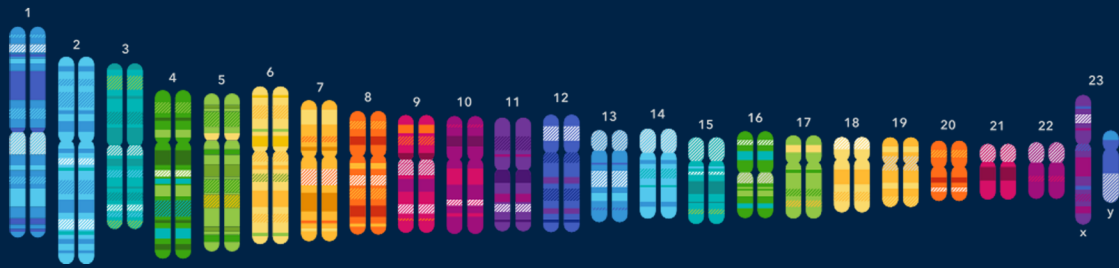
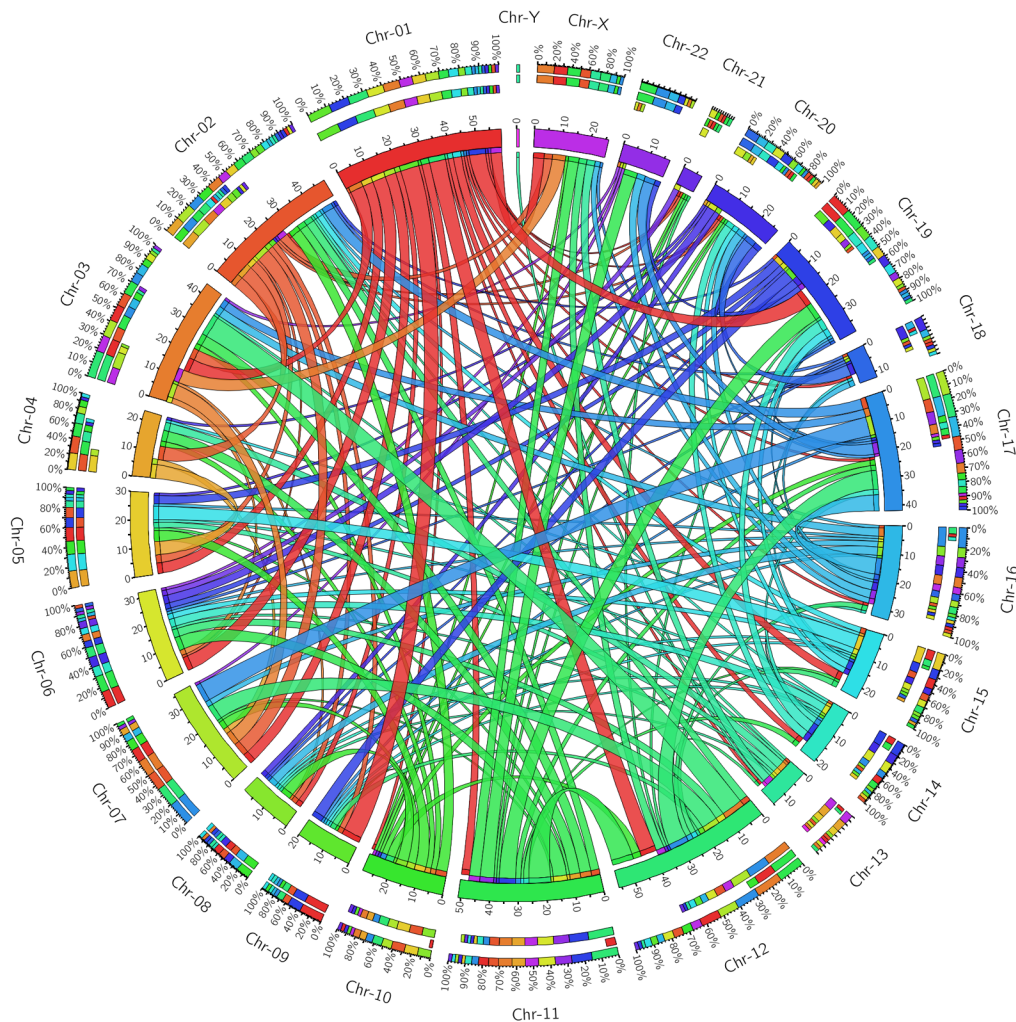


The Importance of Human Duplicated Genes: Insights from Evolutionary Perspective



Thesis submitted for the degree of Doctor of Philosophy (Sc.)
in Biophysics, Molecular Biology and Bioinformatics



By

Debarun Acharya

Bioinformatics Centre, Bose Institute

2018

The Importance of Human Duplicated Genes: Insights from Evolutionary Perspective

**Thesis submitted for the degree of
Doctor of Philosophy (Sc.)
in
Biophysics, Molecular Biology and Bioinformatics**

**By
Debarun Acharya**

**Department of Biophysics, Molecular Biology and Bioinformatics
University of Calcutta**

2018

Dedicated to
My Mother
Mrs. Aradhana Acharya
&
My Wife
Mrs. Aditi Dutta Acharya

Acknowledgement

This PhD thesis represents the beginning of my journey in the vast scientific world of research. With lots of ups and downs, success and failure, acceptance and rejection, agreement and debate, the wait is finally over. Now, it is the time to thank the persons who has played an important role that helped me to achieve my accomplishments, both in academic and nonacademic ways.

First and foremost, I would like to thank my PhD mentor Prof. Tapash Chandra Ghosh for his continuous support, motivation and assistance throughout my PhD research career. He is undoubtedly one of the most kind-hearted person I have ever seen in my life. His faith in me and my abilities has kept me constantly engaged in my research. His vast experience has always made an easy solution for even most complicated problems. His guidance in every academic and non-academic aspect made me hopeful. I feel to be fortunate enough to have a PhD mentor like him.

I would like to show my sincere gratitude to Prof. Pinak Chakraborty, Scientist-in-charge, Bioinformatics Centre (BIC), Bose Institute for his assistance in every aspect during my research work. I would also like to thank Dr. Shubhra Ghosh Dastider, Dr. Zhumur Ghosh and Dr. Sudipto Saha for their kind suggestions whenever I needed an assistance. I am really thankful to all the non-teaching staff of BIC including Sanjib da (Mr. Sanjib Gupta), Tushar da (Mr. Tushar Kanti Bhattacharya), Jiba da (Mr. Jibanananda Mondal) and Sujata di (Mrs. Sujata Roy) for their kind co-operation in any kind of official and non-official work.

I convey my special acknowledgements to my seniors labmates, most of them are now the alumni of our lab, and are the reason behind my critical understanding of the subject and developing my research aptitude. My sincere thank goes to Soumita di (Dr. Soumita Podder), who has helped in every time I have faced trouble in my PhD tenure. Sandip da (Dr. Sandip Chakraborty) was always ready for the research-oriented discussions, even on the go. Arup da (Dr. Arup Panda) was the most sincere and perfectionist and always eager to share his knowledge, Jyotirmoy da (Dr. Jyotirmoy Das) has helped me a lot in all the technical aspects of my work, Tina di (Dr. Tina Begum), the most hard-working lady of our lab, Subarna di (Dr. Subarna Thakur), always ready for good food and scientific fight, Dola di (Mrs. Dola Mukherjee), always tensed, Kamalika di (Dr. Kamalika Sen), the most hygienic person in the lab and Deeya di (Dr. Deeya Saha), the calm and composed lady of our lab has helped me in each and every critical discussion beneficial for the improvement of my research works.

I would also like to acknowledge my junior labmates. Kakali (Mrs. Kakali Biswas Dasgupta), the window shopper and Manish (Mr. Manish Prakash Victor), the Engineer-cum-Biologist of the lab for their valuable help and support to work together for hours tirelessly.

I owe my sincere gratitude to my Parents for supporting me in every aspect of my life with their constant mental and financial support. I also thank my Father-in-law and Mother-in-law for their constant support.

Finally, I am fortunate enough to have my best friend as my life-partner. Aditi (Mrs. Aditi Dutta Acharya) was and is always there for me with her continuous support and encouragements. She is the only person who knows me better than myself. No word is enough to thank her unconditional love and efforts.

