Analysis of Copy Number Variations in Multigene Family in Human Genome

By

Mrs. Sana Ahmed

ISBN: 978-81-936988-0-8

eISBN: 978-81-936988-1-5

Published by

Associated Asia Research Foundation

www.aarf.asia

[Blank Page]

Analysis of Copy Number Variations in Multigene Family in Human Genome

Thesis Submitted to University of Mysore for the Award of the Degree of

Master of Philosophy

In

Zoology

By

Mrs. Sana Ahmed

Under the guidance

Prof.N.B.RamachandraM.Sc., Ph.D.
Coordinator for M. Sc. Genetics
Department of Studies in Zoology
University of Mysore, Manasagangotri
Mysore-570 006

2014

UNIVERSITY OF MYSORE

Department of Studies in Zoology

DR. N. B. RAMACHANDRA M. Sc., Ph.D. Manasagangothri, Mysore-570 006
Professor & Principal Investigator Email: nallurbr@gmail.com
Coordinator for M.Sc. Genetics ramachandra@zoology.uni-mysore.ac.in
Chairman-BOS in Clinical Research & Phone:(O) 0821-2419781 / 2419888
Clinical Data Management (R) 0821-2516056 (M)
9880033687
Chairman-Board of Studies in Zoology (UG&PG) http://www.ramachandralab.com/
Director- University of Mysore Genome Centre
http://umgc.uni-mysore.ac.in/

CERTIFICATE

I hereby certify that the thesis entitled*"Analysis of Human Copy Number Variations in Multigene Famlies*" submitted by **Mrs. Sana Ahmed** for the degree of **Master of Philosophy** in Zoology, University of Mysore, is a record of research work carried out by her during her stay at the Department of Studies in Zoology, University of Mysore, Mysore, under my guidance. It has not previously formed the basis for the award of any other degree/diploma of this or any other university. This research work or part of thereof has not been the basis for the award of any other degree, diploma, associateship, fellowship or any other similar titles.

Dr. N. B. Ramachandra

Guide

Date: 12.06.2014
Place: Mysore

CHAIRMAN

Mrs. Sana Ahmed
M. Phil Student

Department of Studies in Zoology
University of Mysore,
Manasagangotri,
Mysore-570006 India

DECLARATION

I Mrs. Sana Ahmed, declare that this thesis entitled ***"Analysis of Human Copy Number Variations in Multigene Famlies"*** submitted to University of Mysore for the award of degree of Master of Philisophy in Zoology, embodies the result of research work done by me, under the guidance of Prof. N. B. Ramachandra, Co-ordinator for M. Sc, Genetics, Department of Studies in Zoology, University of Mysore, Manasagangothri, Mysore-570006, India.

I further declare that this or part therefore has not been the basis for the award of any other degree, diploma, associateship or any other similar types previously.

(Mrs. SANA AHMED)

Acknowledgement

I am indeed one of the lucky persons to have got this oppurtunity to work under the guidance of Prof. N. B. Ramachandra, my research guide; Co-ordinator for M.sc Genetics; DOS in Zoology; University of Mysore. The enthusiasm he has for his research work was a motivation for me. I am grateful to you sir for your true support, guidance and help from the very first day of my research work and also for his consistent encouragement. I am thankful to you sir for having faith in me and providing me with a free hand to experiment and analyze and for having laid a very strong foundation for my research career.

I am also very much thankful to Prof. H. N. Yajurvedi; present Chairman and the former chairperson; Prof. V. A. Vijayan for

providing me with all the facilities and allowing me to carry out my research work in the Department.

Words of appreciation also go to my good friends and colleagues of Genetics & Genomics Lab, Dr. Avinash M V, Dr. Kusuma L, Megha Murthy, Sangeetha V , Somanna A N, Raviraj V S, Sareh Jahromi for being so supportive and helpful throughout my research work, which helped me to refine my work and make it better. I am also thankful to them for their friendly coordination in laboratory and their timely help during my course. I also thank all the non teaching staff, Department of Studies in Zoology for their cooperation during my research work.

I express my deep sense of gratitude for the patience, cooperation, inspiration of my husband and family members, without which i would have not been able to carry out this research. Thank you so much Mom, Dad, My sister and last but not the least My Husband for their unconditional love, kindness, Support and encouragement.

Mrs. SANA AHMED

ABSTRACT

Copy Number Variations (CNVs) presence alters the transcriptional and translational levels of genes by disrupting the coding structure and this burden of CNV seems to be a significant contributor to the phenotypic variations. Recurrent CNV events evolutionary have resulted in the conversion of genes to multi-genes families. Therefore it was necessary to assess the complexities of CNV burden on the multigenes. In view of this, the following objectives were addressed: To identify the prevalence of CNVs in multigene families in normal cohorts and to identify and associate population specific CNV- multigene families. A total of 1715 individuals from 12 populations were used for the present investigation for the CNV analysis. Analysis was performed using Affymetrix Genome-Wide Human SNP Array 6.0 chip and CytoScan High-Density arrays. A total of 44109 CNVs were identified. These CNVs contain 126190 genes consisting of 15185 singleton genes. Of these, ~56.65% multigenes-CNVs contained ~13.28% multigenes with a total of 902 (~5%) singleton multigenes. These multigene-CNVs were identified in 1703 (99.30%) individuals. This is a maiden report in the analysis of CNVs in multigene families identified in 12 populations across the globe. The findings of CNVs in multigenes contribute to the array of disease phenotypes seen in the families which demonstrate the importance of performing a high resolution assessment of genomic background even after the detection of rare and likely damaging CNVs. The methodology adopted here has a higher genome resolution but the number of the populations chosen for this study is less which becomes a limitation and requires further whole genome study on large extended families to identify more appropriate genes and relevant tests. These will at large help to rehabilitate, treat, control, manage and cure many diseases.

CONTENTS

List of Tables and Illustrations

TABLES

PREFACE

A genetic disorder or any genetic disease is caused by an abnormality in an individual's DNA. These abnormalities can range from minor mutations in a single gene to the duplication and deletion of an entire chromosome or set of chromosomes (Peng et. al, 2007). Genetic disorders are heritable, and are passed down from parents genes. Other defects may be caused by new mutations, due to changes by new or altered mutations or any changes to the DNA. In such cases, the defect will only be heritable if it occurs in the germ line. Genetic disorders may also be complex, multifactorial, or polygenic, meaning that they are likely to be associated with the effects of multigenes in combination with the environmental factors (Zollner et al., 2004). Multifactorial disorders include heart diseases, cancer, mental retardation and diabetes. Although complex diseases often cluster in families, they do not have a clear cut pattern of inheritance. This makes it difficult to determine a person's risk of inheriting or passing on these disorders. Complex disorders are also difficult to study, diagnose and treat because the specific factors that cause most of these disorders have not been identified yet. The most difficult challenge that lies ahead is the precise determination of genetic component of common but genetically complicated diseases.

Multigene families are a group of genes that are derived from the same organisms. These multigenes occur as a result of gene duplication which produces gene in pairs. If both the copies of genes are maintained in the next generation also then it will result in multigene families. A single ancestral gene is believed to give rise to multigene family. The main advantage of multigene family is that the members of the family will have nearly similar functions as they will encode for similar proteins. These multigene families further provide information about gene evolution and how these genes are related to diseases with help of several markers such as Single Nucleotide Polymorphisms (SNPs) and Copy Number Variations (CNVs). Among these markers, CNVs are the most common and powerful markers used in identification of the disease.

CNVs are the most common type of structural variations that are found in humans. CNVs were discovered in humans in early 1990s and were found to be wide spread throughout populations. CNVs account for 5% of the total human genome variations. By definition it can be made out clearly that the CNVs have either or no phenotypic consequences but they might have serious effects if the mutations are deleterious. CNV studies have clearly shown that it has the potential to influence a healthy individual into a diseased one. Hence many efforts are being made to find the co-relation between copy number changes and diseases.

The markers of choice that have emerged for whole-genome scans are SNPs. Although there are multiple sources of genetic variations that occur among individuals, SNPs are the most common type of sequence variation and are powerful markers due to their abundance, stability and relative ease of scoring (White Paper, Affymetrix 2009). Current study estimates of the total human genetic variations suggests that there are over 11 million SNPs with a minor allele frequency of at least 55 (Frazer et al., 2009) The international effort to characterize human haplotypes (Hap Map Project) in four major world populations has identified a standard set of common alleles, SNPs that have

provided the framework for new genomewide studies designed to identify the underlying genetic basis of complex diseases, pathogen susceptibility, and differential drug response (Conrad et al., 2006, Hancock et al., 2008).

In view of this, the following objectives were addressed:

1. **To identify the prevalence of CNVs in multigene families in normal cohorts.**
2. **To identify and associate population specific CNV- multigene families.**

1. REVIEW OF LITERATURE

Human genome is a complex structure in its organization with defined structural entities. The genetic variation sites can be used as markers to identify disease patterns in humans. This approach led to the successful identification of number of useful genes in rare monogenic disorders (Emmanuelle et al., 2008). Since the completion of the human genome sequence, efforts have been made to identify genes which are involved in common complex diseases (Antonarakis and Beckmann, 2006; Beckam et al., 2007). The human genome has been extensively studied when compared to the genomes of other organisms.

1.1 The Human Genome Project

The Human Genome Project is the first project of its kind to sequence the entire human genome. It was built on several debates, some doubting its efficacy and others arguing it would help expedite cancer research and in identifying mutations. Just about the same time, DNA markers were gaining importance in their ability to detect linkages in several disease pedigrees. Several research groups began with the use of DNA markers for human genome which actually

led to the discovery of new genes. These groups also worked to develop chromosomes markers for linkage analysis, with the coverage area attained up to several cMs. (Botstein et al., 1980) and Donis-Kellar et al., (1987) published their first linkage map of the human genome, which paved the foundation for the beginning of the Human Genome Initiative. Two major breakthroughs, one in the field of recombination DNA technology and the other in the development of artificial chromosomes led to the belief that the human genome initiative was feasible. The slow paced discovery of recombinant DNA technology in the mid 1970s followed by the development of several methods to clone large fragments of DNA in 1980s assisted in the cloning of DNA fragments, hundreds to several thousand base pairs in length followed by the development of artificial chromosomes which expanded the size to include large DNA inserts (Botstein et al., 1980). This simplified the cloning of the entire human genomes and reconstructions of the order of the cloned fragments.

Discovery of Polymerase Chain Reaction (PCR) allowed the rapid amplifications of short regions of DNA with much ease thus opening the doors of sequencing

and other analysis with extra ordinary high speed and precision. The Department of Energy (DOE) and National Institute of Health (NIH), USA came together for the first government sponsored genome research program in 1987. Post the initiation of the Human Genome Project, there was a need to mediate international scientific collaboration and to co-ordinate development in the field of genome research between the research groups across the globe. The Human Genome Organization (HUGO) was formed in 1988 to addresses the same (The International Human Genome Mapping Consortium, 2001).

The human genome sequencing project was the primary step towards basic research. The emphasis was on obtaining a complete and highly accurate reference sequence (1 error in 10,000 bases), largely continuous across each human chromosome. Knowing these sequences is critically important for understanding human biology and for applications to the other fields. A "working draft" of the human genome DNA sequences was completed in June 2000 and published in February 2001 (Bentley et al., 2000) (The International Human Genome Consortium, 2001). The working draft comprises of shotgun sequences data from the mapped clones, with gaps and ambiguities unresolved. Drafts sequence provides a foundation for obtaining the high quality finished sequences and also is a valuable tool for researchers hunting disease genes. The Human Genome Project was completed in 2003 with the total funding close to US$ 3.8 billion (Shendure and Aiden, 2012).

