGCAAAGATCTTTCGGGAACCTG
GAGCCCTTGATCTTGTATTGGATGGAAAGAGAAGCATGGGGCTCCATGGTGAATGAGTT
CCCAAGGTGGAAAGAAAATGAAATAAAAGAAGGCTCCATGGTGAATGAGTCT
TAAGGAACTTGCTACTCCAGATTATAATCCTGAAAGGGAAATTACCA
AGAGAAAATGCAATAATGAAAGGTTCCAGAAGCAAAAGGCCCAAGCGGAG
TATTCAGAAGAGGAGCTGAACCTGACAGGCTCTTCCACCAGGTGGAATCT
TGCTTGCAAGCATCTCTCTGTCCTCCCTCAGGCTCCAGGCTCTCAGGCTCT
TCCTTTAAAGACTGATGGCACTTCTAAAGGAGGGGTTTCTCGTGGTTGGCA
CTACAGCCTCAGGCTTTTGTACCACACTCTGGAGGTCCCACATTATTTCTT
ATGATCAGACCCCTTGAGCAGTTGATTGTGTCTAAGAGATTTCATTCAATAA
CATGTAATCGAGCCCTCTGTGGTTGCTTTCTTCTGTGTCACCTTCTCT
ACTGCTTCTTGTATTGGCATCC
Given unknown sequence is blasted against NCBI database to find out homologous sequence

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Max score</th>
<th>Total score</th>
<th>Query coverage</th>
<th>E value</th>
<th>Max ident</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_001469_3</td>
<td>Homo sapiens X-ray repair complementing defective repair in Chinese hamster</td>
<td>1373</td>
<td>1749</td>
<td>58%</td>
<td>0.0</td>
<td>91%</td>
</tr>
<tr>
<td>XR_018440_1</td>
<td>PREDICTED: Homo sapiens similar to ATP-dependent DNA helicase 2 subunit</td>
<td>1049</td>
<td>2386</td>
<td>59%</td>
<td>0.0</td>
<td>100%</td>
</tr>
<tr>
<td>XR_016349_1</td>
<td>PREDICTED: Homo sapiens similar to ATP-dependent DNA helicase 2 subunit</td>
<td>1049</td>
<td>2386</td>
<td>59%</td>
<td>0.0</td>
<td>100%</td>
</tr>
<tr>
<td>XR_018330_1</td>
<td>PREDICTED: Homo sapiens similar to ATP-dependent DNA helicase 2 subunit</td>
<td>1305</td>
<td>1654</td>
<td>58%</td>
<td>0.0</td>
<td>90%</td>
</tr>
<tr>
<td>NM_018715_1</td>
<td>Homo sapiens regulator of chromosome condensation 2 (RCC2), mRNA</td>
<td>660</td>
<td>660</td>
<td>20%</td>
<td>0.0</td>
<td>93%</td>
</tr>
<tr>
<td>XR_019426_1</td>