Abbreviations

Terminologies:

Mya– Million years ago

DNA- Deoxyribonucleic Acid

RNA– Ribonucleic Acid

mRNA– Messenger RNA

tRNA– Transfer RNA

rRNA– Ribosomal RNA

Pu- Purines

Py– Pyrimidines

dN– Nonsynonymous nucleotide substitution per nonsynonymous sites

dS- Synonymous nucleotide substitution per synonymous sites

PPI– Protein–Protein Interaction

P_E- Proportion of Essential genes

WGD- Whole–Genome Duplication/ Whole–Genome Duplicates

SSD- Small–Scale Duplication/ Small–Scale Duplicates

Databases:

GO- Gene Ontology

OGEE- Online Gene Essentiality Database

HGMD- Human Gene Mutation Database

HPA- Human Protein Atlas

Contents

Chapter 1: Introduction.....	1-22
<i>Evolution: the process behind all life forms in the earth.....</i>	<i>1</i>
<i>1.1. Origin of life on Earth.....</i>	<i>2</i>
<i>1.1.1. The ancient environment of the Earth.....</i>	<i>2</i>
<i>1.1.2. Formation of biomolecules essential for life</i>	<i>2</i>
<i>1.2. Cells and the Genetic material.....</i>	<i>4</i>
<i>1.2.1. Cells and Cellular organelle</i>	<i>4</i>
<i>1.2.2. The Structure of DNA.....</i>	<i>5</i>
<i>1.3. From Genotype to Phenotype</i>	<i>7</i>
<i>1.3.1. Genes and Genomes.....</i>	<i>7</i>
<i>1.3.2. Alleles, phenotypes and genotypes</i>	<i>8</i>
<i>1.3.3. From Genes to Proteins- The Central Dogma of Molecular Biology</i>	<i>9</i>
<i>1.4. Mutations.....</i>	<i>11</i>
<i>1.5. Gene Duplication.....</i>	<i>12</i>
<i>1.5.1. The Contribution of gene duplication in Evolution.....</i>	<i>16</i>

<i>1.6. Origin of the proposal.....</i>	<i>18</i>
<i>1.7. Thesis Organization</i>	<i>21</i>
Chapter 2: Resources and Methodology.....	23-32
<i>2.1. Gene Essentiality</i>	<i>24</i>
<i>2.2. Homologous genes- Orthologs, Paralog and Ohnologs</i>	<i>24</i>
<i>2.3. Protein Evolutionary Rate</i>	<i>26</i>
<i>2.4. Pair wise comparison of gene functions.....</i>	<i>26</i>
<i>2.5. Gene expression profile similarity</i>	<i>28</i>
<i>2.6. Micro-RNA target sites</i>	<i>29</i>
<i>2.7. Protein Multifunctionality.....</i>	<i>30</i>
<i>2.8. Disease Genes</i>	<i>31</i>
<i>2.9. Developmental Genes.....</i>	<i>31</i>
Chapter 3: The complex association of gene duplication and gene essentiality: insights from human and mouse genome.....	33-48
<i>3.1. Introduction.....</i>	<i>34</i>
<i>3.2. Materials and Methods.....</i>	<i>36</i>
<i>3.2.1. Gene Essentiality and Gene Duplication</i>	<i>36</i>
<i>3.2.2. Developmental Genes</i>	<i>36</i>
<i>3.2.3. Phyletic Age and Overall Proportion of Essentiality.....</i>	<i>37</i>
<i>3.2.4. Pseudogenization.....</i>	<i>38</i>
<i>3.2.5. Functional Distance.....</i>	<i>38</i>
<i>3.2.6. Micro-RNA Target Sites</i>	<i>39</i>

3.2.7. Evolutionary Rate	39
3.2.8. Statistical Analyses.....	40
3.3. Results and Discussions	40
3.3.1. Gene essentiality and gene duplication in human and mouse.....	40
3.3.2. The proportions of paralog pseudogenization in human and mouse essential genes.....	43
3.3.3. Functional distance of human and mouse essential genes.....	44
3.3.4. Mean micro-RNA target sites in the paralogs of human and mouse essential genes.....	45
3.3.5. The evolutionary rates of paralogs of human and mouse essential genes.....	47
3.4. Conclusion	47
Chapter 4: The importance of whole-genome duplication in human genome evolution.....	49-81
4.1. Introduction.....	50
4.2. Materials and Methods.....	54
4.2.1. Classification of human duplicated genes.....	54
4.2.2. Functional similarity	56
4.2.3. Subcellular localization	56
4.2.4. Gene expression	57
4.2.5. Evolutionary rate	58
4.2.6. Multifunctionality	58
4.2.7. Gene essentiality	59
4.2.8. Disease genes.....	59

4.2.9. Software.....	59
4.3. Results	59
4.3.1. Functional similarity of human SSD and WGD pairs.....	59
4.3.2. Subcellular localization of SSD and WGD pairs	60
4.3.3. Gene expression correlation between SSD and WGD pairs.....	64
4.3.4. Comparison of human whole-genome duplicates with young and old small-scale duplicates.....	65
4.3.5. The difference between Small-scale and Whole-genome duplication in <i>Xenopus tropicalis</i> genome	68
4.3.6. Evolutionary rate of human SSD and WGD genes.....	69
4.3.7. Multifunctionality of human SSD and WGD genes.....	70
4.3.8. Gene essentiality of human SSD and WGD genes.....	72
4.3.9. Disease association of human SSD and WGD genes	73
4.4. Discussions	74
4.5. Conclusions.....	80
Chapter 5: Summary and General Conclusion.....	82-84
References.....	85-98
Publications	99
Reprints	

Chapter 1

Introduction

Nothing in Biology Makes Sense Except in the Light of Evolution."

—Theodosius Dobzhansky (1973)

Evolution: the process behind all life forms in the earth

The term 'evolution' was originated from Latin '*evolutio*' that means 'unrolling' or 'unfolding'. In biology, 'evolution' refers to the changes acquired by an organism over successive generations. Although the foundation of biological evolution is based on Charles Darwin's famous book "*On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*", the word 'evolution' was carefully avoided by Darwin in this book. Evolution depicts the generation of complex organisms from pre-existing simpler ones, upon acquiring changes that are required for such an increase in complexity. Such changes are very slow and begin at the molecular levels. However, the cumulative effect of such changes are large in course of billions of years and is the reason of such a huge variety of existing life forms existing today. In this thesis, we will focus on evolution at the molecular level, that bring changes to the sequence of gene, which is the unit of heredity and serves as the functional unit of life.

1.1. Origin of life on Earth:

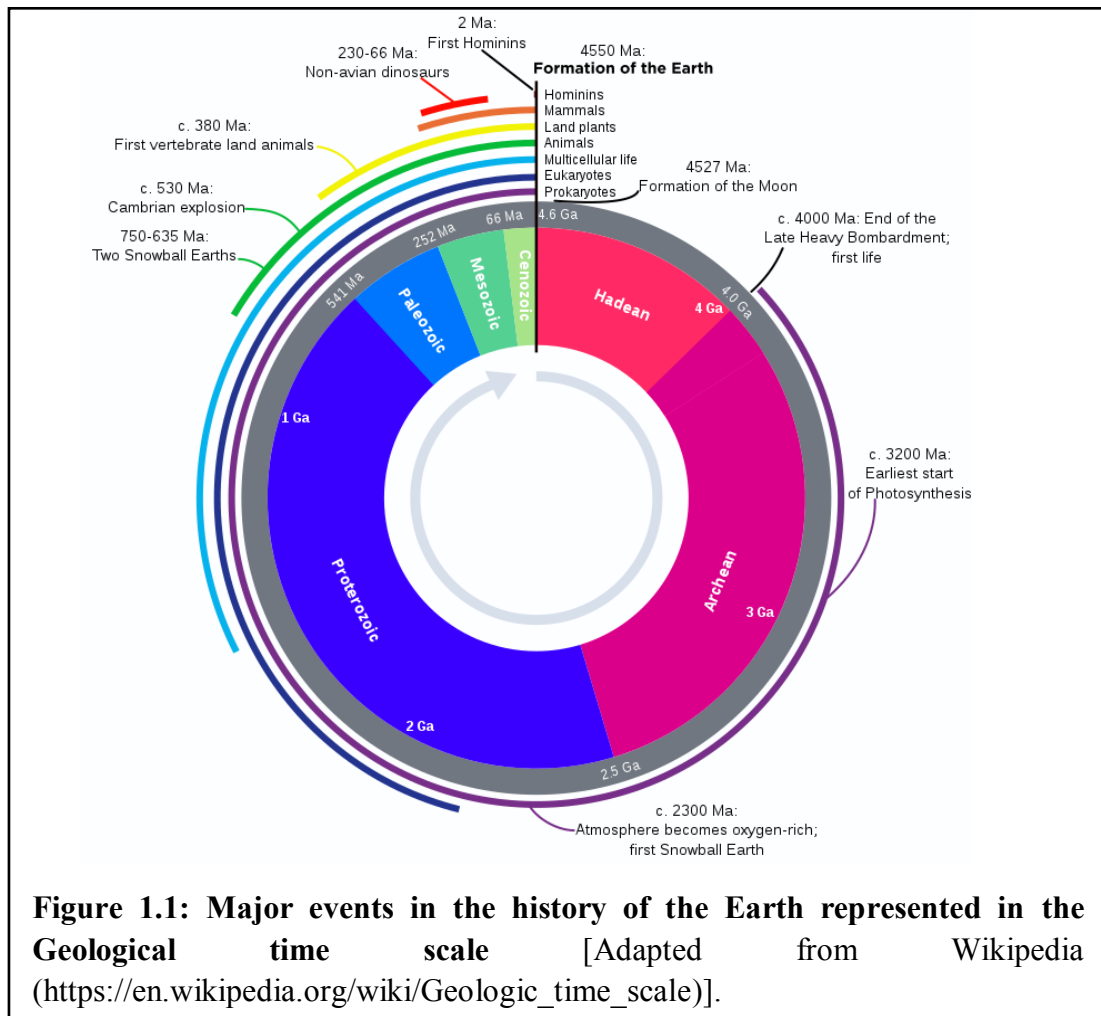
1.1.1. The ancient environment of the Earth:

Starting from the formation of the earth to the origin of life which subsequently resulted in such a massive variety of life-form existing today, the mystery of evolution is not yet completely solved and remains debatable. The evidence obtained from ancient rocks suggest that the earth had originated around 4-4.5 billion years ago from the solar nebula (Bowring, Williams 1999; Dalrymple 2001; Wilde *et al.* 2001) (Figure 1.1). After its formation, the earth was a burning planet with very high temperature, and its atmosphere contained hot gases and vapours of various elements. Elements like carbon, hydrogen, nitrogen, oxygen did not exist in their free state, but remained combined to each other or with other elements. The absence of free oxygen and the ozone layer along with the abundance of hydrogen and vapours generated from volcanic eruptions made the earth very much different from how it is today. Gradually, the earth's temperature dropped, and the surface became a thin and solidified 'crust', with the center being still hot and named 'core'. The solid area separating core and crust is known as 'mantle'.

1.1.2. Formation of biomolecules essential for life:

After the temperature of earth crust dropped down, chemical changes lead to the formation of larger biomolecules from preexisting smaller ones. This includes the formation of simple carbohydrates (monosaccharides), aldehydes, simple fatty acids, glycerol and amino acids. These simple biomolecules then accumulated, reacted and aggregated to form more complex organic compounds including disaccharides, polysaccharides, polypeptides, proteins, fats, purines and pyrimidines. The formation of

nucleic acids was the first big step towards the origin of living beings. The presence of RNA as genetic material in retroviruses suggests that RNA was the first genetic material. In the course of evolution, DNA became the genetic material containing all the genetic information of an organism. Such



information gets translated into proteins that perform all the basic need of the cells. However, the primitive cells were dependent on chemicals to obtain energy and known as Chemoheterotrophs. The abundance of chemotrophs leads to a subsequent decrease in the organic energy source. Thus some of the cells became specialized and developed the ability to photosynthesis (Photoautotrophs). From anaerobic photoautotrophs arose the ancestor of modern plants (Aerobic photoautotrophs). Photosynthesis by

aerobic photoautotrophs leads to the generation of free oxygen molecules in the earth's environment.

1.2. Cells and the Genetic material:

1.2.1. Cells and Cellular organelle:

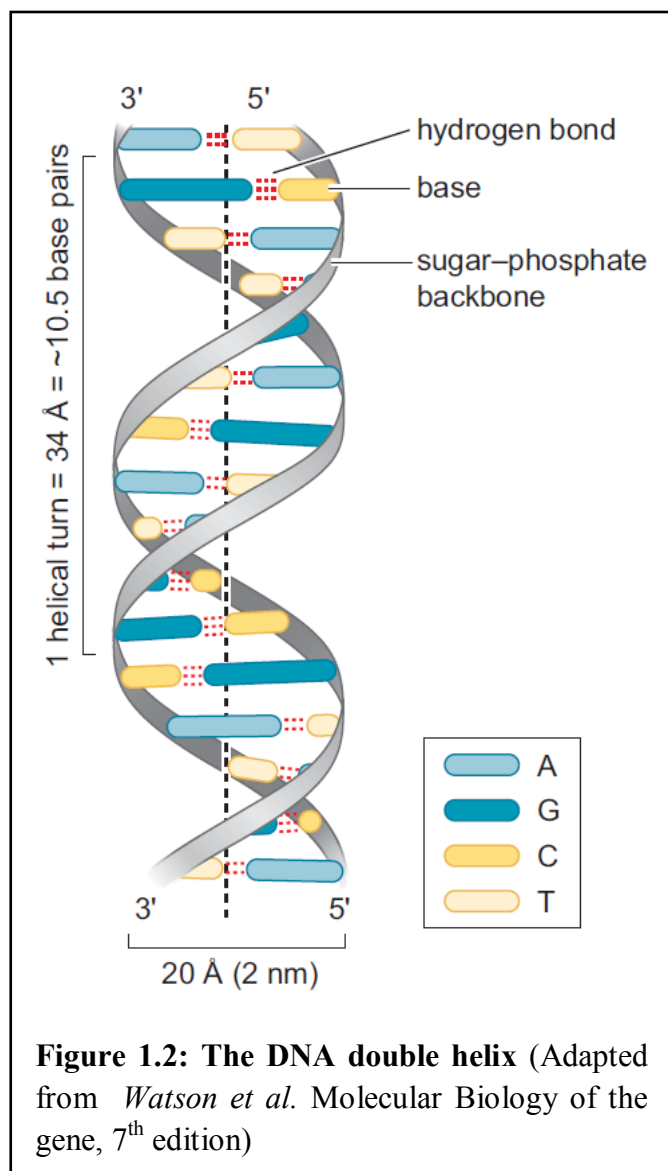
Cells are the smallest units of life composed of different cellular organelle enclosed in a selectively permeable 'Plasma membrane'. The cells are capable of dividing themselves, leading to the formation of daughter cells. The enclosed cellular organelles perform all the basic life processes required for the survival of the cell. The formation of the first living cells was the first sign of life on the earth. These cells eventually gave rise to two types of unicellular organisms- Monera, organisms with cells lacking a distinct nucleus and Protista, representing organisms having a distinct nucleus. Monera became diversified to form prokaryotes while protista became evolved to form eukaryotes. Prokaryotes are unicellular, representing the two domains of life: Bacteria and Archea. Eukaryotes represent a wide variety of organisms, starting from the unicellular protozoans like Amoeba to the most complex multicellular organisms like Humans. Thus, the process of evolution is very complicated, as suggested by the existence of simple unicellular prokaryotes to multicellular plants and animals.

Among the cellular organelle, the largest and most important one is the nucleus. Nucleus contains the genetic material of an organism and is responsible for the characteristic features of the cell. The genetic material become transferred to the daughter cells during cell division and hence, the daughter cells resemble the characteristic features of its parents. Although, RNA is considered the first biomolecule that evolved as genetic material, and still plays the same role in some retroviruses, DNA serves as the genetic

materials in both prokaryotes and eukaryotes. The DNA contains all the genetic information of a cell and has the capability of self-replication. Regions on DNA (known as genes) serve as a template to 'transcribe' messenger RNAs (mRNA) containing the desired information carried by the DNA. The mRNA gets 'translated' into proteins that perform the function of the region of DNA.

1.2.2. The structure of DNA:

The structure of a DNA was revealed by James Watson and Francis Crick in 1953, and confirmed with the X-ray crystallographic analyses of Rosalind Franklin, consequently leading to the 1962 Nobel Prize in Physiology or Medicine to Watson, Crick and Maurice Wilkins. Their study revealed that the DNA is a polymer composed of monomeric units known as nucleotides comprising a nitrogenous base, a pentose (deoxyribose) sugar, and a phosphate molecule. There are four types of nucleotides classified on the basis of the nitrogenous bases:



Adenine(A), Guanine(G), Thymine(T) and Cytosine(C). A and G are known as Purines (Pu), whereas T and C are known as Pyrimidines (Py). The structure of DNA is a double-stranded helix of 34 Angstrom (Å) or 3.4nm pitch and 20Å (2 nm) diameter. A DNA molecule comprises two antiparallel strands made up of nucleotides where the sugar and phosphate form the backbone of each strand via an ester linkage (phosphodiester bond), and nitrogenous bases are arranged in a staircase-like fashion inside the double helix (Figure 1.2). Both ends of a DNA contain a free 5' and 3' end in their terminal sugar residues. The hydrogen bonds between the nitrogenous bases of these two strands stabilize the double helix structure. Among the bases, A always pairs with T and G always pairs with C with two and three hydrogen bonds, respectively. Thus, a DNA typically contains an equal amount of Pu (A+G) and Py (T+C), a phenomenon known as Chargaff's rule, named after Austrian chemist Erwin Chargaff (Chargaff, Lipshitz, Green 1952).