The International Genome project was a hierarchical mapping and sequencing strategy to construct the working draft of the human genome. This clone based approach involved generations of overlapping series of clones that covered the entire genome. These overlapping series are fingerprinted on the basis of the patterns of fragments generated by restriction enzyme digestion. This whole genome clone based map assisted the sequencing of the human genomes in many ways. It started with the use of fingerprinted BAC maps which helped selecting the clones for sequencing. On the other side, challenges in sequence assemblage were minimized by restricting random shotgun sequencing to individual clones (International Human Genome Sequencing Consortium, 2004). This way of accurate fingerprinting and sizing each clone enabled the accuracy of the shotgun sequence assembly. Similar clone based map approach were even employed during the sequencing of the Drosophila melanogaster, Saccharomyces cerevisae, Caenorhadbitiselegans and Arabidopsis thaliana (Bentley et al., 2000).

During the last 10 years, the use of molecular markers, revealing polymorphisms at the DNA level, has been playing an increasing part in genetic studies. Amongst others, the micro satellites DNA marker had been the most widely used, due to its easy use by simple PCR, followed by a denaturing gel electrophoresis for allele size determinations and to the high degree of information provided by its large number of alleles per locus (Collins et al., 1998; Dib et al., 1996). The term “polymorphism” is often used in rather vague and facile ways. Technically, a polymorphic locus is one whose allele or variants are such that the most common variant among them occurs with less than 99% frequency in the population at large (e.g. if the locus is bi allelic, the rarer allele must occur with a frequency greater than 1% population) (The International SNP Map Working Group, 2001; Collins et al., 1998). However, the use of polymorphism in modern genetics initiatives ultimately emanated from the studies of physiological and biochemical variations, such as that exhibited by protein isoforms and blood group antigens. These variations are thought to arise from actual DNA sequence variations and when this was confirmed, a number of crucial questions arose.

First; what kind of alterations exists in the genome that can be understood to impact phenotypic variations? Second; how are such variations maintained and what is their behavioral populations? Third, how can one "link relevant genetic variations with phenotypic variations?" And fourth, how are such alterations ultimately translated into overt biochemical and phenotypic variations? In theory, the answer to these questions seem straight forward to obtain merely conducting studies, examining the association of DNA variants with either the presence or absence of different phenotypes among individuals or among individuals from different populations. In practice, however things have proven more difficult for many reasons. One simple reason for this difficulty is that the very definition of a DNA variant ranges from a single base pair to several hundred base pairs.

(The International SNP Map Working Group, 2001). Thus it is not always straight forward to actually identify a specific genomic site that results in phenotypic variation.

Human Genome Sequence Variation Goals are; to develop technologies for rapid, large scale identification and scoring of SNPs and other DNA sequence variants, identify common variants in the coding regions of the majority of identified genes during the 5-year period, create a SNP Map of at least 100,000 markers, develop the intellectual foundations for the studies of sequence variation and create public resources of DNA samples and cell lines.

1.2 Single Nucleotide Polymorphisms (SNPs)

Although more than 99% of human DNA sequences are the same across the populations, variations in DNA sequence can have a major impact on how humans respond to disease; environmental insults such as bacteria, viruses, toxins and chemicals and drugs and other therapies. Variations in the name of a new marker type, named SNP gained importance even though it is only a bi allelic type of marker. SNP's (pronounced "snips"), are the most common type of genetic variations among people (Shastry, 2002). Each SNP represents a difference in a single DNA building block, called a nucleotide. For example a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in certain stretch of DNA. The first wave of information from the analysis of the human genome revealed SNPs to be the main source of genetic and phenotypic human variations (Brookes, 1999). Polymorphisms arise as a result of mutation. The different types of polymorphisms are typically referred to by the type of mutation that created it (Lercher and Hurst et al., 2002). The simplest types of polymorphisms result from a single base mutation which substitutes one nucleotide for another. For example, the first systematic studies of single base variations were pursued through the identification of restriction enzymes sites, where a single base pair change could result in the loss or gain of a restriction site. Digestion of a piece of DNA containing the relevant site with an appropriate restriction enzyme could then distinguish alleles or variants based on resulting to as “restriction fragment length polymorphisms (RFLP) (Botstein et.al.1980). Other SNP’s which do not directly create or destroy a restriction site, have been identified, often by creating restriction sites via PCR primer design, by oligonucleotide probing or by direct sequencing (Taylor et.al.,2001). Recent technological advances have greatly improved the ease and sophistication of such identification processes. Although the frequency with which SNPs occur over the genome is certainly much greater than that of RFLPs alone, precise estimates are difficult to determine and often vary across different populations and genomic regions. Some studies have suggested that SNPs can be found, on average, every 0.3-1 kilo bases (kb) within the genome, although most data used to address the question of SNP frequency were derived from studies of SNPs within specific genes and thus are likely biased. Thus, it is not clear whether or not current estimates can be extrapolated to the rest of the genome and to population other than those studied. Whatever the actual frequency of SNPs across the genome is, it is known to be greater than any other type of polymorphisms. SNPs occur normally throughout a person's DNA. They occur once in every 300 nucleotide on average, which means there are roughly 10 million SNPs in the human genome. Most commonly, these variations are found in the DNA between genes (Wang and Moult, 2001). They act as biological markers, helping scientists locate genes that are associated with disease. When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the genes function. Most SNPs have an effect on health and development. Researchers have found SNPs that may help predict an individual's response to certain drugs, susceptibility to environmental factors such as toxins and risk of developing particular diseases (Wang and Moult, 2001). Current estimates of the total human genetic variation suggest that there are over 11 million SNPs with a minor allele frequency of at least 5% (Howland, 2012).

1.3 Copy Number Variations (CNVs)

The advent of genome scanning technologies has now uncovered an unexpectedly large extent of what can be termed as "structural variation' in the human genome (Itsara et.al., 2010). This comprises microscopic and more commonly, sub microscopic variants, which include deletions and large scale copy number variants collectively termed CNVs or copy number polymorphisms

(CNPs) as well as insertions, inversions and translocations (Itsara et.al.,2010). Rapidly accumulating evidence indicates that structural variants can comprise millions of nucleotides of heterogeneity with every genome and are likely to make an important contribution to human diversity and disease susceptibility (Malhotra and Sebat, 2012). Genomic structural variants (SVs) are abundant in humans, differing from other forms of variation characterization, the nucleotide resolution architecture of most SVs remain unknown. Unbalanced SVs or CNVs, involving large scale deletions, duplications and insertions form one of the least well studied classes of genetic variations (Malhotra and Sebat, 2012). The fraction of the genome affected by SVs is comparatively larger than accounted for by SNPs implying significant consequences of SVs on phenotypic variation. SVs have been associated with diverse diseases, including autism, Schizophrenia and Crohn's disease. Furthermore, locus specific studies suggest that diverse mechanism may form SVs de novo, with some mechanisms involving complex rearrangements resulting in multiple chromosomal breakpoints (Girirajan et al., 2011; Malhotra and Sebat, 2012).

CNVs account for a major proportion of human genetic polymorphism and have been predicted to have an important role in genetic susceptibility to common disease (Girirajan et al., 2011). Chromosomal rearrangements can cause particular rare disease and syndromes and recent reports have suggested a role for rare CNVs either individually or in aggregate in susceptibility for a range of common diseases, notably neurodevelopmental diseases (Girirajan et al., 2011). So far there have been relatively few reported associations between common diseases and common CNVs, which might simply reflect incomplete catalogues of common CNVs or the lack of reliable assays for their large scale genotyping. The human genome is enriched in interspersed segmental duplications that sensitize approximately 10% of our genome to recurrent micro deletions and micro duplications as a result of unequal crossing over. Studies of common complex genetic disease show that a subset of these recurrent events plays an important role in autism, schizophrenia and epilepsy. The genomic hot spot model may provide a powerful approach for understanding the role of rare variants in common disease (Sharp et al., 2006).

DNA CNVs represent a considerable source of human genetic diversity. A global map of CNV in the human genome has been drawn up which reveals not only the ubiquity but also the complexity of this type of variation (Girirajan and Eichler, 2010). Thus two human genomes may differ by more than 20Mb and it is likely that the full extent of CNV still remains to be discovered. Nearly 3000 genes are associated with CNV (Kehrer-Sawatzki, 2007). This high degree of variability with regard to gene copy number between two individuals challenges definition of normality. Many CNVs are located in regions of complex genomic structure and this currently limits the extent to which these variants can be genotyped by using tagging SNPs. However, some CNVs are already amenable to genome wide association studies so that their influence on human phenotypic diversity is understood. The enormous extent of this type of variation has been documented by Redon et al., (2006) who established a genome wide CNV map in humans. In this hallmark study, which has the highest coverage reported so far, 270 human individuals from four populations (the Hap Map collection) were analyzed using SNP arrays and comparative genomic hybridization with arrays of genomic clones covering 94% of the euchromatin portion of the human genome (Whole Genome Tile Path(WGTP)array). Both methods complement each other, although SNP analysis tends to detect smaller CNVs whereas the WGTP platform identifies larger CNVs (>40kb). The latter approach is however more effective in tracing CNVs in genomic regions of complex structure, which are not sufficiently tagged by SNPs. If the results of both the methods are combined, ~43% of all copy number variable region s (CNVRs) were detected by both methods and in more than one individual. The 1447 CNVRs identified over 12% (360Mb), far more than the amount of genetic material contained in the largest human chromosome (Redon et.al., 2006). The average length of the genome found to be under copy number variables was 24 Mb on the WGTP array and 5Mb using SNP-based platform. This indicates that two individual human genomes may differ by more than 20 Module to the CNV. Compared to the 5-6 million genotyped SNP in phases 1 and II of Hap Map, the CNV mediated variation could thus be five times higher than the variations due to SNPs alone. The majority of CNVs identified by Redon et.al. (2006) are biallelic polymorphisms. However more complex patterns of variations were also observed, particularly in the region enriched with segmental duplications (Redon et al., 2006). In total 24% of CNVRs are associated with segmental duplications suggesting that non allelic homologous recombination between duplications has been frequently involved in the genesis of these CNVs.

1.4 The Haplotype Mapping Project (Hap Map)

Since understanding the relationship between the genotype and phenotype is one of the central goals in biology and medicines, the International Hap Map project was initiated to catalogue both allele frequency and to identify the co-relations patterns between the nearby variants in other words linkage disequilibrium (LD) across several populations. For 3.5 million SNPs the International Hap Map Project has made an effort to characterize human haplotypes (Hap Map Project) in four major world populations identifying a standard set of common allele SNPs that have provided the frame work for the new genome wide studies designs to identify the underlying genetic basis of complex diseases, pathogen suscepecibility, and differential drug response (The

International hap Map Consortium, 2005.; Sabeti et al., 2007).

The international Hap Map Project was a partnership of scientist and funding agencies from Canada, China, Japan, Nigeria, United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. The International Hap Map Project is collaboration among researchers at academic centers, non-profit biomedical research groups and private companies (Sabeti et al., 2007).

Most of the common haplotypes occur in all human populations; however, their frequencies differ among populations (The International Hap Map Consortium, 2010). Therefore, data from several populations are needed to choose tag SNPs. Pilot studies found sufficient differences in haplotypes frequencies among populations samples from Nigeria (Yoruba), Japan, China and the U.S.(residents with ancestry from Northern and western Europe, collected in 1980 by the center d'Etude du Polymorphism Huamin (CEPH) and used for other human genetic maps) to warrant developing the Hap Map with large scale analysis of haplotypes in these populations. The Hap Map developed with information from these populations has been useful for all populations in the world. About 4.9 terabytes (TB) of DNA sequence was generated from the samples of the Hap Map Project (The International Hap Map consortium 2010).