<td>PREDICTED: Homo sapiens similar to ATP-dependent DNA helicase 2 subunit</td>
<td>1116</td>
<td>1211</td>
<td>47%</td>
<td>0.0</td>
<td>96%</td>
</tr>
<tr>
<td>XR_018655_1</td>
<td>PREDICTED: Homo sapiens similar to ATP-dependent DNA helicase 2 subunit</td>
<td>1131</td>
<td>1407</td>
<td>58%</td>
<td>0.0</td>
<td>86%</td>
</tr>
<tr>
<td>XR_018487_1</td>
<td>PREDICTED: Homo sapiens similar to ATP-dependent DNA helicase 2 subunit</td>
<td>1127</td>
<td>1217</td>
<td>47%</td>
<td>0.0</td>
<td>96%</td>
</tr>
<tr>
<td>XR_016730_1</td>
<td>PREDICTED: Homo sapiens similar to ATP-dependent DNA helicase 2 subunit</td>
<td>1125</td>
<td>1402</td>
<td>58%</td>
<td>0.0</td>
<td>86%</td>
</tr>
<tr>
<td>XR_016467_1</td>
<td>PREDICTED: Homo sapiens similar to ATP-dependent DNA helicase 2 subunit</td>
<td>1122</td>
<td>1217</td>
<td>47%</td>
<td>0.0</td>
<td>96%</td>
</tr>
<tr>
<td>XM_372262_5</td>
<td>PREDICTED: Homo sapiens similar to ATP-dependent DNA helicase 2 subunit</td>
<td>1305</td>
<td>1654</td>
<td>58%</td>
<td>0.0</td>
<td>90%</td>
</tr>
<tr>
<td>XR_018635_1</td>
<td>PREDICTED: Homo sapiens similar to ATP-dependent DNA helicase 2 subunit</td>
<td>1131</td>
<td>1407</td>
<td>58%</td>
<td>0.0</td>
<td>86%</td>
</tr>
<tr>
<td>XR_016487_1</td>
<td>PREDICTED: Homo sapiens similar to ATP-dependent DNA helicase 2 subunit</td>
<td>1127</td>
<td>1217</td>
<td>47%</td>
<td>0.0</td>
<td>96%</td>
</tr>
<tr>
<td>XR_016730_1</td>
<td>PREDICTED: Homo sapiens similar to ATP-dependent DNA helicase 2 subunit</td>
<td>1125</td>
<td>1402</td>
<td>58%</td>
<td>0.0</td>
<td>86%</td>
</tr>
<tr>
<td>XR_018426_1</td>
<td>PREDICTED: Homo sapiens similar to ATP-dependent DNA helicase 2 subunit</td>
<td>1116</td>
<td>1211</td>
<td>47%</td>
<td>0.0</td>
<td>96%</td>
</tr>
</tbody>
</table>