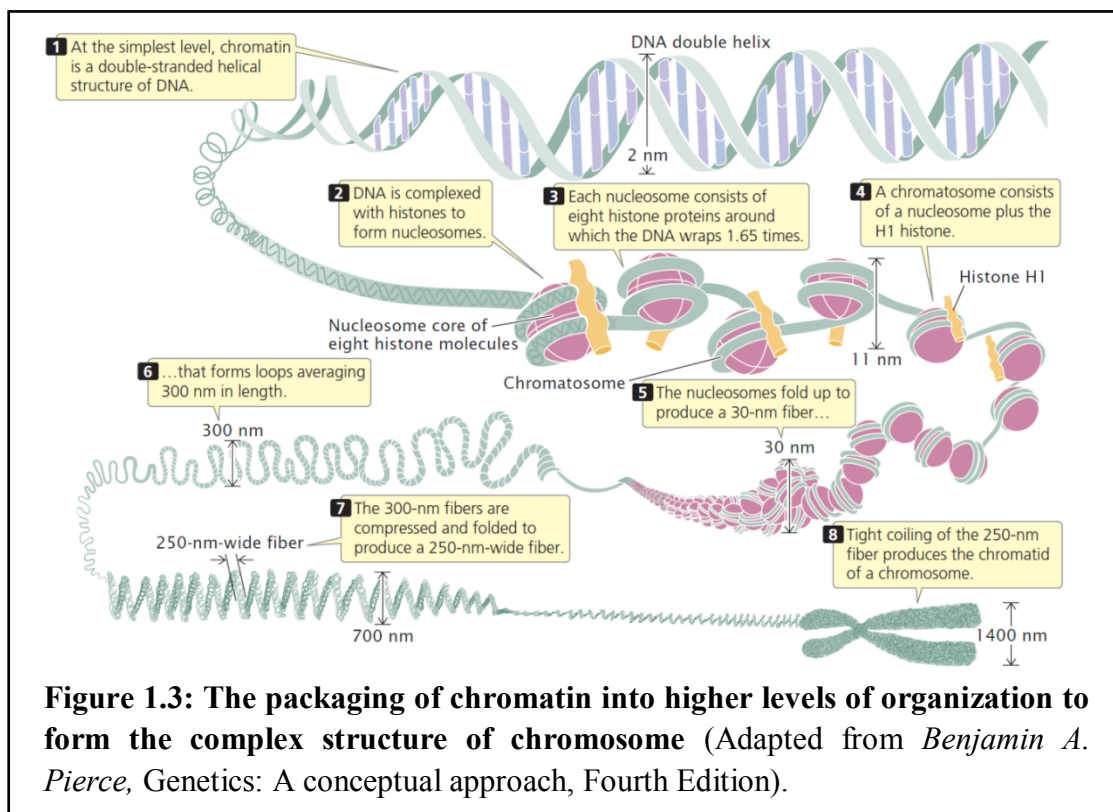


Figure 1.3: The packaging of chromatin into higher levels of organization to form the complex structure of chromosome (Adapted from *Benjamin A. Pierce, Genetics: A conceptual approach, Fourth Edition*).

However, DNA is a very long molecule as it carries all the genetic information of an organism. The human DNA totals to a length of about 3 meters, which is huge when compared to the diameter of the cell, which measure up to a maximum value of 1mm in ovum, the largest cell of the human body. However, the DNA fits inside the nucleus in a very condensed manner with the help of many proteins which 'pack' the less condensed DNA into more condensed structure, the chromosomes. The DNA packaging involves histone proteins, which are basic in nature and forms a stable octameric core surrounded by the 146bp of double-stranded DNA. The nucleosomes eventually form higher order structures and finally forms the chromosome (Figure 1.3) The presence of nucleosome is not only important for efficient packaging of DNA, but it also helps to regulate gene expression.

1.3. From Genotype to Phenotype:

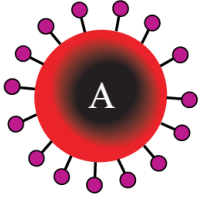
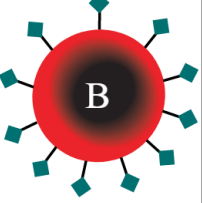
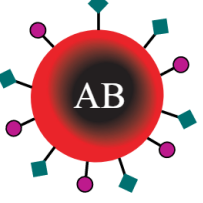
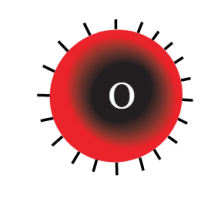
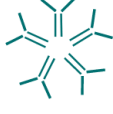

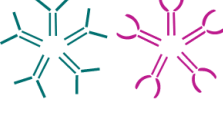



1.3.1. Genes and Genomes:

Gene is the molecular unit of heredity located in a particular region or 'locus' on the DNA or chromosome and governs a particular characteristic of an organism. The number of genes varies greatly among organisms, ranging from as few as ~500 in the *Mycoplasma genitalium* (Fraser *et al.* 1995) to ~23000 in humans. The record of highest number of genes is held by the near-microscopic freshwater crustacean *Daphnia pulex* (water flea) which contains ~31000 genes (Colbourne *et al.* 2011). The complete set of genes present in an organism is known as the 'genome' of that organism. Fundamentally, all the cells in an organism's body are identical in terms of their number of genes. Genes determine the genotype of an organism and govern the phenotypic characters (traits) of that organism. The genes within an organism come from its parents, by a process known as 'inheritance'.

1.3.2. Alleles, phenotypes, and genotypes:

A gene may have different forms, known as 'alleles' that are responsible for the variation of a particular phenotypic trait across different individuals of the same species. These alleles are situated in the same locus in homologous chromosomes, and thus, a diploid organism having two sets of identical chromosome can possess maximum two variants of a gene. For example, let's recapitulate the famous work by Sir Gregor Johan Mendel in pea plant for one gene two allele character (Monohybrid). Here, 'T' is the allele that encodes products responsible for regulating the length of the plant (TT= Tall). However, its mutant (recessive) allele 't' is not capable of doing so, and leads to the short height of plants having only 't' alleles in both their homologous chromosomes (tt= dwarf). However, the presence of a single 'T' allele is enough for the normal plant phenotype, as seen in the plants having genotype 'Tt' shows characters like 'TT' genotype, that is they are phenotypically 'tall' plants. Therefore both TT and Tt are phenotypically same, but genotypically different, as only one 'T' allele is responsible for performing all functions that are responsible for 'Tall' phenotype, and is considered as a dominant allele. However, a population (defined by the number of conspecific individuals at a spatiotemporal interval) comprising diploid species may possess more than one allele, a phenomenon known as 'Multiple allelism', with only a maximum of two variants of a gene being present in an individual's genome. One such example includes the ABO-Blood group system, where three different alleles are present in a population that determines the blood type. These are A, B, and O, that are surface antigens present on the surface of human red-blood corpuscles. Here, both A and B are individually dominant over O and are codominant with each other (Figure 1.4).