1.5. 1000 genome project

Since the completion of the Human Genome Project, advances in human genetics and comparative genomics have made it possible to gain increasing insight into the nature of genetic diversity (The 1000 Genomes Project Consortium, 2012). Improvement in sequencing technology such as the "next-gen" sequencing have sharply reduced the cost of sequencing. To understand how processes like the random sampling of gametes, structural variations, CNVs retro elements, SNPs and natural selection have shaped the level and pattern of variation within species and between species, and also in order to continue providing a deep characterization of human genome sequence variation to investigate the relationship between the genotype and phenotype, the 1000 Genomes project was proposed. It is the first project to sequence the genomes of a larger number of people to provide a comprehensive resource on human genetic variations. It was planned to sequence the genomes of at least one thousand anonymous participants from a number of different ethnic groups within three years, using newly developed technologies which were faster and less expensive. The project unites multidisciplinary research teams from institutes around the world, including China, Italy, Japan, Kenya, Nigeria, Peru, the United Kingdom, and the United States. Each Institute contributes to the enormous sequence dataset and to a refined human genome map. These teams sequenced samples from African Caribbean in Barbados, Hap Map African ancestry individuals from South Western US, Chinese Dia in Xishuangbanna, China, CEPH individuals, (CHB) Han Chinese in Beijing, Chinese in metropolitan Denver, (CHB) Han Chinese South Columbia, Hap Map Finnish individuals from Finland, British individuals from England and Scotland (GBR), Hap Map Guajarati individuals from Texas, Iberian populations in Spain, JPT Japanese individuals Kinh in Ho Chi minh City from Vietman, (LWK) Luhya individuals, Hap Map Maasai individuals from Kenya, Hap Map Mexican individuals from LA,

California, Peruvian in Lima from Peru, Puerto Rican in Puerto Rico, Tuscan individuals, (YRI) Yoruba individuals (Altshuler et al., 2010).

The aim is to determine genotypes and provide precise haplotypes information on all forms of human DNA polymorphisms from numerous human populations (Altshuler et al., 2010). Five major population groups with ancestry from Europe, East Asia, South Asia, West Africa and the Americans are chosen to characterize over 95 % of variants that are in genomic regions and are accessible to current high throughput sequencing. Analysis to identify and genotype sequence changes differed with the variant types. All the three projects shared the workflows with the following for features of discovery, filtering, genotyping and validation (The 1000 Genome Project Consortium, 2012).

The 1000 Genome project conducted studies in three subprojects. In the trio project, coverage of ~42X was obtained per individual across six individuals with about 2.3 GB., and provided 3.6 X coverage per individual across 179 individuals identifying 14.4 million SNPs, 1.3 million inDels and over 20,000 CNVs (The 1000 Genomes Project Consortium, 2012). In the exons sub project, about 56X

coverage was obtained per individual, the highest compared to the other two sub projects obtained across 697 individuals identifying 12,758 SNPs and 96 inDels. The variations identified in all these projects were found to be unevenly distributed across the genome, specifically in regions such as HLA region and subtelomeric regions. Populations with African ancestry contributed the largest number of variants and contained the highest fraction of novel variants, reflecting the greater diversity in African populations. An excess of lower frequency variants in the exons project was observed, reflecting purifying selection against weakly deleterious mutations and recent population growth (Altshuler et al., 2010)

The characteristics that make SNPs useful markers for genetic studies also make SNPs powerful markers for additional biological applications such as the analysis of populations and admixture structure (Johnson et al., 2001; Kidd et al., 2008) and DNA copy number changes, whereas the CNV markers are useful for detecting loss of heterozygosity, deletions, uniparental disomy and gene amplifications (Kidd et al., 2010; Medford et al., 2008). The use of microarray chips (such as Affymetrix Genome -Wide SNP 6.0, Illumina Bead Array) for genome wide scans identifies mainly two types of structural genome variations, SNPs and CNVs. These variations are widespread in the human genome and are a significant source of human genetic variation accounting for disease and population diversity.

1.6 The Encode project

Interpreting the human genome sequence is one of the major scientific endeavors. In February 2001, when the human genome reference was initially released (Lander et al., 2001), understanding of the encoded contents was surprisingly limited. It was perplexing to many in the scientific community when they realized that the human genome contains only ~21,000 distinct protein coding genes (Claverie, 2011; Clamp et al., 2007; Hollon, 2001; Pennisi, 2003), as other less complex species like the nematode Caenorhadbitis elegans were known to have a similar number of protein coding genes (Hillier et al., 2005). It quickly became apparent that the developmental and physiological complexity would not be explained solely by the number of protein coding genes, and the quest to understand the contents of the human genome began full force. The

Encyclopedia of DNA Elements (ENCODE) Project was launched in September of 2003 with the daunting task of identifying all the functional elements encoded in the human genome sequence. To accomplish this task, the National Human Genome Research Institute (NHGRI) organized The ENCODE Project Consortium, which consisted of an international group of scientists with diverse expertise in experimental and computational methods for generating and analyzing high-throughput genomic data (The Encode Project Consortium 2004). During the initial four years, the consortium conducted a pilot project which focused on annotating functional elements in a defined 1% of the human genome consisting of ~30 Mb divided among 44 genome regions. On June 14, 2007, a report summarizing the findings of the pilot project revealed pervasive transcription of the human genome, with the majority of nucleotides represented in transcripts in at least a limited number of cell types at some time (The ENCODE Project Consortiums 2007). Many of these transcripts comprised novel non coding RNA genes. Importantly, The ENCODE Pilot Project assigned function to 60% of the evolutionary constrained bases in the 44 genomic regions and identified many additional functional elements seemingly unconstrained across mammalian evolution. Integration of the various experimental data generated by the ENCODE Pilot Project provided further insights into connections between chromatin structure (modifications and accessibility) and gene expression (The ENCODE Project Consortium 2007; Koch et al., 2007; Thurman et al., 2007; Zhang et al., 2007) and the timing of replication (Karnani et al., 2007).

1.7. Micro Array Analysis

Whole-genome genotyping (WGGT) arrays have become an important tool for discovering variants that contribute to human disease and phenotypes. The two primary applications of this technology, genome-wide association studies (GWAS) and CNV analysis, have helped researchers begin to unravel the complex genetic architecture behind diseases such as diabetes and Crohn's disease, and traits such as hair and eye color. Microarray based gene chip offer researchers the flexibility to genotype samples with hundreds of thousands to millions of markers that deliver dense genome wide coverage with the most

up-to-date content. Markers on the chips are strategically selected to provide maximum coverage of the genome for both association testing and copy number detection. Whether polymorphisms come from publically available databases or new discoveries, it allows researchers to combine existing marker sets with new, unique content on a single chip, increasing efficiency and cost effectiveness in study design. Designed to make large scale genotyping affordable, these customizable arrays include an informative backbone of tag SNPs enabling researchers to tailor studies to specific populations or research goals. These microarrays can also be used to quickly and easily obtain baseline sample database for sample tracking and QC, as well as perform a variety of downstream applications such as common variant and CNV detection studies. Genome-wide association studies, which predisposing allele co-segregates with a particular allele of a SNP, have been hampered by the lack of whole- genome genotyping methodologies (Bush and Moore, 2012). As new genotyping technologies develop, coupled with ongoing studies into LD patterns and haplotypes block structure across the genome, improvements in the design and power of

association studies was feasible (Hirschhorn and Daly, 2005; Hindorff et al., 2009). The Genome -Wide Human SNP 6.0 microarray chip is an affordable tool to examine the role of CNV in disease by combining high powered SNP genotyping with highly accurate and sensitive detection of copy number state across the human genome. The combined density of SNPs and non-polymorphic probes on the SNP array provides high coverage of CNVs in particular those containing few SNP's that can be genotyped and enable to detect up to 10 times more copy number changes than competing platforms.

The SNP Array 6.0 contains more than 900,000 non polymorphic probes and 906,600 SNP probes for copy number analysis. All probes on the array are designed to test sequences present on Nsp I or Sty I restriction enzyme fragments of ~200 - 1,100 bps that are amplified using the Genome-Wide Human SNP Nsp /Sty Assay Kit 5.0/6.0. To obtain highly accurate and precise copy number intensity measurements, the linearity of response to copy number dosage was used in conjunction with probe spacing considerations to select a set of high- performing copy number probes that evenly span the genome and

specifically target regions of known CNVs. Probes for assessing copy number were empirically selected using a screen on a 13 million -probe array set designed against the fraction of the genome present on Nsp I fragments of a particular size.

GWAS seek to identify variation in the human genome that underlies a particular disease, drug response, diagnosis or prognostic outcome. The cataloging of human variation and subsequent association analysis has traditionally focused on SNPs. The assessment of common SNP variation in human disease has proven fruitful; more than 50 common variants have been found to be associated with disease such as type 2 diabetes, cardiac and immunological disease. However, recent work has demonstrated that other types of genomic variation - such as CNPs and CNVs (Girirajan et al., 2011) play a significant role in determining phenotype in common diseases and are likely to be found at reasonably high frequencies in the population at large. CNVs accounts for up to 4 Mb of normal genetic differences, compared to roughly 2.5 Mb for SNP variation (Gamazon et al., 2011). Many examples of disease are known to be

associated with copy number changes, including psoriasis, autism, lupus glomerulonephritis, as well as HIV infection and progressions. However, cost-effective, genome -wide methods for analysis of CNVs often fail Hardy-Weinberg and Mendelian inheritance checks, and are therefore not represented on most commercially available microarrays that principally interrogate SNPs.

1.8 Genome-Wide Copy Number Scans for Complex Diseases.

The extreme genetic heterogeneity of common complex disease, including autism and schizophrenia, and the high de novo mutations rate have hindered linkage studies of inherited susceptibility loci. Initial discoveries of CNVs in complex disease were based on locus-specific case-control association models in which deletions and duplications for a particular region of the genome were expected to be more represented in cases than healthy controls. The discovery of CNVs in autism and schizophrenia was also based on a new mutation model where only de novo variants were considered pathogenic (Sebat et al., 2007, Xu et al., 2008). Two pioneering studies exemplify this model. Sebat et al., (2007)

and colleagues utilized the representational oligonucleotide microarray analysis of CNV detection methodology in families affected with autism and comparing it with control families. They identified enrichment for de novo variants in cases compared with unaffected controls. This study also observed a higher incidence of deletions in cases with autism when both CNVs in controls were duplications. Similarly, Walsh et al., (2008) tested the total genome wide mutational burden in individuals with schizophrenia and compared them with ancestry matched controls and identified that, compared with only 5 % frequency in control, CNVs of recent origin accounted for 15 % adult onset (p=0.0008) and 20% of early onset schizophrenia (p=0.0001).

Pathway analysis showed that these mutations disrupted genes involved in neuronal signaling networks including glutamate and neuregulin pathways. Some studies did not find enrichments for large, rare CNVs in cases compared with controls, as described by Walsh and colleagues (2008). For example, Shi et al., (2008) analyzed 155 cases with schizophrenia and 187 matched controls- all recruited from the Han Chinese population in Shanghai. No significant

enrichment was observed for rare CNVs (>100 kbp) in case cohort compared with controls in this study. No increase in rare CNV (.500 kbp) burden was also reported by Ikeda and colleagues (2010) after analyzing patients with schizophrenia and controls subjects from Japan. Similarly, Glessner and colleagues (2010) analyzed CNV data from schizophrenics and healthy controls of European ancestry and then followed up the positive findings in another set of cases of schizophrenia and controls. They were unable to replicate the previously reported over representation of rare CNVs affecting many genes in schizophrenia compared with controls. Need and colleagues (2009) also tested the large CNV enrichment in schizophrenia cases along with matched controls. They could not provide strong support for the hypothesis that schizophrenia patients have a significantly greater load of large (>100 kbp), rare CNVs. It is important to note that such observations could be due to differences in sample quality, cell lines artifacts, probe resolution, GC content, lack of genotype information, sub phenotype characterization and clinical heterogeneity, age of onset of disease, and platform -specific biases. In all of these studies, previously

reported loci, including 1q21.1, 15q11.2, 15q13.3 and 22q11.2 and NRXNI deletions, were supported, although at a nominal significance.