Multiple sequence alignment result using ClustalW is attached in the appendix

With the help of bioinformatics approaches, a novel gene could be discovered in the Xq22.1 region of the human X-chromosome. The sequence, structure and a possible role could also be predicted using several databases and tools. Biotechnology tools finally confirmed the results obtained from the previous gene prediction software's and thus, supported the possibility of a novel gene in the Xq22.1 region.

2.5.4 Novel Algorithm for identifying processed and non processed pseudogenes in Xq22.1

The chromosomal positions and the Pseudogene id of Xq22.1 were shortlisted in the table below.
<table>
<thead>
<tr>
<th>Chromosome Start/End</th>
<th>Short ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>95003452 to 95005580</td>
<td>O14917_1</td>
</tr>
<tr>
<td>95397304 to 95397773</td>
<td>P05092_76</td>
</tr>
<tr>
<td>966628869 to 96629322</td>
<td>P05092_77</td>
</tr>
<tr>
<td>96964020 to 96965678</td>
<td>Q14409_1</td>
</tr>
<tr>
<td>94316369 to 94318169</td>
<td>P12956_3</td>
</tr>
<tr>
<td>97224722 to 97225063</td>
<td>Q99880_1</td>
</tr>
<tr>
<td>97858274 to 97858606</td>
<td>Q16695_6</td>
</tr>
<tr>
<td>95507712 to 95508489</td>
<td>Q95478_6</td>
</tr>
<tr>
<td>99218221 to 99218769</td>
<td>Q9UDT4_111</td>
</tr>
<tr>
<td>98333374 to 98333575</td>
<td>Q9UHS9_1</td>
</tr>
<tr>
<td>98574203 to 98574448</td>
<td>Q9UHQ7_2</td>
</tr>
<tr>
<td>95398178 to 95400060</td>
<td>Q60216_1</td>
</tr>
<tr>
<td>95452296 to 95453379</td>
<td>P09651_49</td>
</tr>
<tr>
<td>95484643 to 95485590</td>
<td>P09651_50</td>
</tr>
<tr>
<td>96722638 to 96724405</td>
<td>P41743_1</td>
</tr>
<tr>
<td>97375970 to 97375970</td>
<td>Q969E4_1</td>
</tr>
<tr>
<td>98516740 to 98517357</td>
<td>Q969E4_2</td>
</tr>
<tr>
<td>95553789 to 95554525</td>
<td>Q9NUM9_1</td>
</tr>
<tr>
<td>96782622 to 96783213</td>
<td>Q9NWJ3_1</td>
</tr>
<tr>
<td>97799914 to 97800194</td>
<td>Q9H387_77</td>
</tr>
<tr>
<td>94483752 to 94484971</td>
<td>Q9NY97_1</td>
</tr>
<tr>
<td>98329992 to 98332069</td>
<td>Q9NWV2_1</td>
</tr>
<tr>
<td>98496435 to 98496773</td>
<td>Q96AT4_1</td>
</tr>
<tr>
<td>98501364 to 98502112</td>
<td>Q96HQ6_1</td>
</tr>
<tr>
<td>96868959 to 96869616</td>
<td>Q9BVZ1_3</td>
</tr>
<tr>
<td>97389170 to 97389511</td>
<td>Q9BXY8_1</td>
</tr>
<tr>
<td>96140251 to 96140807</td>
<td>Q43674_2</td>
</tr>
<tr>
<td>94752897 to 94753785</td>
<td>P08865_53</td>
</tr>
<tr>
<td>96905623 to 96906736</td>
<td>Q03924_1</td>
</tr>
<tr>
<td>98448806 to 98449693</td>
<td>P08865_54</td>
</tr>
</tbody>
</table>

Table 32- Pseudogene identified using novel algorithm

For each potential pseudo gene (or fragment) match, a number of steps have been taken to assess its validity as a pseudogene. These steps include checking for over counting and repeat elements, overlap on the genomic DNA with other homolog and cross
referencing with exon assignments from genome annotations. The resulting pseudogenes or pseudogenic fragments have then been assigned to the paralog family of the most homologous gene (or assigned to a singleton gene if the probe gene has no obvious paralog). Further verification was done using some of the online tools. (http://www.ensembl.org/Homo_sapiens/mapview?chr=X, http://www.pseudogene.org). Significant alignments were further analyzed for processed and non processed pseudogene. The first 26 compromise of processed and the remaining non processed pseudogene (in italics).
DISCUSSION
IV. Discussion

Statistical analysis indicated 8% isoform genes are present in Xq22.1, which is completely occupied by an Autophagy-related cysteine endopeptidase-2 protein (APG4A). APG4A has three isoforms (Isoform a, Isoform b, Isoform C).

The isoforms, SNP analysis for protein coding, the validation, variation status and impact of it were studied for APG4A using NCBI, SNP database. Protein structure prediction was done using MODELLER; active site prediction was done by using PASS and screening of suitable ligand for the protein using NCI database.

Complete annotation of it revealed interesting results which were incorporated for drug designing of APG4A. The novel gene prediction methods on this region are successful and couple of genes has been experimentally validated in vivo. Some of the pseudogenes were explored in Xq22.1 got very promising results.