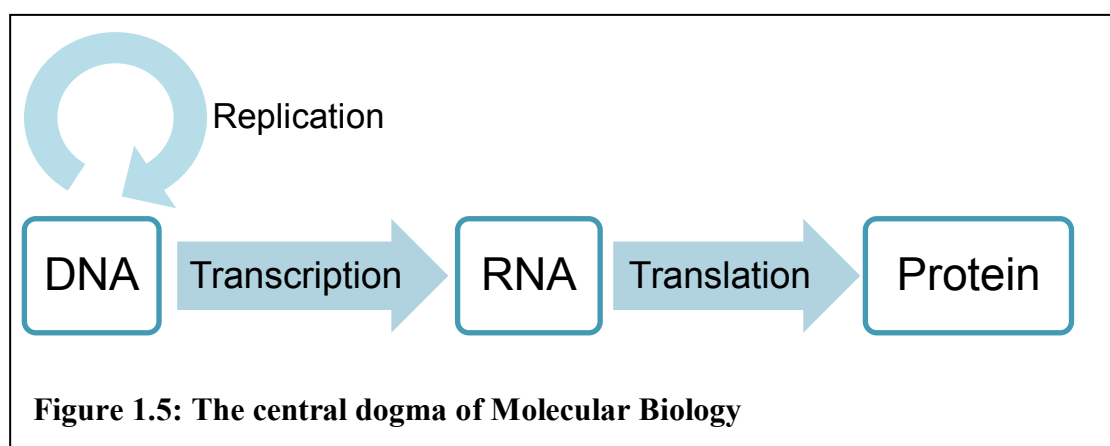
However, the genotype is not the sole determinant of the phenotypic traits, as most of the traits are governed by gene-gene and gene-environment interactions (Via, Lande 1985). But genes play the most important role in regulating and maintaining the cellular functions. Moreover, the function of a gene can be modified by the accumulation of mutations that changes the nucleotide sequences of the gene, which may prove neutral, beneficial or even detrimental to the organism, as we will see later in this chapter.

	Group A	Group B	Group AB	Group O
Red blood cell type				
Antibodies in Plasma	 Anti-B	 Anti-A	None	 Anti-A and Anti-B
Antigens in Red Blood Cell	 A antigen	 B antigen	 A and B antigens	None
Figure 1.4: The ABO blood group system in humans (Adapted from https://commons.wikimedia.org/wiki/File:ABO_blood_type.svg)				

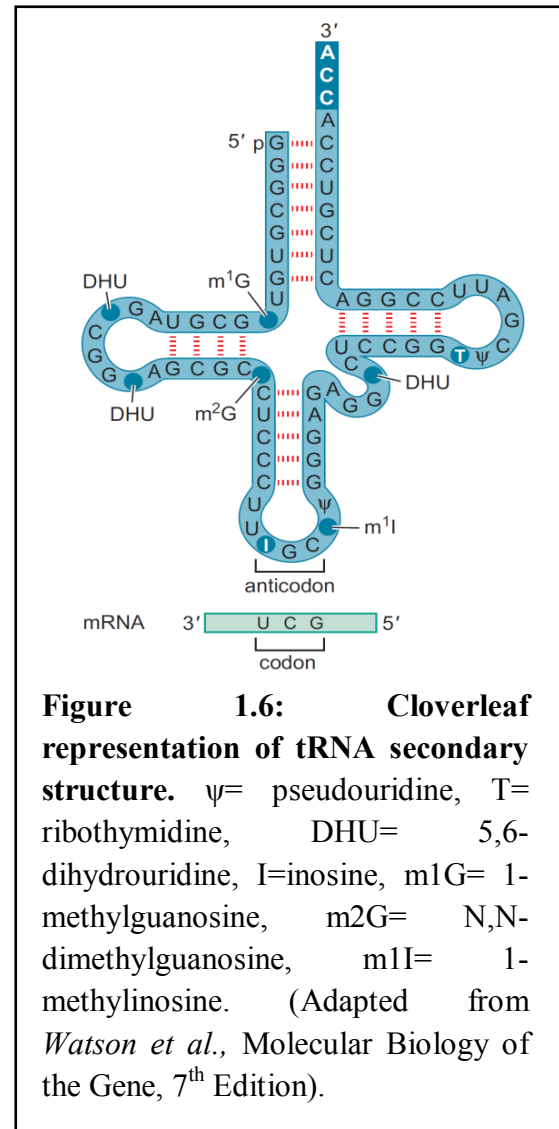
1.3.3. From Genes to Proteins- The Central Dogma of Molecular Biology:

As stated earlier, genes determine the phenotypic trait of a cell or an organism. For example, in a normal human being, all the cells contain same type and amount of genetic material, starting from the epithelial cells to the tubular muscle fibers and the shapeless white blood cells. However, they differ in terms of the genes that are ‘expressed’ within that particular cell type, a feature that makes those cells ‘unique’ and different from each other.

Each cell divides to give rise to daughter cells, genetically identical to the parent. Thus, the whole genetic material of a parental cell must be multiplied and shared equally among the daughter cells so that they contain the same amount of genetic material as their parent. The genetic material or DNA is capable of making copies of itself by the help of certain proteins, among which the most important ones are DNA polymerase, Helicase, Single-strand binding proteins, and DNA topoisomerase. This process is known as DNA replication, and it leads to the formation of two double-stranded DNA molecules from one such DNA after each round. However, the message of DNA is transferred from nucleus to cytoplasm via a cellular messenger known as mRNA (messenger RNA), which is transcribed from DNA with the help of DNA-dependent RNA-polymerases and other enzymes, by a process known as ‘transcription’. The mRNAs carry all the genetic information from the DNA and ‘translates’ the same in the form of polypeptides or proteins. Thus, the DNA makes DNA, as well as RNA via the process of replication and transcription, respectively; and the RNA produces proteins via the process of translation, a phenomenon known as the Central dogma of molecular biology (Figure 1.5). Stated by Francis Crick in 1958, the central dogma describes the direction of flow of genetic information within a cell(Crick 1970):



The central dogma also indicates that as the DNA is the most vital molecule and is central to all cellular functions, it should be maintained in a more protected state and the message of DNA is 'transcribed' in the form of RNA (messenger RNA). The formation of polypeptides or proteins is more complex, as it involves decoding of the message carried by mRNA via ribosomes, with the help of transfer RNAs (tRNAs) that transfer the information from nucleic acids to amino acids. The nucleotide sequence of the mRNA determines the amino acid sequence of a protein by proper tRNA-mRNA



codon-anticodon base pairing, where codon represents trinucleotide sequence on the mRNA and anticodon sequence is present in the anticodon arm of tRNA (Figure 1.6).

1.4. Mutations:

Mutations are changes acquired in the DNA sequence that serves as the raw materials of genetic variation and thus, in an important process of evolution. The term 'mutation' was coined by Hugo deVries, who represented it as large, sudden and spontaneous inheritable changes that occur suddenly in naturally reproducing populations. More recent definition of mutation

considers it as any hereditary change in the genetic makeup of an individual other than that which may be caused by the simple recombination of genes. Such changes lead to the formations of different variants of a gene, known as alleles. Mutations may occur at the level of gene structure or composition (referred to as gene mutations or point mutations), at the chromosomal level resulting in an alteration in chromosome structure or number (known as chromosomal mutations). Additionally, mutations may occur at somatic levels, affecting one or more tissues in the body and limited to an individual; or at germ-line tissues, appearing in the gametes and propagating to the next generation. However, only the germ-line mutations are inherited and play crucial roles in evolution. Therefore in this thesis, the term 'mutation' refers to the germ-line mutations. Although mutations bring changes in the gene structure, not all the mutations become expressed and showed up in the individual. Dominant mutations are expressed in the mutant organism(s), whereas recessive mutations may remain hidden for generations. Most of the mutations are recessive, a fraction of such mutations are even lethal, and becomes fatal for the organisms homozygous for the mutant allele. Mutations are classified according to their mutagenic effect on proteins' structure, function significance of mutation and the extent of mutation (Table 1.1).

1.5. Gene Duplication:

It was in 1936 when Calvin Blackman Bridges reported that duplication of a chromosomal segment leads to a severe reduction of eye-size in *Drosophila melanogaster* (Bridges 1936). Subsequent studies aiming towards the evolutionary significance of gene duplication has drawn much attention (Stephens 1951; Ohno, Wolf, Atkin 1968; Nei 1969). In 1951, S.G. Stephens

comprehended the significance of gene duplication in evolution, as an increment of genomic loci, to produce more gene required for increasing complexity in the course of evolution. He hypothesized that if mutations are required to develop novel gene functions, it is likely to hamper pre-existing ones, and gene duplication provides a shield to this, by preserving both the novel and ancestral forms and functions of the genes (Stephens 1951). Afterwards, in 1970, Ohno postulated that gene duplication is the major driving factor that brings about genome evolution.

