CHARACTERIZATION AND STATISTICAL ANALYSIS OF GENES ON HUMAN X-CHROMOSOME Xq22.1

VIDYA NIRANJAN¹, RIAZ MAHMOOD¹, PARESH KUMAR SINGH², KONTHOIJAM DEEPIKA², RAMESHWAR.G.K², JAYARAMA REDDY³

St. Joseph’s Post Graduate and Research Centre
¹Department of Bioinformatics and Biotechnology, Kuvempu University, Sankarghatta ,India
²Oxford College of Science, Bangalore, India.
³St. Joseph’s Post Graduate and Research Centre
E-mail: niranjan.vidya@gmail.com

ABSTRACT:

Human Xq22.1 is a genomic region, which has 20 mammalian rapidly evolving genes and involved in numerous genetic diseases such as XLMR, Fabry disease, pre mature ovarian failure, cancer etc. Many complex diseases may arise from quantitative, rather than qualitative differences in the gene products. Review, characterization and statistical analysis is performed on protein coding genes in human genomic region, Xq22.1. The analysis is based on literature review, molecular weight, molecular class, biological process, gene localization, gene expression, splicing events and mutational study. Role in biological processes, cellular component and molecular class of most of the genes in human genome Xq22.1 is unknown. Tracing the specific genes will help us in understanding the genetics of human disease associated with Xq22.1. Study on splicing events and SNP studies revealed interesting results on APG4A, the gene responsible for autophagy.

1. INTRODUCTION:

20 mammalian rapidly-evolving genes on human chromosome Xq22.1 in various protein families BEX, WEX and GASP were identified(1). (2) identified a novel human homeobox gene ESXR1 with that of mouse Esx1 homeobox gene. ESXR1 has been localized to human Xq22.1-q22.3, 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. (3) discovered a new member of the RAB family was identified through a public database search, which revealed that the gene is divided into three exons and spans approximately 7.2kb of the genome DNA of chromosome Xq22.1-q22.3 region.

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 (4). The analysis of protein coding genes were done using a combination of bioinformatics and experimental verification using HPRD (www.hprd.org) on molecular weight, molecular class, molecular function, biological process, localization , expression, isoforms, Proteolytic Cleavage and Single Nucleotide Polymorphisms (SNPs) of 40 genes reported.

Review on disease association on Xq22.1

Pandita A eta.al gave insight on the pathogenesis of gangliogliomas and associated rare malignant progression(5). The gene encoding the human protein is located on chromosome Xq22.1, involved in premature ovarian failure. Merhi ZO et.al revealed that premature ovarian failure due to Xp duplication and Xq deletion(6). Clarke NF eta.al 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 Xq (Xq22.1-->qter) and terminal Xp deletion accompanying sonographic detection of uterine growth restriction (7). A novel NXF gene cluster on Xq22.1, a good candidate gene X-linked Mental Retardation (XLMR)(8,9) and can be classified into syndromal and non syndromal and the observation confirmed the suspicion of a gene involved in growth hormone regulation being localized in Xq22.1.

(Scientist identified that Fabry disease which is a lysosomal disease caused as the result of the gradual accumulation of Glycosphingolipid in the alpha-galactosidase A enzyme in Xq22-1(10,11,12,13). Eng CM et.al (14, 15, 16) found an X-linked inborn error of glycosphingolipid catabolism 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 but 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.

(Autophagy-related cysteine endopeptidase A (APG4A), Ras related protein Rab-40A (RAB40A), Translocase of inner mitochondrial membrane 8 homolog A (TIMM8A) is located in Xq22.1. These genes 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 homeostatis 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 our detailed analysis is accentuated in a 4.2 Mb region of the human X-chromosome, Xq22.1 and in detail on APG4A.

RESULTS:

STATISTICAL ANALYSIS:

2. MOLECULAR WEIGHT ANALYSIS:

Statistical analysis is performed Molecular weight Vs Gene Symbol.

It shows 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 KIAA043, with highest levels in brain, kidney, and ovary.
SMPX (Small muscular protein) has a 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.