2.5.1 Characterization and statistical analysis of genes on human X-chromosome, Xq22.1

The genomic region of Xq22.1 contains 40 protein coding genes out of a total of 46 genes. These protein coding genes were then statistically analyzed based on molecular weight, molecular class, biological process, gene localization and gene expression. Molecular analysis shows GPRASP1 has the highest molecular weight (156845 Da) and SMPX has the lowest (9560 Da).
In this genomic region, the molecular class of most of the genes remains to be unknown. Transcription regulatory protein is one of the major molecular classes observed in this genome segment. The statistical analysis of gene localization shows that most of the gene localizations are unknown. Many protein coding genes are localized in Nucleus and Cytoplasm. Some genes are localized in mitochondria and secretary granules. Most of the biological processes in the genome are unknown. Xq22.1 plays a major role in Nucleic acid metabolism. Protein metabolism and Signal transduction are also observed in this genomic segment. These genes were found to be expressed maximum in the testis. The protein coding genes were also found to be expressed in the Heart, Prostate, Spleen, Skeletal muscle, Brain and Ovary. Gene expression was found to be minimal in B cell and Small intestine tissues. Phosphorylation was the only Post translational modification observed in genomic segment Xq22.1. Out of the 40 protein coding genes, SYTL4 was found to be the only plasma protein. The statistical analysis reported localization, biological process and gene expression of the protein coding genes in the human genome Xq22.1. Isoforms and SNP analysis of APG4A mutation at the functional level could bring about a breakthrough in autophagy discovery.

The statistical analysis confirmed the characterization of each protein coding gene present in human genome Xq22.1. Analysis gives us an insight in identification and characterization of proteins in Xq22.1 and will throw light on origin of various genetic diseases. Basic research on Autophagy and APG4A may reveal probable drug target for APG4A.
2.5.2 Drug target for APG4A

The probable active site was identified as Cys 77 and suitable ligand was identified as 1, 2-dicarboxyethanethiolate, which shows considerable interaction with Cysteine residue. Various analogs of the original ligand are also selected and docked with the receptor. The binding energy of those analogs also increases the confidence of using those ligands as candidate for this particular protein. Further extensive research work on this protein will leads to a new milestone in Autophagy research.

One advantage of incorporating flexibility is that it enables a search without bias introduced by the initial model. This bias normally influences both the orientation and conformation of the ligand in the active site, which usually represents a local minimum generally more convenient and practical. A relationship, once quantified, can be used to estimate the properties of other molecules simply from their structures and without the need for experimental determination or synthesis.

The accurate prediction of enzyme-substrate interaction energies is one of the major challenges in computational biology. Protein-ligand docking methods aim to predict the binding energy of the protein-ligand complex given the atomic coordinates. In such calculations, both the protein and ligand can be treated as rigid bodies; alternatively, the ligand, the protein, or both molecules, can be completely or partially flexible.

In this work we have extensively worked on gene responsible for autophagy, a process of intracellular bulk degradation in which cytoplasmic components including organelles are sequestered within double-membrane vesicles that deliver the contents to the lysosome/vacuole for degradation. Reduced levels of autophagy have been described in
some malignant tumors. It was found that APG4A which is recently sequenced have a wide role in suppressing various cancers. The work performed, includes identifying the genomic localization, tissue specificity, functions, gene expression level in diseased and normal cell, relation with various cancers, understanding its role in various pathways, protein structure prediction, active site prediction and screening of suitable ligand for the protein. The novel gene APG4A of Homo sapiens was studied extensively via Insilco. It is believed that with a suitable target for inducing autophagy through APG4A, would be a real breakthrough in cancer therapy.

2.5.3 Identifying Novel gene and verifying experimentally using RT-PCR

Thus, using bioinformatics tools, a novel gene could be predicted and also, the amino acid sequence of the protein thought to be encoded by it could be retrieved from Augustus. This information was further utilized to model the protein even before the results could be obtained using the biotechnology methodology and techniques. The results obtained were confirmed in the laboratory which reassured that the predicted protein was a nuclear protein involved in DNA repair mechanisms and expressed only when there is damage to DNA of the cell, e.g. when the cells are subjected to certain kind of rays which alter the structure of DNA.

2.5.4 Novel Algorithm for identifying processed and non processed pseudogenes in Xq22.1

Processed Pseudogenes those matches with our analysis comprise >70% length of closest matching Human Ensemble Database Sequence in a continuous segment. It was also
possible to contribute classified Pseudogenes than that available on various online data. The hope behind this work is that it might reveal many secrets which might be a surprise to the scientific community. This work will also offers new insight on the evolutionary history of human genes and the stability of genome as a whole, reason for its variations in diseased conditions.