Fundamentally, gene duplication is a genetic mutation that leads to an increase in the genetic material of an organism by doubling of gene(s). With many of the deformities associated with such type of genetic mutation (Dickerson, Robertson 2012; Veitia, Bottani, Birchler 2013; Malaguti, Singh, Isambert 2014; McLysaght *et al.* 2014), gene duplication is the major driving force of genome and organism evolution as it supplies raw genetic materials for structural modifications like mutation, leading to functional modifications and the origin of new functions. The possible modes of duplication include: 1. unequal crossing over, which happens during chromosomal rearrangement, 2. Retrotransposition; where the messenger RNA of a gene gets reverse transcribed, and the complementary DNA is integrated back into the genome; and 3. Polyploidization, where the whole set of chromosomes get duplicated. The evolutionary time scale depicts that all life forms present today were originated from simplest unicellular organisms upon acquiring further complexities. Such complexities include development of multicellularity, tissue-level organization, complex metamorphosis and embryogenesis, regulated cellular processes and transition from asexual to sexual reproduction. However, such transition is never easy, as it requires many biological factors and cellular machinery, of

Table 1.1: Types of Mutations

Classification of mutations	Types of mutation	Description
Based on the effects of mutations on proteins' structure	Nonsense Mutations	Point mutation leading to premature stop codon in the mRNA e.g., UGG > UAG trp > STOP
	Missense Mutations	Point mutation leading to the formation of a different amino acid leading to altered protein product. Effects may be minor to drastic, or even lethal. e.g., GAG > GUG glu > val
	Neutral mutations	Point mutation resulting different, but chemically similar amino acid. Leads to little or no harmful effect in the resultant protein. e.g., AAA > AGA lys > arg (Both are basic in nature)
	Silent mutations	Point mutations leading to no change in the amino acid sequence of the encoded protein. e.g., UUG > CUG leu > leu
	Frameshift mutations	Mutations leading to the change in the reading frame. Results from insertions/deletions that are not multiples of three nucleotides. Leads to the formation of a completely different Protein product. The earlier in the DNA sequence the mutation occurs, the encoded proteins become more altered. e.g., AUG AAC CUA CUG... > AUG AAG CCU ACU G... met - asn - leu - leu...> met - lys -pro - thr...

Table 1.1: Types of Mutations (continued)		
Classification of mutations	Types of mutation	Description
Based on the functional significance of mutations	Loss of function mutation	Mutations resulting in gene product having less or no function.
	Gain of function mutations	Mutations leading to the gain of new functions.
	Dominant negative mutations	Mutations leading to the production of gene products acting antagonistically to the wild-type gene product
	Lethal mutations	Mutations resulting in the death of the organisms carrying the mutant copy of the gene.
	Reversion/ Back mutation	Mutation that restores the wild-type function in a previously mutated gene.
Based on the extent of mutations	Small-scale mutations	<p>Mutations affecting one or a few nucleotides.</p> <p>Are of three types-</p> <ul style="list-style-type: none"> i) Point mutations ii) Insertions iii) Deletions.
	Large-scale mutations	<p>Mutations leading to a change in the chromosomal structure.</p> <p>Are of two types-</p> <ul style="list-style-type: none"> i) Duplications (or amplifications)- creates multiple copies of gene(s) or chromosomal regions, increasing the products of the gene(s) located within the region. ii) Deletions- Removal of chromosomal regions, leading to loss of the genes located within those regions.

which proteins play an important part. Proteins are nitrogenous biomolecule composed of one or more long chains of amino acid residues and are encoded by genes located in the chromosome(s). The number of genes in an organism may vary from a few hundred genes in some bacteria to more than 20000 genes in humans. Therefore, such a huge variation in the number of genes corresponds to the huge variation in the complexity of these organisms. However, such an increase in gene number requires an efficient evolutionary mechanism that helps in generating new genes that are retained in the course of evolution. Gene duplication supplies raw genetic materials required for functional innovation that are key to evolution. Additionally, large evolutionary transitions requiring many genes are also achieved by large-scale gene duplications, that leads to the duplication of many genes, chromosomal segments, whole chromosomes and even the whole-genome.

1.5.1. The Contribution of gene duplication in Evolution:

Susumu Ohno was among the first evolutionary biologist to explain the 'Evolution by gene duplication' hypothesis (Ohno, Wolf, Atkin 1968; Ohno 1970). He also postulated the possible outcomes of gene duplication, stating that in most of the cases gene duplication are unfavorable as it leads to the generation of 'useless duplicates' that leads to nonfunctionalization, or functional disruption (Ohno 1970). However, a vast number of duplicated genes throughout all three domains of life indicate that such duplicates may have been retained. Such retention of duplicates are favored in the circumstances like increased gene dosage advantage, where the subsequent increase in gene product is favorable for the organism (Innan, Kondrashov 2010). In the long term, these duplicated gene copies may serve as backup copies or diversify to acquire novel function [Neofunctionalization] or to

partition ancestral functions after complementary degenerative mutations [Subfunctionalization] (Clark 1994; Force *et al.* 1999). Whereas the sub- and neofunctionalization helps in functional specialization and generation of novel functions, respectively, and helps in genome evolution, the duplicated genes that remain structurally and functionally similar helps to increase genetic robustness against deleterious mutations (Gu *et al.* 2003; Liang, Li 2009). However, as proteins work together to perform certain biological function(s), and are involved in a protein-protein interaction (PPI) network that serves all of their functions. Thus, from the perspective of PPI-network, gene duplication may not be favorable, as it increases the amount of encoded protein product (dosage) for that protein only. This creates a disparity in the protein interaction network, as the interacting partners of the concerned gene produces normal dosage of their protein products. Such disparity is known as 'dosage imbalance' and is dependent on the connectivity of a protein in the PPI network, known as the centrality-lethality rule (Jeong *et al.* 2001). Thus, in a PPI network, the relative dosage of all of its participants should be maintained to accomplish optimum functionality. However, in higher organisms like humans, gene dosage is strictly regulated by the presence of efficient dosage regulatory mechanisms (Li, Musso, Zhang 2008). The presence of dosage-balanced duplicated copies, therefore, is advantageous, as it provides shield against deleterious mutations without any disparity in the protein interaction network.

1.6. Origin of the proposal:

Gene duplication provides raw genetic materials for genome innovation and evolution (Stephens 1951; Ohno, Wolf, Atkin 1968; Ohno 1970; Zhang 2003). Although every gene has its own function in an organism's life, there are genes essential for its survival and reproduction. These genes, known as essential genes, causes organismal sterility and/or lethality upon their deletion (He, Zhang 2006b; Liao, Zhang 2007; Makino, Hokamp, McLysaght 2009). Essential genes are associated with essential biological functions and are detected by analyzing loss of function mutations. However, there are many genes considered as nonessential despite performing such vital functions, due to the functional restoration in loss of function mutants by their duplicated copies. Therefore, the duplication of vital genes provides a selective advantage due to the increased robustness against deleterious mutations. However, such a duplication event may not be favorable, as proteins work together to perform certain biological functions, and duplication increases the amount of encoded protein product (dosage) for that protein only (He, Zhang 2006b). This creates a stoichiometric imbalance within the protein interaction network, as the interacting partners of the concerned gene maintain their normal protein dosage. Such disparity, known as 'dosage imbalance' is even more pronounced for essential genes, as they are highly connected (hub-like) in protein-protein interaction network (He, Zhang 2006a). Dosage imbalance is one of the main reasons of lower duplicability of essential genes in the lower unicellular eukaryotes (He, Zhang 2006b), largely due to the lack of dosage regulatory mechanisms (Springer, Weissman, Kirschner 2010). However, in higher organisms like humans, gene dosage is strictly regulated by the presence of efficient dosage regulatory mechanisms (Li, Musso, Zhang 2008). The presence of dosage-balanced

duplicated copies, therefore, is advantageous, as it provides a shield against deleterious mutations without any dosage imbalance in the protein interaction network.

Additionally, there is also evidence of gene duplication while keeping the dosage of protein interaction network in balance. The extent of gene duplication ranges from Small Scale Duplication (SSD; usually involving one or a few gene) to large-scale duplication that may comprise the duplication of the whole genome (Whole Genome Duplication or WGD). These two extents of duplication affect their associated protein-interaction network differentially (Lynch, Conery 2000; Freeling, Thomas 2006; Hakes *et al.* 2007; Makino, McLysaght 2010; Fares *et al.* 2013). In WGD, all the proteins within a PPI network become duplicated at the same time, thus resulting in a stable stoichiometric balance of the participant proteins even after the duplication. In SSDs, however, the duplicated gene(s) form more protein than the non-duplicated participants of the PPI network, creating an imbalance in the whole network. SSD occurs at any time and the duplicates formed by it are retained in the course of evolution, upon favorable circumstances. WGDs, in contrast, are much rarer in their occurrence in eukaryotes, being most common and widely studied in the evolution of plant genome (Stebbins 1971; Blanc *et al.* 2000; Wendel 2000; Adams, Wendel 2005a). An ancient whole-genome duplication in the genome of the simplest unicellular eukaryote yeast (Wolfe, Shields 1997; Dujon *et al.* 2004; Kellis, Birren, Lander 2004) leads to the comparison of the SSDs and WGDs. Such a comparison has revealed various differences (Guan, Dunham, Troyanskaya 2007; Hakes *et al.* 2007; Fares *et al.* 2013). Yeast WGD pairs are functionally more similar to each other than SSD-pairs, which is independent of their sequence similarity (Guan, Dunham, Troyanskaya 2007; Hakes *et al.* 2007). Additionally, SSDs