Table 1: Genome sequencing projects on humans till date, modified after Aiden (2012)

SI. No.	Genome Sequencing Projects	Project Duration
1.	The Human Genome Project	1990-2003
2.	J.Craig Venter (HuRef)	2003-2006
3.	James D. Watson	2003-2006
4.	African Individual (NA18507)	2007
5.	Chinese individual	2007
6.	Seong-Jin Kim (a Korean individual)	2009
7.	Korean Individual 2	2009
8.	African Individual (NA18507)	2007
9.	Acute Myeloid Leukemia Genome	2007-2010
10.	69 Genomes By Complete Genomics	2011-2012
11.	52 year old Indian	2008
12.	1000 Genome Project	2011-current

1.9 Multigene families

Multigene families comprises of genes that are identical or having similar sequences and the similarity can be either for the entire sequence or partial, limited to specific domains. Recurrent CNV events evolutionarily have resulted in the conversion of genes to multigene families. These multigene families are generated by continuous genomic rearrangements caused by duplications, deletions and inversions in the genome. These multigene families are seen scattered across chromosomes or localized at one place (Nei and Rooney, 2005). There are >10 such multigene families in human genome. The genes for alpha and beta chains of the mammalian hemoglobin molecule are coded by multigene families on chromosomes 16 and 11. Multigene families of actins, Immunoglobulin's, interferon's, tubulins, hemoglobin's and histones are seen scattered and conserved (Bhowmick et al., 2007). The most prominent in the human genome, the r-RNA genes, alone has 2000 genes for 5S r-RNA. The most overrepresented category was the family consisting of alpha-amylase multigene family (MY1 and AMY2) located on chromosome 1. The second most frequent being the uridine glucoronosyltransferase (UGT) gene family divided into two sub families, UGT1 and UGT2 located on 2q37 and 4q13 respectively. Similar to UGT 1, the region for the protocadherin beta (PCDHB) gene family encodes 16 different proteins with variable N-termini. Olfactory Receptor (OR) gene family among these multigene families is the largest, occurring in both human and lower primates (Huminiecki and Wolfe, 2004).

1.10 Evolution of Multigene Families

Gene duplication is the major cause of the evolution of multigene families Chromosomes with more different genes are said to be more fit than others. Mutations either advantageous or harmful and duplications or deletions also affect the evolution pathway. Gene duplication with successive variations is thought to have played very important role in the evolution. If a gene is duplicated, the selective control becomes less for the extra copy, and it evolves with a different function, while the original function of the gene is kept in the other copy. Thus, gene duplication with variations is one of the simplest ways to acquire a new function and is thought to be employed many times during the

evolution. Gene duplications are not restricted to ancient times. Changes of gene number are observed in closely related species in many taxa. Thus, gene duplications are still ongoing evolutionary processes.

1.11 Multigene diseases

Inflammatory bowel diseases (IBDs) – a broad classification that includes Crohn's disease and ulcerative colitis – are multigenic disorders that are linked to various environmental components. Despite this complexity, considerable progress has been made in unraveling their pathogenesis. The genetics of IBDs point to roles for epithelial barrier function, innate and adaptive immunity in pathogenesis, while key environmental factors include commensally bacteria which may provoke impairment of regulatory immune responses in the host. The advances made in understanding IBDs show that it is possible in dissecting the pathogenesis of complex disorders. In view of this, analysis of genetic variations in multigene families of different populations has been carried out in the present study.

2. MATERIALS AND METHODS

For this study, a total of 1715 individuals involving 43 normal members from randomly selected twelve families residing in Karnataka, India with different age group members ranging from 13-73 years, 270 HapMap samples covering CEU (CEPH collection), CHB (Han Chinese in Beijing, China), JPT (Japanese in Tokyo, Japan) and YRI (Yoruba in Ibadan, Nigeria) populations, 31 Tibetan samples, 155 Chinese samples, 472 of Ashkenazi Jews replicate I (AJI), 480 of Ashkenazi Jews replicate II (AJII), 204 individuals from Taiwan, 55 from Australia and 64 from New World population (Totonacs and Bolivians), were selected for the CNV analysis in the genome. 5ml EDTA blood was collected from each member of the Indian study group and genomic DNA was extracted using Promega Wizard® Genomic DNA purification kit. The isolated DNA was quantified by Bio-photometer and gel electrophoresis. This research was approved by the University of Mysore Institutional Human Ethics review committee (IHEC). Written informed consent was obtained from all sample donors and the IHEC approved the sample consent procedure. Written informed consent was obtained from parents/guardians in the cases of participants being minors. The 270 individuals sample data from the four populations was obtained from the International HapMap Consortium (The International HapMap Consortium, 2003). The samples for the HapMap come from a total of 270 people: the 30 both-parent-and-adult-child trios from the Yoruba people in Ibadan, Nigeria, 45 unrelated Japanese individuals in Tokyo, 45 unrelated individuals Han Chinese in Beijing, and the 30 both-parent-and-adult-child trios from CEPH. The raw, unprocessed data from Affymetrix Genome Wide SNP 6.0 array for the 31 individuals from Tibet population, submitted by Simonson et al, (2010) and the remaining populations except India were obtained from the Array Express Archive of the European Bioinformatics Institute.

Genotyping

Genome-wide genotyping was performed using an Affymetrix Genome-wide Human SNP Array 6.0 chip and AffymetrixCytoScan® High-Density (HD) Array having 1.8 million and 2.6 million combined SNP and CNV markers with the median inter- marker distance of 500-600 bases. These chips provide maximum

panel power and the highest physical coverage of the genome (Affymetrix, Inc. Data Sheet, 2009). Genotyping quality was assessed using Affymetrix Genotyping Console Software. Copy Number Analysis Method offers two types of segmenting methods, univariate and multivariate. These methods are based on the same algorithm, but use different criteria for determining cut-points denoting CNV boundaries.

Algorithms for Copy Number state calling BirdSuite (v2)

Bird Suite (Birdsuite Algorithm, 2010) is a suite originally developed to detect known common CNPs based on prior knowledge, as well as to discover rare CNVs, from Affymetrix SNP 6.0 array data. To do this, it incorporates two main methods; the "Birdsuite" algorithms and the "Canary" (White Paper: Affymetrix®Canary Algorithm, 2008). The Birdsuite algorithm uses a Hidden Markov model (HMM) approach to find regions of variable copy number in a sample. For the HMM, the hidden state is the true copy number of the individual's genome and the observed states are the normalized intensity measurements of each array probe. CNV calls from the Canary and Birdsuite algorithms were

collated for each sample, and kept as long as they met the following criteria: i) Birdsuite calls with a log10 of odds (LOD) score (Odds Ratio) greater than or equal to 10 (corresponding to an approximate False Discovery Rate of ~5%), ii) Birdsuite calls with copy number states other than 2 were retained; iii) Canary CNP calls with CN states different from the population mode were retained.

Canary

CNP analysis was performed using the Canary algorithm. Canary was developed by the Broad Institute for making copy number state calls in genomic regions with CNPs. Canary algorithm computes a single intensity summary statistic using a subset of manually selected probes within the CNP region. The intensity summaries are compared in aggregate across all samples to intensity summaries previously observed in training data to assign a copy number state call.

Genotyping Console

After processing CEL files and the Birdseed to call genotypes, we used the Genotyping Console (GTC v.3.0.2) to detect CNVs from the Affymetrix 6.0 array for samples that passed initial QCs. The default parameters of >1kb size and

>5 probes in this algorithm were used.

Data Analysis

Genome-wide CNV study was carried out using SVS Golden Helix Ver. 7.2 (Bozeman, 2013) and Affymetrix Genotyping Console software as prescribed in their manuals (Affymetrix, 2005; 2007; 2008). Eigenstrat method was used to avoid possibility of spurious associations resulting from population stratification. Bonferroni correction was employed for multiple testing and the corrected data were then used for CNV testing. Bonferroni methods for population data genotyped on the Affymetrix 6.0 platform was α = 0.05 thresholds between 1 × 10-7 and 7 × 10-8.

Analyzing the collated data from both BirdSuite and Canary algorithms increased the stringency on those meeting the CNP calls with a log10 of odds score greater than or equal to 10 corresponding to a False Discovery Rate of ~5%. All SNPs that were called using Birdseed v2algorithm had a Quality Control (QC) call rate of >97% across individuals. All the subjects and members with SNPs that passed SNP QC procedures were entered into the CNV analysis.

Filters were set for ID call rates for the overall SNPs to identify IDs with poor quality DNA, if any. The CNV calls were generated using the Canary algorithm. In AGCS, contrast QC has to be >0.4 to be included in the CNV analyses. In this study, contrast QC observed was >2.5 across all samples showing a robust strength. To control for the possibility of spurious or artifact CNVs, we used the EIGENSTRAT approach of Price et al., (2006). This method derives the principal components of the correlations among gene variants and corrects for those correlations in the testing. By choosing as many components as the number of markers, we obtained greatly reduced effects and thereby obtained nothing from the correction tests. Therefore we selected the maximum number of markers and the number of samples less one (N-1) as the principle components. It consisted of obtaining the components themselves and their corresponding eigenvectors for N - 1 principal components, where N is the total number of samples in the dataset. We removed 55 individuals from the study group because they were extreme outliers on one or more significant EIGENSTRAT axes and further dropped 543 CNVs in the members selected for the study for not meeting the required QC measures. CNVs were considered validated when there was a reciprocal overlap of 50% or greater with the reference set. Though the Jaccard statistic is sensitive to the number of CNVs called by each algorithm (ideally each two algorithms would detect similar number of CNV calls), the relative values between the different comparisons of algorithms/platform/site are very informative. All the overlap analyses performed have handled losses and gains separately except when otherwise stated, and were conducted hierarchically. The calls from the algorithms that were called in both were not considered; instead, they were collated so that the relative values between the different comparisons of algorithms/platform/site are still very informative.

3. RESULTS

A total of 44109 CNVs investigated from 1715 individuals across 12 populations were identified. These CNVs contain 126190 genes consisting of 15185 singleton genes. Of these, 24,991 multigenes-CNVs (~56.65 %) contained 16759 (~13.28%) multigenes with a total of 902(~5%) singleton multigenes. These multigenes-CNVs were identified in 1703 (~99.30%) individuals) (Figure 1). Multigene-CNVs were found ranging from 51-62% across all populations with Hap Map Africa being the highest at 62% followed by Taiwan and Hap-Map CEU (Europe) at 56%; China-CHB and New World at 55% and Australia at 51%. Duplication CNVs containing multigenes was significantly higher (74.33%) compared to deletion (25.67%). However, Hap Map populations excluding YRI showed higher multigene deletion CNVs than duplication CNVs. The duplication multigene-CNVs were in the range of 38-78% with the highest in Ashkenazi Jews-II (78.97%) and the lowest in CHB (38.49%). The deletion multigene-CNVs were found highest and lowest in CHB (61.51%) and Ashkenazi Jews-II (21.03%) respectively.