3. MOLECULAR CLASS:

These following molecular classes are with percentage identified in Xq22.1 are discussed below.

In Xq22.1, molecular classes of most of the genes are unknown (37%). Now at present, transcription regulatory protein (TCEAL1, TCEAL2) is a major molecular class (12%) identified in Xq22.1. Pillutla RC eta.al1999 (17) revealed that TCEAL1 is the gene which consists of three exons and two introns that codes 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.
4. GENE LOCALISATION:

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% genes 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 nucleus and alternative localization is cytoplasm (3%).

5. BIOLOGICAL PROCESS:
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) are the biological processes observed in genome Xq22.1 as 10%. 5% of biological process observed as transport (GPRASP1, SYTL4). Metabolic pathways (TIMM8A) and cell growth factor maintenance (MID2) are the biological processes observed as least (3%) in Xq22.1.

6. GENE EXPRESSION:

In gene expression analysis maximum gene expression about 30% is observed in Testis (TAF7L, FSHTRH1, APG4A, TCEAL3, NGFRAP1 and ESX1L).
Next to that, 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) shows similar gene expression 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 in the Ubiquitous (RAB9B). Finally the least expression of gene i.e. about 1% is observed in B cell (CSTF2) and small intestine (MID2).

7. ISOFORM ANALYSIS IN Xq22.1:
Statistical analysis of the genes present on Xq22.1 indicated that 8% isoform genes were present in this region. This percentage included an Autophagy-related cysteine endopeptidase-2 protein (APG4A). APG4A is has a three isoforms denoted as Isoform a, Isoform b and Isoform C.
8. GENETIC MAPPING by CLUSTALW Results:

APG4A

ISOFORMS
APG4A Isoform a (398 aa)

APG4A Isoform c (321 aa)

APG4A Isoform b (336 aa)

9. PROTEOLYTIC CLEAVAGE:

We now turn 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 protein and the 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 HsApg4A. 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 HsApg4A, 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 variation occur 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; and 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 SNP maps will help them identify the multiple genes associated with such complex diseases 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.

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 retroposable 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.

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

11 VALIDATION ANALYSIS:
SNPs assays validated directly by the submitter through the validation section showed the type of evidence used to confirm the variation.
Validation status of APG4A gene variant may be found by more than one method. 15% validation status of APG4A was found by 2-hit allele. 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.

12. DISCUSSION:

The genome Xq22.1 contains 46 genes in that 40 genes are protein coding genes. The statistical analysis is performed on Protein coding genes in human X-Chromosome, Xq22.1 based on molecular weight, molecular class, biological process, gene localization and gene expression. Molecular analysis shows GPRASP1 has a highest molecular weight (156845 Da) and SMPX has a lowest molecular weight (9560 Da).

In genome Xq22.1, the molecular classes of most of the genes are unknown. Transcription regulatory protein is one of the major molecular classes observed in this genome segment. The statistical analysis of gene localization shows most of the gene localization is unknown. Many protein coding genes are localized in Nucleus and Cytoplasm. Some genes are localized in Mitochondria and Secretary Granule. Most of the biological process in the genome is unknown. Nucleic acid metabolism is played a major role in Xq22.1. Protein metabolism and Signal transduction also observed in this genome.

The majority of the gene expression is observed in Testis. The protein coding genes are also expressed in Heart, Prostate, Spleen, Skeletal muscle, Brain and Ovary. The least gene expression is observed in B cell and Small intestine.

Phosphorylation is the only Post translational modification observed in genome Xq22.1. In 40 protein coding genes, SYTL4 is an only one plasma protein.

The statistical analysis reported localization, biological process and gene expression of the protein coding genes in the human genome Xq22.1. Isoform and SNP analysis of APG4A mutation at functional level can be a breakthrough in autophagy discovery.

13. CONCLUSION:

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

14. Acknowledgment: We acknowledge Kalai arasi and Prem kumar for their valuable help in data mining.
REFERENCES


16. Eng CM, Desnick RJ. Molecular basis of Fabry disease: mutations and polymorphisms in the human alpha-