also diverge more at their subcellular localization than the WGDs (Fares *et al.* 2013). Also, yeast SSD genes were found to contain a higher proportion of essential genes than WGD genes (Guan, Dunham, Troyanskaya 2007; Hakes *et al.* 2007). Additionally, there are concrete evidence of two rounds of whole-genome duplication during the evolution of early vertebrates (Zhou, Cheng, Tiersch 2001; Dehal, Boore 2005; Brunet *et al.* 2006a). Such a genome duplication provided the raw materials for such an extensive species diversity of vertebrates (Zhou, Cheng, Tiersch 2001; Dehal, Boore 2005) and hence, is an important process in vertebrate evolution (Allendorf, Thorgaard 1984; Dehal, Boore 2005). The accumulation of human genomic and proteomic data plays an important role for in-depth analysis of human genes. Based on these data and the studies mentioned above, my proposed work aims to explore the role of human duplicated genes in human genome evolution under the following objectives:

Objectives:

- Determining the proportions of essential genes retained as singletons and as duplicates in humans.
- Interpretation of the functional role and evolutionary significance of duplicated gene copies by comprehensive analysis using human essential duplicated genes.
- Exploring differences in evolutionary and genomic attributes in human duplicated genes arising from small-scale duplication with those originating from whole-genome duplication.
- Investigating the fate of small-scale and whole genome duplicates to explore the long-term evolutionary consequences of vertebrate whole genome duplication.

- Understanding the evolutionary rate differences between human small-scale and whole genome duplicates.
 - Estimation of the functions of human whole-genome duplicate genes and evaluation of the importance of such functions to understand their importance in human genome.
-

1.7. Thesis Organization:

The whole work was carried out in Prof. Tapash Chandra Ghosh's laboratory at the Bioinformatics Centre, Bose Institute, Kolkata, India. This thesis integrates my published works during my Ph.D. tenure. The thesis starts with **Chapter 1: Introduction**, explaining the background to the study; followed by the **Chapter 2: Resources and methodology**, summarizing a clear description of the databases, materials and protocols used for the whole work. Then the individual chapters representing specific topics under the study are described.

Chapter 3 explores the duplication pattern of essential genes in human and mouse. Here, we have observed that the human duplicated genes contain a higher proportion of essential genes, whereas in mouse, a higher proportion of essential genes are retained as singletons. Our cross-species comparison comprising mouse and human essential duplicate genes reveal that human essential duplicates are functionally less diverged and evolutionarily more conserved than that in mouse, revealing their ability to serve as backup copies to increase robustness against gene-deletion. We have also observed a higher enrichment in micro-RNA target sites among the paralogs of human essential genes than the mouse counterparts, representing the role of micro-

RNAs in maintaining the gene dosage of backed up duplicated copies of human essential genes.

Chapter 4 aims to reveal the importance of whole genome duplication in human evolution. Here, we compared the human small-scale duplicates (SSD) with their whole-genome duplicates (WGD). The latter class of duplicated genes had originated during the genome duplication occurred early in vertebrate evolution and thus our study aims to explore the long-term evolutionary fate of these duplicates. We observed a lower functional similarity, lower subcellular colocalization and lower coexpression among the WGD pairs, indicating that these duplicates tend to diverge more in their function and expression. Further, our detailed functional analysis suggests that the WGD duplicates are associated with more variety of functions and functions that are crucial for the survival of humans.

Chapter 5 gives a general summary of our work and conclusion from the above studies.

The Thesis ends with the literatures that have been followed and cited for the above-mentioned works and the reprints of our works relevant to the thesis followed by those that are not related to the thesis.

Chapter 2

Resources and Methodology

“A theory can be proved by experiment; but no path leads from experiment to the birth of a theory.”

— Albert Einstein (1976)

With the advent of new genomic tools and accumulation of genomic data from large-scale genomic experiments and a vast number individual studies, genome-scale high throughput data analysis became feasible. This leads to large-scale genomic analysis in a wide-range of organisms, rather than focusing in a particular gene group or species, resulting in the extraction of meaningful biological information. It also helps in the comparative genomic analysis in similar or diverse species to study the similarities or differences in their evolutionary genomic properties. Additionally, the development of powerful statistical tools have made the validation of the hypothesis that are put forward in a more powerful and rigorous way.

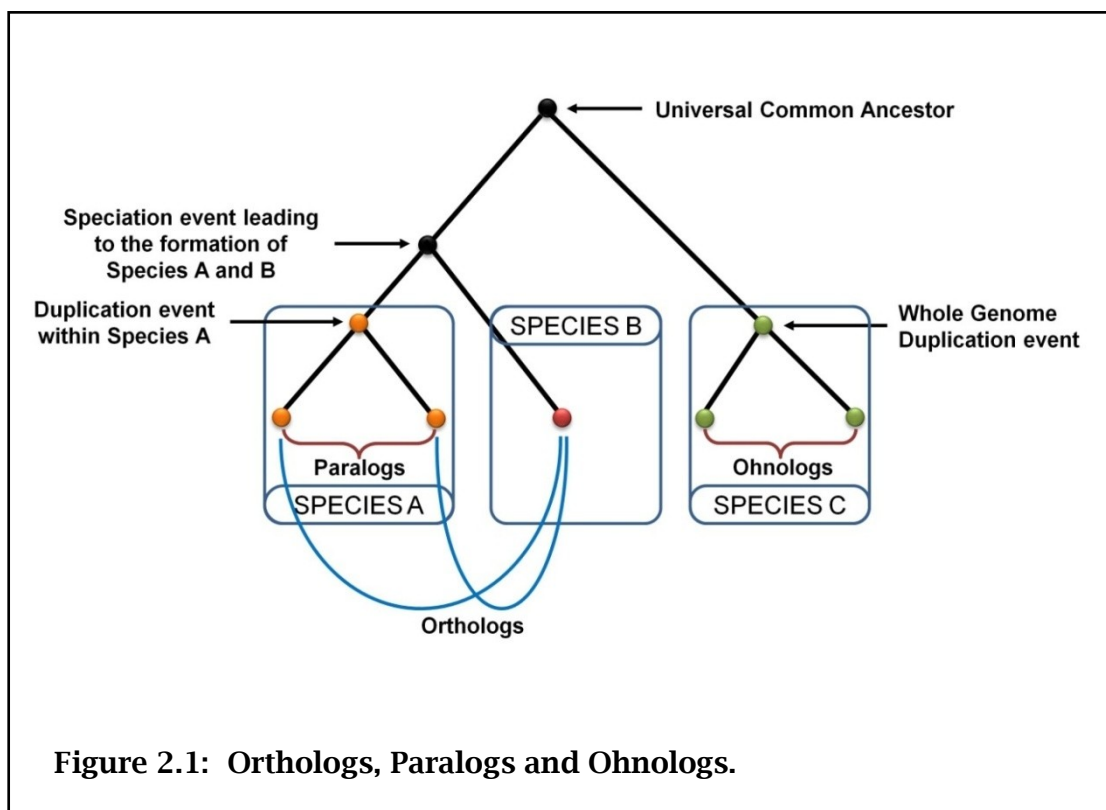
2.1. Gene Essentiality:

In every organism, there are a number of genes crucial for the purpose of its survival and reproduction and is indispensable for its genome. These genes are collectively known as essential genes. The essential genes are thus indispensable to the viability of the organism (Seringhaus *et al.* 2006; Hwang *et al.* 2009; Wang, Peng, Wu 2013), and are associated with basic cellular and molecular functions, thereby representing the minimal set of genes required for cellular survival (Juhas, Eberl, Glass 2011; Yang *et al.* 2014). The identification of essential genes has been a major goal in genomic analysis and is usually done by disrupting the function of the gene by experimental approaches involving gene-knockouts (Giaever *et al.* 2002), conditional knockout (Roemer *et al.* 2003), gene-knockdown by RNA-interference (RNAi) (Kamath *et al.* 2003; Cullen, Arndt 2005; Silva *et al.* 2008; Chen, Zhang, Long 2010), and gene-trap mutagenesis (Blomen *et al.* 2015) *etc.* For our study, the essential genes in human and mouse genome were retrieved from the Online Gene Essentiality database (OGEE) (<http://ogeedb.embl.de/>)(Chen *et al.* 2012b) which estimates gene essentiality based on si-RNA mediated gene-knockdown fitness effect in humans, and gene-knockout fitness effect in humans. OGEE contains 1528 (7.38%) human essential genes based on genome-wide experiments data on 20,684 genes and 2618 (43.36%) mouse essential genes based on genome-wide experiment on 6038 genes.