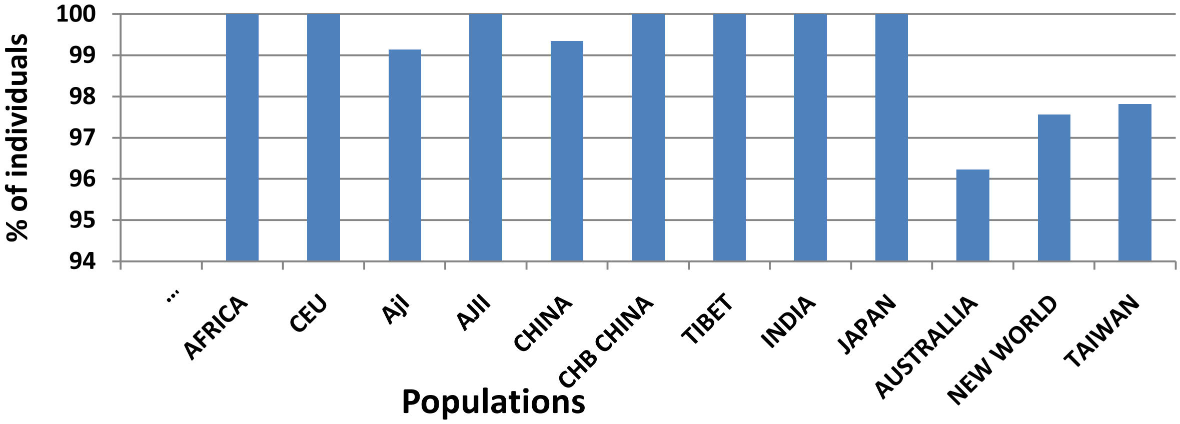

Figure 1: Number of individuals (in %) with multigene-CNVs across 12 populations.

Chromosome-wise distribution of multigene-CNV Count

Chromosome 4 showed the highest CNV burden on multigene across all populations followed by chromosomes 8, 14, 15, 17 and 22. Chromosomes 6, 13, 20, 21 and Y showed negligible presence of multigene-CNVs for all populations; population specific multigene-CNVs presence or absences were also observed. Chromosome wise multigene-CNV size distribution across 12 populations

Chromosomes 8, 14, 15 and 22 showed the highest CNV burden on multigene across all populations followed by chromosome 1, 4, 9, 10 and 17. Further, chromosomes 2, 5, 7, 8, 12, 16, 19 and X showed multigene-CNV presences for all populations (Figure 2). Chromosomes 3, 6, 11, 13, 18, 20, 21 and Y showed negligible presence of multigene-CNVs for all populations.

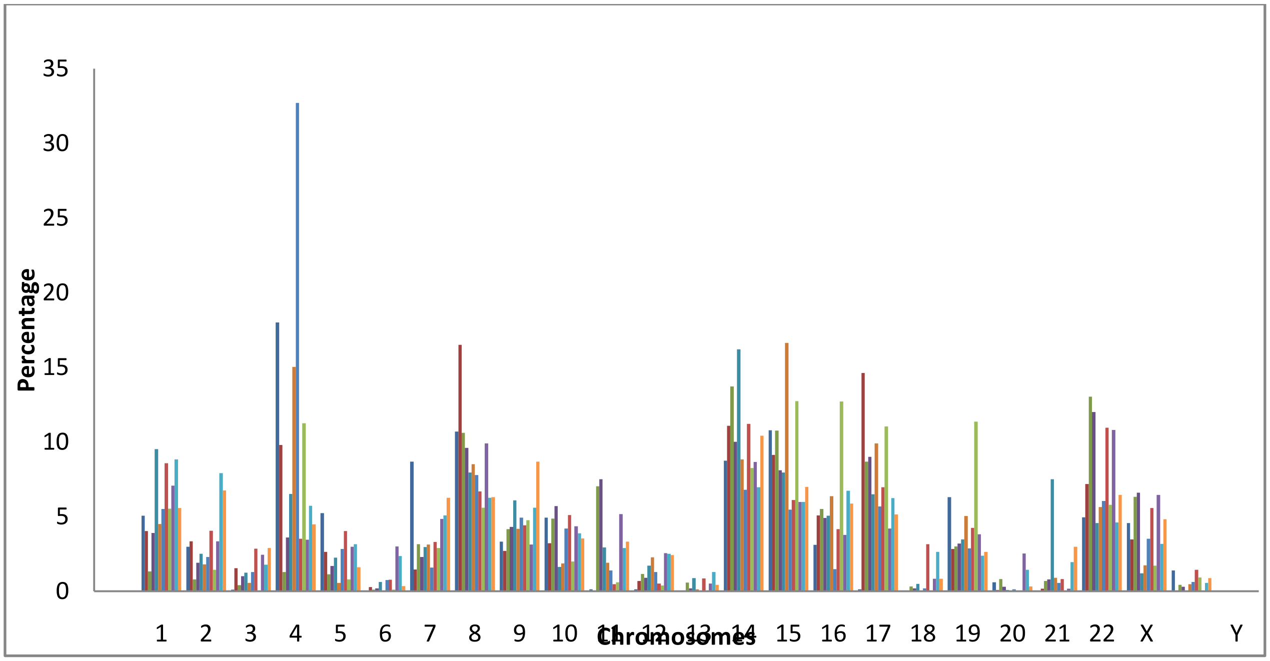

Figure 2: Chromosome wise Population wise size distribution of multigene-CNVs with each colour depicting individual population.

Multigene-CNV concentration in duplication and deletion regions

Multigenes were found over represented in duplication regions in two populations, and deletion regions in three populations viz., CEU, China and JPT. The highest gene content under duplication CNVs was observed for Australia and AJ II (Figure 3), followed by India, Taiwan Tibet and AJI. The highest gene content under deletion CNVs was observed in China (60.6%) and JPT (57%).

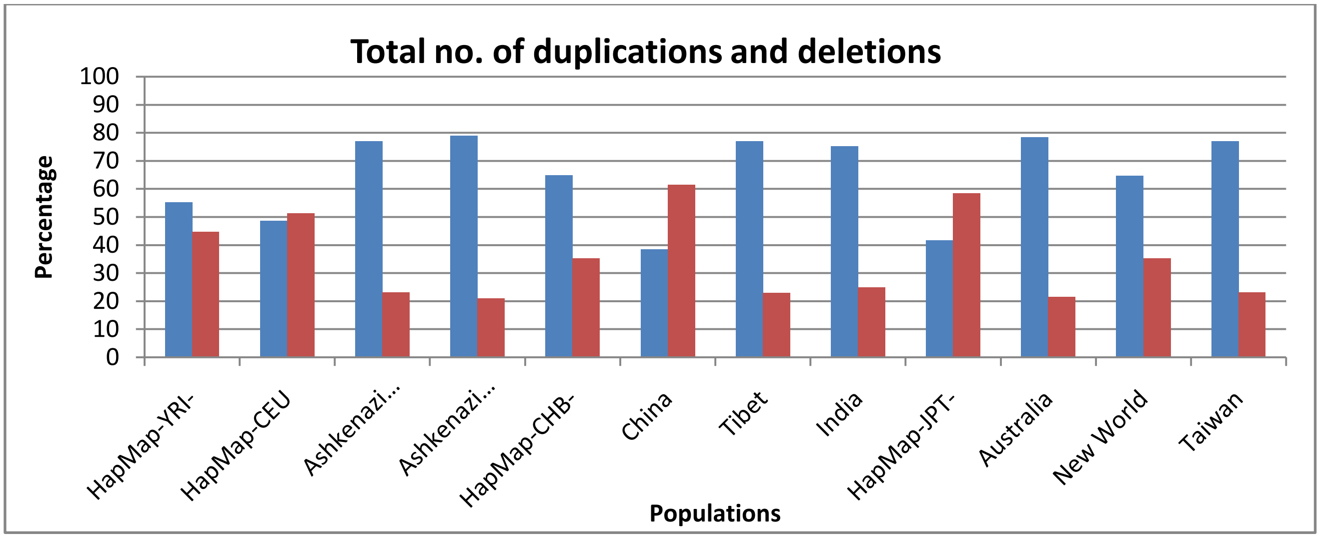

Figure 3: Total number of duplications and deletions in multigene-CNVs

Copy Number (CN) State of Multigene-CNVs

CN states for all duplication and deletion multigene-CNVs were assessed based on the 0, 1, 2, 3 and 4 states, where the numerical value represents its corresponding allele presence in the genome (Table 2). CHB showed the highest followed by CEU and YRI while Taiwan stood least in complete removal of both allelic segments (CN=0) in the genome for all multigenes. JPT showed the highest followed by CHB and CEU while AJ II showed least loss of one allele state in the multigenes bearing regions. CN state in the multi genes of sex chromosomes (CN=2) was observed high in AJ I followed by AJ II and New world and least in CEU. Nearly all populations showed a high number of CN=3

autosomal duplication multigene-CNVs ranging from 33-67%. The homozygous duplication (CN=4) state was seen across all populations distinctly, with highest in Taiwan followed by India and AJ I and least dual duplication CN state in CHB (Figure 4).

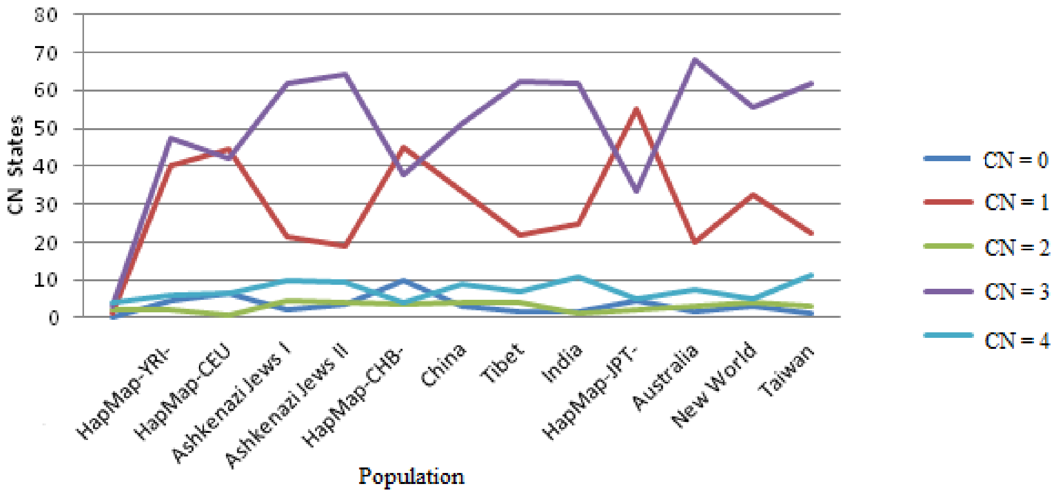

Figure 4: Distribution of CN sates across 12 populations

<u>Table 2: Total number of Copy Number (CN) States across 12 populations</u>

POPULATION	**CN=0**	**CN=1**	**CN=2**	**CN=3**	**CN=4**
Hap Map-YRI-Africa	**4.52**	**40.21**	**2.17**	**47.28**	**5.79**
Hap Map-CEU-Europe	**6.34**	**44.57**	**0.83**	**41.9**	**6.34**
A J- I	**2.03**	**21.56**	**4.68**	**61.65**	**10.06**
A J -II	**3.36**	**18.91**	**4.29**	**64.07**	**9.34**

Hap Map-CHB-China	9.62	45.18	3.34	37.65	4.18
China	2.91	33.2	3.9	51.23	8.73
Tibet	1.8	21.96	4.09	62.45	6.97
India	1.76	24.77	1.03	61.79	10.61
HapMap-JPT-Japan	4.34	55.07	2.17	33.33	5.07
Australia	1.53	20.08	2.99	67.86	7.52
New World	2.87	32.4	4.27	55.43	5.01
Taiwan	1.39	22.23	3.12	61.89	11.35

Sex bias in multigenes-CNVs

Multigene-CNV gene presences were observed to be biased in male and female genomes in several populations. Africa, CHB, CEU, AJ-II, China, Taiwan and Australia showed males (48-67%) to be carrying more multigenes under CNVs than females (≤40%), while, AJ-I, Tibet and India showed higher content in females (˜60%) compared to males (≤40%). The HapMap JPT showed balanced multigenes in both sexes, and New World population did not contain female samples, but show the largest multi-gene content in males, whereas,

Australian males showed the equal multigenes content under CNVs (Figure 5).

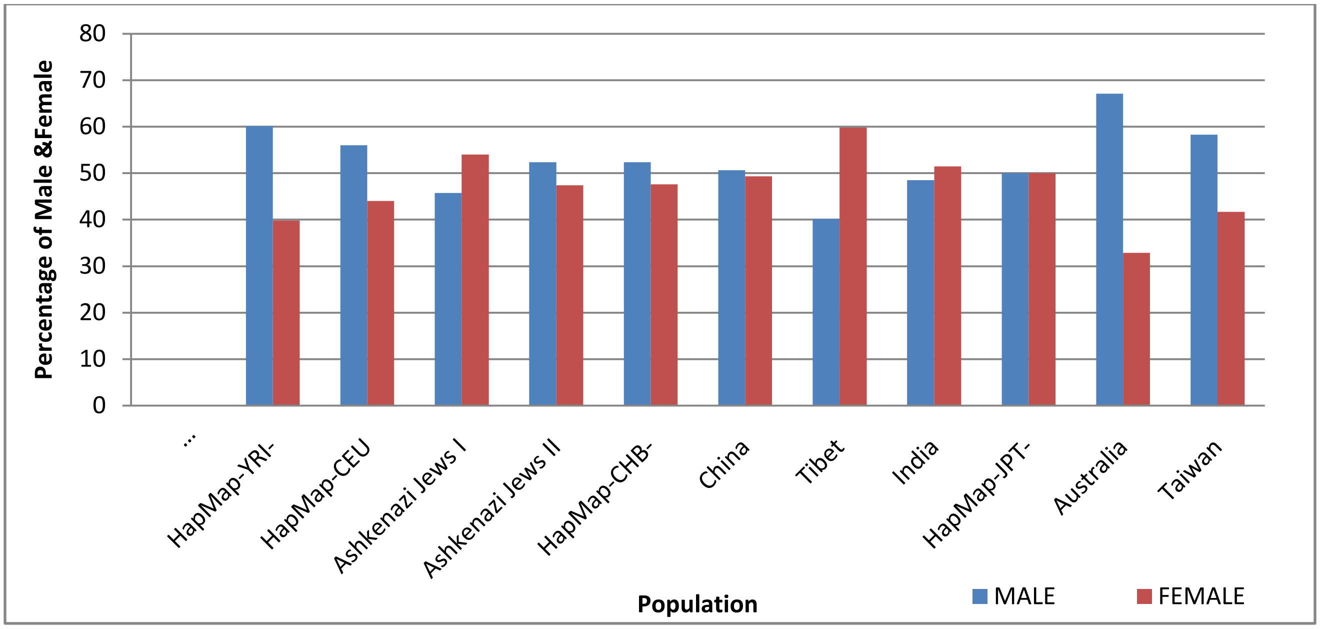

Figure 5: Total Number of multigene-CNV in males and females

Chromosome p and q arm wise multigene-CNVs size distribution

Multigene-CNVs were found distributed across both p and q arms based on CNV size in nearly all the chromosomes except a few. Deletions in p arm of chromosome CHB, Tibet, New world, Japan, India, CEU was found to be low. However 100% deletions was seen in 11p of Africa populations Highest duplications was seen on 16p arm. In other population of Hap-Map CEU no deletions were observed in 2p and 2q arm of the chromosomes, no duplications was seen in 3p and no deletions was seen on 3q arm. On the chromosome 4

there was no duplication nor deletion of the p arm. 100% duplications was observed on 7p, 12p and 20p arms, whereas 22q arm showed 100% deletions only.

AJ-1 showed no p arm duplication or deletion on chromosomes 13, 14, 15 and 22 while it was observed in other chromosomes. Chromosome 18 and 20 showed no deletions of either p or q arm. X chromosomes showed no p arm deletions. AJ-II showed no p arm deletions on chromosomes 10, 18 and 20, whereas no q arm deletion was seen on chromosomes 6, 16 and 21. Neither duplications nor deletions of p arm was been seen on Chromosomes 13, 14, 15 and 22 and only duplications was seen on Y chromosomes. p arm deletions was not observed on chromosomes 6, 10, 11; 21 and Y in China. Neither duplications nor deletions of p arm was seen on Chromosomes 13, 14, 15, 20 and 22 and only duplications was observed on p and q arm of chromosome 18.

HapMap-CHB showed 100 % duplications and deletions on Chromosomes 2 and 19. Further, no p arm deletions were observed on chromosomes 7 and 12. Neither duplications nor deletions of p arm was seen on chromosomes 4, 10, 11, 14, 15, 17, 21 and 22. Chromosomes 5 and 13 showed 100% deletions of multigene-CNVs on q arm whereas no q arm deletions were observed on chromosome 8 and 9. Chromosome 16 showed no q arm duplications. Chromosomes X and Y has no deletions of either p or q arm. Tibet showed no p arm deletions on chromosomes3, 5, 7, 10, 11, 12 and 21. Neither duplications nor deletions of p arm was seen on chromosomes 14 and 15. Chromosome 16 shows no deletion of q arm. No deletions of either p or q arm was seen on chromosomes 18 and only duplications of p and q arm was seen on chromosomes 20 and Y.

India showed 100 % duplications and deletions on chromosomes 1, 2, 4, 5 , 9, 17, 19 and X. No p arm deletions was observed on chromosomes 3, 4 and 7. No deletions of either p or q arm was seen on chromosomes 6, 11, 12 and 18 whereas chromosome 8, 16 and Y showed no deletion of q arm. Further, chromosome 10, 14, and 15 shows no deletion and duplication of p arm. 100 % duplication was seen on of q arm of chromosome 13. Chromosome 20 showed 100% duplication of p arm. Duplications of q arm were seen only on chromosomes 21 and 22. HapMap-JPT showed 100 % duplications and deletions on Chromosomes 1, 9 and 17. No p arm deletions was observed on chromosomes 4 and 5. No deletions of either p or q arm was seen on chromosomes 2. Chromosome 6 and 10 showed no deletion of q arm whereas chromosome 7, 14, 15 and 22 shows no deletion and duplication of q arm. 100% duplication was seen on of p arm of chromosome Y. No duplications of p and q arm were seen on chromosomes 19 and 16 respectively. p arm deletions were not observed on chromosomes 4, 6, 7, and 10 in Australian population while chromosomes 13 and 18 showed 100% of q duplications whereas chromosome Y was totally absent.

No p arm deletions were observed on chromosomes 5, 7, 11 and 20 in New world. Chromosomes 10, 13, 14, 15, 21 and 22 showed no p arm duplications neither deletions. Whereas Chromosome 16 showed no q arm deletions and chromosomes Y showed duplications only of p and q arm respectively. Taiwan showed no p arm deletions and duplications on chromosomes 13, 14 and 22 while chromosomes 18 showed p and q arm only duplications.

4. DISCUSSION

Identification of the multigene-CNVs across diverse populations helps understand the organization, distribution pattern of multigene-CNVs, evolutionary dynamics of the multigene sub-genome (Basten and Ohta, 1992) and account for differences in the expression of genes (Niimura and Nei, 2008). There have been only few multigene-CNV investigations on the CNVs obtained through bioinformatics (Waterhouse et al., 2009) identification from several databases and HapMap populations. However, there have not been many, which comparatively include populations across all continents to study notable variations on the genome, particularly on genes using different ethnic backgrounds. The current study represents the first drafts of population-specific multigene-CNVs map as well as a cross-population map. Here we present a comprehensive global multigene-CNV spectrum by identifying 902 singleton multigenes from total of 16759 multigenes from 24,991 multigene-CNVs across 12 populations using Affymetrix high resolution arrays. The CNVs identified in this study are highly consistent, because of the higher stringency adopted in both

the selection and validation of CNVs using multiple algorithms. Although these samples are well characterized, no medical information (except for Hap Map) was obtained. The multigenes and CNVs were co-active and conferred tremendous burden in the genome. Genetic diversity in humans affects both disease and normal phenotypic variation. Presence of CNVs alters the transcriptional and translational levels of overlapping or nearby genes by disrupting the coding structure or by altering gene dosage thereby conferring differential susceptibility to complex diseases (Pazin, 2013) and presence of multigene-CNV further adds to the complexities on the regulation of the genome. We detected gradual increase in the multigene-CNV counts from the Old World populations towards the New World populations indicating selective pressure of CNV occurrences in New World populations contributing to increased genetic diversity.

Multigene-CNVs were observed in all parts of the genome. CNVs across the human genome have been found to be associated with normal genetic heterogeneity as well as for a number of diseases and disorders. Previous studies have identified CNVs more in numbers. Multigene CNVs can considerably

alter their dosage, which would then affect the expression levels of several genes. We identified both multigene and non-multigene duplication CNVs to be higher than deletion CNVs in several populations, while only few populations showed increased deletions. The prevalence of higher multigenes-duplication CNVs might be due to the lesser damage it offers by protecting against deletions. Though, multigene duplication CNVs have been associated in many diseases, (Girirajan et al., 2012) most of them are regarded due to partial duplication disruptions of the multigene family. No consistent pattern was identified for the multigenes duplication CNV distribution across ethnicities since extremes of both multigene-CNV types were found elevated in ethnic groups residing within the continental boundaries.

Chromosome-wise distribution of Multigenes-CNVs

Chromosome 14 harbors high concentration of multigenes-CNVs followed by chromosomes 4, 8, 15 and 22. The reason behind such chromosomal selection for multigene-CNVs to frequent is largely speculative. The chromosomal region, 15q13-q14, including the α7-nicotinic acetylcholine receptor gene, CHRNA7, is a replicated region for schizophrenia (Sarah H. Stephens, Alexis Franks). These genes were found largely under CNVs across all populations, however, population specific multigene-CNV presences were also observed for India, Tibet and New World. Several multigene-CNVs have been implicated in pathways, phenotypes and diseases while various multigenes-CNVs are found to be either deleted or duplicated in diverse types of cancers (Jankowitz, and Lee, 2013) Gene signatures which are based on multigene profiling assays have been developed for the purpose to better define the prognosis and prediction of therapy results in early-stage breast cancer. The observed varied dominance of chromosomal multigene-CNVs may help in understanding the molecular basis of human phenotypic diversity. There were few instances without multigene-CNV presences in chromosomes, for some populations. Okihiro syndrome results from truncating mutations in the SALL4 locus on the chromosome 20q13.13-q13.2. Deletions of the whole SALL4 coding region as well as single exons deletions are also a common cause of Okihiro syndrome and indicate haplo insufficiency as the disease causing mechanism. The phenotypes caused by SALL4 deletions are not different from those caused by point mutations. No multigene deletion including SALL4 has been documented to date. Chromosomes 11 showed the mean burden, this can be due to the fact that people normally inherit one copy of chromosome 11 from each parent. For most genes on this chromosome, both copies of the gene are expressed, or "turned on," in cells. For some genes in the 11p15.5 region, however, only the copy inherited from a person's father (the paternal copy) is expressed. For other genes, only the copy inherited from a person's mother (the maternal copy) is expressed. These parent-specific differences in gene expression are caused by a phenomenon called genomic imprinting. Researchers have determined that changes in genomic imprinting disrupt the regulation of several genes located at 11p15.5, including CDKN1C, H19, IGF2 and KCNQ1OT1. Because these genes are involved in directing normal growth, problems with their regulation lead to overgrowth and the other characteristic features of Beckwith-Wiedemann syndrome. X chromosome showed multigene-CNV presences for all populations. The reason for this can be that, a number of multigenes have been identified that escape mammalian

X chromosome inactivation and are expressed from both active and inactive X chromosomes. The basis for escape from inactivation is unknown and, a priori, could be a result of local factors that act in a gene-specific manner or of chromosomal control elements that act regionally. Models invoking the latter predict that such genes should be clustered in specific domains on the X chromosome, rather than distributed at random along the length of the X, hereby defining a unique multigene domain on the proximal short arm that is transcriptionally active on the inactive X chromosome (Willard .and Miller, 1998) and chromosome Y showed negligible presence of multigene-CNVs for all populations, because of gene conversion that if not proper can lead to male infertility (Vollrath et al., 1992)

Higher duplication events were observed than deletions in several studies (Todd and Vodkin, 1996); the unusual nature of the locus suggests that its dominant alleles may represent naturally occurring examples of homology-dependent gene silencing and that the spontaneous deletions erase the gene-silencing phenomena. Similarly the genes under these CNVs also tend to be overrepresented in duplication CNVs compared to deletion CNVs (Köster et al.,1988). Australia and Ashkenazi Jews, followed by India, Taiwan and Tibet showed higher gene content in multigene duplication CNV events, while gene overrepresentation in deletion CNVs were limited to only China, CHB and Japan. These indicate the genetic relationships and diversity between and within populations. The proportion of human genetic variation due to differences between populations is modest, and individuals from different populations may be genetically more similar than individuals from the same population. Yet sufficient genetic data can permit accurate classification of individuals into populations. This causes genetic clusters to correlate statistically with population groups when a number of alleles are evaluated. Comparing human populations taken from different continents meaning that only between 10 and 15 % of genetic differences between individuals are attributable to their geographic origins. This difference is relatively small compared to many other large mammal species spread among different continents Different clines align around the different centers, resulting in more complex variations than those observed comparing continental groups and thus propose that multigene-CNVs are inherent variants and should be considered as probable candidates while performing genotype-phenotype association studies.

Multigene-CNV size and multigene-CNV count distribution deviated significantly across population and revealed contrasting contributions towards the burden of CNVs (Martin et al., 1998) Though we found duplication multigene-CNVs to be overrepresented than deletion CNVs, however, further analysis on multigene-CNV size, revealed larger size deletions to be encompassing multigenes compared to duplication CNVs while only few populations were found with high duplication sized multigene-CNVs. This probably indicates that, CNV type, count and size factors are independent of each other and should be collectively analyzed to understand the complexities of multigene in populations.

Ashkenazi Jews, JPT and Taiwan contained the highest number of singleton multigenes under CNVs, than the remaining populations. The inconsistency in total multigenes distribution against singleton multigene counts across chromosomes was observed due to the regular frequenting of CNVs in regions bearing one gene across many individuals in all populations and was not due to the presence of multiple multigenes and this pattern was observed across multiple chromosomes within the populations. These singleton genes can be used for genome wide association studies and can also be used as disease specific Tags.

Sex bias

Gender plays a pivotal role in the human genetic identity and is also manifested in many genetic disorders. Multigene-CNV presence showed sex bias in several populations, largely dependent on ethnicity. There was a significant difference in multigenes distribution under CNVs in both male and female genome. These findings imply gender based phenotypic differences. Many human genetic phenotypes, including those related to olfaction, developmental delay and intellectual disabilities have shown similar sex bias along with recombination rates (Shadravan, 2013; Girirajan et al., 2011, 2012).

As males and females share highly similar genomes, the regulation of many sexually dimorphic traits is constrained to occur through sex-biased gene regulation. There is strong evidence that human males and females differ in terms of growth and development in utero and that these divergent growth strategies appear to place males at increased risk when in sub-optimal conditions. Since the placenta is the interface of maternal-fetal exchange throughout pregnancy, these developmental differences are most likely orchestrated by differential placental function. To date, progress in this field has been hampered by a lack of genome-wide information on sex differences in gene expression.

It was necessary to evaluate the degree of similarity of all the identified multigene-CNVs to accurately assess the genome changes in order to correlate with duplication and deletion events. HapMap and China showed significant gains in the multigenes sub-genome with decreased 0 and 1 deletion CN states indicating homozygous gain of parts of multigenes sub-genome, while remaining populations showed lower duplication. Australia showed highest male populations than other populations and exhibited burden of multigenes-CNVs on

sex chromosomes compared to any other population, which indicates the phenotypic differences of the Australian population can largely be attributed to the multigenes present on sex chromosomes since they all exhibited differences even with shared ethnicities. Comparison of the 5 CN states of the non-multigenes CNVs with the CN states of multigenes-CNVs revealed striking similarities indicating that the multigenes sub-genome is under the force of CNV dynamics of the genomes.

Based on the inheritance status, YRI showed a fold higher inheritance rate compared to India, whereas New World did not show any female inheritance of multigene-CNVs. The rate of de novo CNV occurrences were similar, while "Unknown" cluster showed unequal rates with Australia being the highest, in males followed by YRI, Taiwan and India. Though, earlier studies have reported diverse CNV transmission and de novo event rates in probands with several neurodevelopment phenotypes and monozygotic twin studies (Ehli et al., 2012). Multigenes -CNV frequency bias was observed on the CNV transmissions from maternal genome only, showing major contributions from deletion CNVs than

duplications. This could be due to existence of such biased transmissions that greatly signifies the role of parental transmission in gene regulation and counters some distinct phenotypic features observed in previous studies [van den Ouweland] and in certain gender-specific disease manifestations.

Each chromosome is divided into two sections (arms) based on the location of a narrowing (constriction) called the centromere. By convention, the shorter arm is called p, and the longer arm is called q. The chromosome arm is the second part of the gene's address. The position of the gene on the p or q arm is based on a distinctive pattern of light and dark bands that appear when the chromosome is stained in a certain way. The position is usually designated by two digits, we found CNVs close to telomeres and centromere regions. Multigene-CNVs in telomere region were found in both p and q arms and did not show any definite pattern of distribution in the populations and also no single factor was found that might explain this pattern of CNV distribution but were significantly overrepresented in number. Though, earlier studies have reported contradictory findings on the order of representation of CNVs in telomere regions, however,

the present study indicates slight inclination towards the higher representation (Sharp et. al., 2005; Nguyen et. al., 2006). Population specific dominance across p and q arms of chromosomes were also observed and showed the highest across all populations. AJ-1 peaked in the duplication CNVs of 20q region. Higher duplications were also observed in 14q of India.

5. SUMMARY

1. The Human Genome Project (HGP) was an international collaborative research program whose main goal was the complete mapping and sequencing of the whole genes present in human beings.

2. HGP researchers had three major aims: to determine the order of the sequences of all bases in our genome, making the maps which can locate the genes on our chromosomes and making of the Linkage Maps through the help of which inherited traits can be tracked over generations.

3. The HGP has revealed that there are about 20,500 human genes which are 99.99% accurate but the sequence is not completely complete and can neither be.

4. The HGP has laid the foundation for many various different projects and also paved way for findings of the different types of variations found in humans.

5. Of all the variations found in human the most important one is the "Copy Number Variations" or CNVs.

6. CNVs indicate the presence of additional or absence of segments of

chromosomes in individuals that found to be abundant in humans and are found to be associated with some diseases. CNVs account for up to 5% of the total human genome.

7. The CNVs location in our genome is of utter most importance because an additional copy of the gene might result in additional expression if the gene found to be actively functional and it might create an entirely new Chimeric protein. A partially expressed second copy of the gene might stop the expression of the first copy by inserting within it. Hence, many efforts have been made to find co relation between copy number variations and multigenes and its association with different types of diseases.

8. A growing body of evidence suggests that CNVs or structural variations across the genome is common and likely contributes to human diseases.

9. Even in studies that have shown negative linkage, the possible contribution of undetected CNVs cannot be dismissed.

10. Multigenes are the group of genes that are derived from the same organisms and encodes proteins with same sequences. These multigenes are

the results of gene duplications that produces genes in pairs. If both the copies of the genes are maintained in the next generation also then it produces a "Multigene Family"

11. There are many multigene families in humans such as Globin Genes; Histocompatibility antigens, Actins, Tubulins etc.

12. Understanding the roles of CNVs in multigenes could help to diagnose and treat many diseases more effectively and rapidly than is currently possible.

13. Genetic linkage analysis has identified regions of the genomes that are inherited and might be related in causing different diseases.

14. A whole genome scan is a very powerful tool for identification of any type of changes in an individual with diseases. Genome Wide scan can be performed using several tools at varying resolutions. The advancement in microarray well bead based chip technologies led to the development of chips containing millions of SNPs and CNP probes. Comparatively these advanced DNA Chips offer a high resolution at base pair level intermarker distances compared to earlier Mega Base resolutions.

15. A total of 44109 CNVs investigated from 1715 individuals across 12 populations were identified. These CNVs contain 126190 genes consisting of 15185 singleton genes. Of these, ~56.65% multigenes-CNVs contained ~13.28% multigenes with a total of 902 (~5%) singleton multigenes. These multigenes-CNVs were identified in 1703 (99.30%) individuals.

16. This is a maiden report in the analysis of copy number variations in multigene families identified in 12 populations across the globe.

17. The findings of CNVs in multigenes contribute to the array diseased phenotypes seen in the families which demonstrate the importance of performing a high resolution assessment of genomic background even after the detection of rare and likely damaging CNVs.

18. An understanding the biology of the complex cognition is major challenge to which genetics can provide crucial clues.

19. The methodology adopted here has a higher genome resolution but the number of the populations chosen for this study is less which becomes a limitation and requires further whole genome study on large extended families

to identify more appropriate genes and relevant tests. These will at large help to rehabilitate, treat, control, manage and cure many diseases.

6. REFERENCES

1. Affymetrix, Inc. (2008) User manual: Genotyping Console Software 2.1.

2. Affymetrix, Inc. Data Sheet: Genome Wide Human SNP Array 6.0 (2009).

3. Anderson S, Bankier AT, Barrell BG,et.al. (1981) Sequence and organization of the human mitochondrial genome, Nature. 1981290(5806):457-65

4. Armengol L, Villatoro S, González JR et al. Identification of copy number variants defining genomic differences among major human groups. PLoS One. 2009; 30; 4(9):e7230.

5. Basten CJ and Ohta T. (1992) Simulation Study of a Multigene Family, with Special Reference to the Evolution of Compensatory Advantageous Mutations, Genetics. 247-252.

6. Beckman JS and Estivilli X. (2007) Copy number Variants and genetic traits: closer to the resolution f phenotypic to genotypic variability. Nat. Rev. Genet. 8, 639-46.

7. Bentley DR. (2000) The Human Genome Project- an overview. Med Res rev 20, 189-86.

8. Benyamin B and Visscher PM. (2009) Family based genome wide association studies. 170-72.

9. Bochulova E.G., Huang N. et al. (2010) Large, rare chromosomal deletions associated with severe early- onset obesity. Nature 463, 666-670.

10. Bourgain C, Genin et al. (2007) Are genome-wide association studies that we need to dissect thebgenetic component of complex human diseases? Eur.j.Hum. Genet. 15,260-3

11. Bozeman MT: Golden Helix, Inc. SNP & Variation Suite (Version 7.x) [Software]. Available: http://www.goldenhelix.com (Accessed 2013 Jan 13).

12. Butler JL, Osborne Locke ME, Hill KA et al. (2012) HD-CNV: Hotspot Detector for Copy Number Variants. Bioinformatics. 15: 262-263.

13. Charlesworth B. (1991) The evolution of Sex chromosomes. Science 4997, 1030-1034.

14. Clamp M, Fry B, Kamal M et al. (2007) Distuishing protein- coding and non coding genes in the human genome. Proc. Natl. Acad. Sci.104, 19428-19433.

15. Claverie JM (2001) Gene Number. What if there are only 30,000 human genes? Science 291, 1255-1257.

16. Conrad DF, Pinto D et al. (2010) Origins and functional impact of copy number variations in Human Genome. Nature 464, 704-702.

17. Database of Genomic Variations website. Available: http://dgv.tcag.ca/dgv/app/home (Accessed 2013 Jan 13).

18. Ehli EA, Abdellaoui A, Hu Y et al. (2012) De novo and inherited CNVs in MZ twin pairs selected for discordance and concordance on Attention Problems. Eur J Hum Genet 20: 1037-43.

19. Feuk L, Carson AR and Scherer SW (2006) Structural variations in the human genome, Nature Reviews Genetics 7, 85-97. Doi:10:1038/nrg 1767.

20. Gautam P, Jha P, Kumar D et al. (2012) Spectrum of large copy number variations in 26 diverse Indian populations: potential involvement in phenotypic diversity. Hum Genet.;131: 131-43.

21. Girirajan S, Brkanac Z, Coe BP et al. (2011) Relative Burden of Large CNVs on a Range of Neuro-developmental Phenotypes. PLoS Genet 7 (11): e 1002334 doi:10:1371/ journal. pgen. 1002334.

22. Girirajan S, Eichler EE. (2010) Phenotypic variability and genetic susceptibility to genomic disorders. Hum Mol Genet.;15;19(R2):R176-87.

23. Girirajan S, Rosenfeld JA, Coe BP et al. (2012) Phenotypic heterogeneity of genomic disorders and rare copy-number variants. N Engl J Med.; 367(14):1321-31.

24. Girrirajan S, Campbell CD et al. (2011) Human copy number variation and complex genetic disease, Annu Rev Genet 45, 203-26.

25. Hollox EJ, Huffmeier U, Zeeuwen PL et al. (2007) Diet and the evolution of human amylase gene copy number variation. Nat. Genet. 39: 1256-1260.

26. Iafrate AJ, Feuk L, Rivera MN, et. al. (2004) Detection of large scale of variations in the human genome. Nature Genetics 36, 949-95.

27. International Hap-Map Consortium 2001. A physical map of the human genome. Nature 409, 934-941.

28. Itsara A, Wu H, Smith JD, et.al. (2010) De Novo rates and selection of large copy number variations Genomes Res. 20: 1469-1481.

29. Jankowitz CR and Lee AV. (2013) The Evolving Role of Multi-Gene Tests in Breast Cancer Management. Oncology (Williston Park) 27(3):210, 212, 214.

30. Jobling MA, Lo LC et al. (2007) Structural Variations on the short arm of the human Y chromosome: recurrent multigenes deletions encompassing Amelogenin Y. Hum. Mol. Genet. 16, 307-316.

31. Kaback DB et al. (1999) Chromosome size-dependent control of meiotic reciprocal recombination in Saccharomyces cerevisiae: the role of crossover interference.

Genetics. 152 (4):1475-86.

32. Kehrer-Sawatzki H. (2007) What is the difference between copy number variations makes. Bioassay 29 311-313.

33. Ken L, Majid F. (2001) Myotonic dystrophy–a multigene disorder, Brain Res Bull. 2001 Oct-Nov 1;56(3-4):389-395

34. Kidd JM, Cooper GM et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453, 56-64.

35. Kim HL, Iwase M, Igawa T et al. (2012) Genomic Structure and Evolution of Multigene Families: "Flowers" on the Human Genome. Int J Evol Bio; doi:10.1155/2012/917678.

36. Kirov G. (2010). The role of copy number Variation in schizophrenia. Expert Rev Neuro 10, 25-32.

37. Köster M, Pieler T, Pöting A et al. (1988) The finger motif defines a multigene family represented in the maternal mRNA of Xenopus laevis oocytes, EMBO J. 1988 Jun;7(6):1735-

38. Krzywinski M, Schein J, Birol I et al. (2009) Circos: an Information Aesthetic for Comparative Genomics. Genome Res. 19: 1639-1645.

39. Lee C, Iafrate AJ, Brothman AR. (2007) Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. Nature Genetics. 39: S48-S54.

40. Lekkerkerker CG and Boland JC. (1962) Representation of a finite graph by a set of intervals on the real line. Fund. Math. 51:45-64

41. Li H, Durbin R. (2011) Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.

42. Lin CH, Li LH, Ho SF et al. (2008) A large-scale survey of genetic copy number variations among Han Chinese residing in Taiwan. BMC Genet. 9:92.

43. Lin CH, Li LH, Ho SF et al. (2012) A large-scale survey of genetic copy number variations among Ha Kim HL, Iwase M, Igawa T, et al. Genomic Structure and Evolution of Multigene Families: “Flowers” on the Human Genome. Int J Evol Bio; doi:10.1155/2012/917678.

44. Liu X, Cheng R, Ye X et al. (2013) Increased rate of sporadic and recurrent rare genic copy number variants in Parkinson's disease among Ashkenazi Jews. Mol Genet Genomic Med.; doi: 10.1002/mgg3.18

45. Lou H, Li S, Yang Y, et al. (2011) A Map of Copy Number Variations in Chinese Populations. PLoS ONE 6, (11).

46. Maeda N and Smithies O. (1986) The Evolution of Multigene Families: Human Haptoglobin Genes, DOI: 10.1146/annurev.ge.20.120186.000501

47. Martin F, Stefan S, Daniel D, et al. (1998) Controlled proliferation by multigene metabolic engineering enhances the productivity of Chinese hamster ovary cells, Nat Biotechnol. 1998 May;16(5):468-72.

48. McElroy JP, Nelson MR, Caillier SJ et al. (2009) Copy number variation in African Americans. BMC Genetics;10:15.

49. Nguyen DQ, Webber C, Ponting CP. (2006) Bias of selection on human copy-number variants. PLoS Genet.; 2(2):e20. doi:10.1371/journal.pgen.0020020

50. Niimura Y. (2012) Olfactory Receptor Multigene Family in Vertebrates: From the Viewpoint of Evolutionary Genomics, Curr Genomics.; 13(2): 103–114,

51. Ohno S. (1967) Sex chromosome and Sex linked Genes. Springer Books, Berlin.

52. Pazin M. (2013) Interpreting Variation in Human Non-Coding Genomic Regions Using Computational Approaches with Experimental Support. 228-239.

53. Peng B, Amos CI et. al. (2007). Forward- time stimulations of human populations with complex diseases. PLoS Genet. Mar 23;3(3):e47, 407-420.

54. Price AL, Patterson NJ, Plenge RMet al. (2006) Principal components analysis

corrects for stratification in genome-wide association studies. Nat Genet. 38 (8): 904-909.

55. Redon R, Ishikawa S, Fitch KR et al. (2006) Global variation in copy number in the human genome. Nature; 444 (7118):444-454.

56. Rozen S, Skaletsky H, Marszalek JD, et al. (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature 423, 873-876.

57. Sebat J, Lakshmi B, Malhotra D et al. (2007) Strong association of de novo copy number mutations with autism. Science 316: 445-449.

58. Sebat J, Lakshmi B, Troge J, et al. (2004) Large-scale copy number polymorphism in the human genome. Science 305: 525-528

59. Shadravan F. (2013) Sex bias in copy number variation of olfactory receptor gene family depends on ethnicity. Front Genet. 4:32. doi: 10.3389/fgene.2013.

60. Sharp AJ, , Locke DP, McGrath SD et al. (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77: 78-88.

61. Simonson TS, Yang Y, Huff CDet al. (2010) Genetic evidence for high-altitude adaptation in Tibet. Science 329, 72-75.

62. Stephens SH, Franks A and Leonard S. (2012) Multiple genes in the 15q13-q14 chromosomal region are associated with schizophrenia, Psychiatr Genet. 2012 doi:

10.1097

63. Takahata N. (1993) Allelic genealogy and human evolution. Mol Biol Evol 10: 2-22.

64. The 1000 Genomes Project Consortium, Abecasis GR, Altshuler D et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-73.

65. The ENCODE Project Consortium. (2011) A users guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 9, e1001046.

66. The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74.

67. The International Hap Map Consortium (2003) The International Hap Map Project. Nature 426: 789-796.

68. The International Hap-Map Consortium. (2005) A Haplotype Map of the Human Genome. Nature 437,.1299-1320..

69. The International Hap-Map Consortium. (2010) Integrating common and rare genetic Variations in diverse human populations. Nature 467, 50-58.

70. Todd JJ and Vodkin LO. (1996) Duplications That Suppress and Deletions That Restore Expression from a Chalcone Synthase Multigene Family, Plant Cell, 1996

Apr;8(4):687-699.

71. Tomoko O. (1980) Evolution and Variation of Multigene Families. Lecture Notes in Biomathematics. 639-46.

72. Tomoko O. (2008) Gene Families: Multigene Families and Super families, DOI: 10.1002/9780470015902.a0005126

73. van den Ouweland JM, Lemkes HH, Trembath RC et al. (1994) Maternally inherited diabetes and deafness is a distinct subtype of diabetes and associates with a single point mutation in the mitochondrial tRNA(Leu(UUR)) gene. Diabetes 43(6): 746-51.

74. Veerappa AM, Vishweswaraiah S, Lingaiah K, et al. (2013) Unraveling the Complexity of Human Olfactory Receptor Repertoire by Copy Number Analysis across Population Using High Resolution Arrays. PLoS ONE; 8(7): e66843.

75. Vollrath D, Foote S, Hilton A, et al. (1992) The human Y chromosome: a 43-interval map based on naturally occurring deletions, Science. 1992 Oct 2; (5079):52-9.

76. Waterhouse AM, Procter JB, Martin DM, et al. (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 2009 May 1;25(9):1189-91

77. Wei Ning C. (2014) Enhanced Genetic Tools for Engineering Multigene Traits into Green Algae, PLoS One. 2014 Apr 7;9(4):e94028.

78. White Paper: (2008) Affymetrix Canary Algorithm Version 1.0., 1-7.

79. Willard HF and Miller AP. (1998) Chromosomal basis of X chromosome inactivation: identification of a multigene domain in Xp11.21-p11.22 that escapes X inactivation. Proc Natl Acad Sci U S A. 1998 Jul 21;95(15):8709-14.

80. Wu J, Grzeda KR, Stewart C, et al. (2012) Copy Number Variation detection from 1000 Genomes Project exons capture sequencing data. BMC Bioinformatics. 17;13:305. doi: 10.1186/1471-2105-13-305.

81. Yang Y, Chung EK, Wu YL, et al. (2007) Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): Low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am. J. Hum. Genet.;80:1037-1054.

82. Young JM, Endicott RM, Parghi SS, et al. (2008) Extensive copy-number variation of the human olfactory receptor gene family. Am J Hum Genet.; 83(2):228-242.

83. Zhang F, Gu W, Hurles ME, et al. (2009) Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet.;10:451-81

84. Zhang Y-B, Li X, Zhang F, et al. (2012) A Preliminary Study of Copy Number Variation in Tibetans. PLoS ONE; 7:7.