2.2. Homologous genes- Orthologs, Paralogs and Ohnologs:

Homologous genes refer to the genes originated from a common ancestor and are identified by their structural and functional similarities. There are two types of homologous genes- Orthologs are defined as homologs originated as a result of speciation and thus refer to the homologous genes

in different species. In contrast, Paralogs are homologs originated as a result of gene duplication within species and thus refer to homologous genes within the same species. Ohnologs however, refer to the duplicated genes originated from whole-genome duplication events and named after the renowned Evolutionary Biologist, Susumu Ohno (1928-2000), who first proposed the two rounds of WGDs in the genome of vertebrate ancestor. Thus, ohnologs simply refer to the paralogous genes originated from whole-genome duplication event (Figure 2.1). For our studies, we obtained the



orthologous and paralogous genes from the biomart interface of Ensembl genome browser (Cunningham *et al.* 2014; Flicek *et al.* 2014) (<http://www.ensembl.org/biomart/martview>). Ohnologs in the human genome were obtained from the database 'OHNOLOGS, a repository of genes retained from whole genome duplications in the vertebrate genomes' (<http://ohnologs.curie.fr/>) (Singh, Arora, Isambert 2015a), along with a previously published data of Makino and McLysaght (Makino, McLysaght

2010). In both the datasets, human ohnologs and families were constructed using a quantitative multiple-genome comparative approach.

2.3. Protein Evolutionary Rate:

Since the discovery of the amino acid sequences of homologous proteins in the 1950s and 1960s, the rate of protein sequence evolution has been of enduring interest to evolutionary biologists (Zuckerlandl, Pauling 1965; Kimura 1968). The protein evolutionary rate refers to the changes acquired over time in a protein's sequence. Such changes play an important role in protein evolution and are used to reconstruct the evolutionary history of these species. In practice, the measurement of protein evolutionary rate involves the sequence alignment of the DNA encoding that protein and its orthologous DNA sequence from ancestral species. After this alignment, the nucleotide substitution is measured. The nucleotide substitutions are subdivided into two parts- (1) Synonymous, that does not bring any change to the amino acid of its encoded protein; and (2) Nonsynonymous, that changes the amino acid of its encoded protein. Kimura used the ratio (ω) of synonymous substitution per synonymous sites (dS) and the non synonymous substitution per nonsynonymous sites (dN) as the evolutionary rate of the protein (Kimura 1968). This ratio (ω) is also used to measure the selection pressure, where $\omega=1$ signifies no or neutral selection, $\omega<1$ denotes negative or purifying selection and $\omega>1$ indicates positive or diversifying selection. In our study, we obtained the dN and dS of mouse and human genes from the Ensembl biomart (Cunningham *et al.* 2014; Flicek *et al.* 2014) and calculated the dN/dS ratio.

2.4. Pairwise comparison of gene functions:

The functional similarity and functional divergence compares the function of any of two or more genes and are opposite to each other. The functional similarity refers to the proportion of functions that are shared among the genes in concern, whether functional divergence calculates the proportion of functions that are different between those genes. Both the values range from 0 to 1, and are calculated using the Gene Ontology terms of the genes in concern. Gene Ontology (GO) uses GO-term classifications to describe gene function, and relationships between these terms. It classifies the functions of all genes in three aspects:

1. GO Molecular Function, associated with molecular activities of the gene products.
2. GO Biological Process, dealing with the pathways and larger processes of which the gene in concern takes part.
3. GO Cellular Component, describing the subcellular locations where gene products are active.

The GO term 'GO Cellular Component' also has a very significant role, as it describes the subcellular components in which the encoded protein(s) of a gene is localized. Thus, it is widely used to study the similarity, and/or divergence of two or more genes in their subcellular protein localization, using the same set of formula used to calculate the functional similarity and divergence, respectively. The formulae are as follows-

1. Bayesian data integration method:

$$\text{Functional Similarity } (i, j) = \frac{2 \times S(i, j)}{[\text{GO terms}(i) + \text{GO terms}(j)]}$$

Where '*i*' and '*j*' are duplicated pairs, '*S (i,j)*' represents the Gene Ontology terms shared between the duplicated pairs '*i*' and '*j*'. The values range from 0

to 1, where '0' means lowest functional similarity (complete functional divergence) and '1' means highest functional similarity (No functional divergence) (Table 2.1).

2. Czekanowski-Dice distance formula for functional distance (Baudot, Jacq, Brun 2004)

$$\text{Functional Distance } (i, j) = \frac{\text{Number of Terms } (i) \Delta \text{ Number of Terms } (j)}{\text{Number of } [\text{Terms } (i) \cup \text{Terms } (j)] + \text{Number of } [\text{Terms } (i) \cap \text{Terms } (j)]}$$

Here, i and j denote a gene and its paralogous gene within a species. Terms (i) and Terms (j) are the lists of the GO terms for individual genes. ' \cup ' and ' \cap ' denotes the nonredundant and common GO id count, respectively, of the two genes. ' Δ ' is the symmetrical difference between the GO term sets of two genes, i.e. ' $(\cup - \cap)$ '. The values range from 0 to 1, but here, '0' means the lowest functional distance (complete functional similarity) and '1' means the highest functional distance (No functional similarity) (Table 2.1).

Table 2.1. The calculation of functional similarity and functional distance								
Gene 1	Gene 2	Gene 1 GO ids	Gene 2 GO ids	Common GO ids	Nonredundant GO ids	Functional Similarity	Functional Distance	
i	j	GO00001, GO00002, GO00003.	GO00003, GO00004, GO00005, GO00006.	GO00003	GO00001, GO00002, GO00003, GO00004, GO00005, GO00006.	(Bayesian data integration method)	(Czekanowski-Dice distance formula)	
COUNT		3	4	1	6	$=(2 \times 1) / (3 + 4)$ $= 0.286$	$= 5 / (6 + 1)$ $= 0.714$	

2.5. Gene expression profile similarity:

Pearson Correlation Coefficient (r) was used to calculate the expression profile similarity of two or more genes with their expression levels in several different tissues (Liao and Zhang, 2006). For our analyses, we obtained the

RNA-seq gene expression data of human from two databases- **(1) Human protein atlas** Release 9 (<http://www.proteinatlas.org/>) and **(2) EMBL-EBI Expression Atlas** (<http://www.ebi.ac.uk/gxa>). Both the databases provide experimental RNA-seq gene expression data in human tissues. The Pearson correlation coefficient used to determine the expression profile similarity within the duplicated pairs –

$$\text{Pearson correlation coefficient } (r) = \frac{N \sum ij - (\sum i)(\sum j)}{\sqrt{[N \sum i^2 - (\sum i)^2][N \sum j^2 - (\sum j)^2]}}$$

Where ‘*i*’ and ‘*j*’ are paralogous pairs, ‘*N*’ represents the total number of tissues, ‘ $\sum ij$ ’ is the sum of the products of paired expression signal intensities, ‘ $\sum i$ ’ sum of expression signal intensities for gene ‘*i*’, ‘ $\sum j$ ’ is the sum of expression signal intensities for gene ‘*j*’, ‘ $\sum i^2$ ’ is sum of squared expression signal intensities of gene ‘*i*’, ‘ $\sum j^2$ ’ is sum of squared expression signal intensities of gene ‘*j*’.

2.6. Micro-RNA target sites:

Micro-RNAs, as their name suggests, are small RNA molecules that do not encode any protein. Instead, they regulate the gene expression by interfering at the post-transcriptional level. The micro-RNAs are found mainly in plants, animals, and some viruses, and are transcribed from the micro-RNA gene, by the help of RNA polymerase II. The primary transcript of micro-RNA genes undergoes several post-transcriptional modifications inside nucleus as well as after its export to the cytosol via RAN-GTP mediated protein exportin, by the nuclear and cytoplasmic exonucleases Drosha and Dicer, respectively.

A mature micro-RNA is single-stranded ~22 nucleotide molecule that regulates the gene expression at the posttranscriptional level either by repressing translation or by degrading its target mRNA. This phenomenon

requires complementary base-pairing of the messenger RNA and the seed region of the microRNA.

For our study, we obtained the micro-RNA target sites of human and mouse genes from the TargetScan database Release 6.2 (<http://www.targetscan.org>) (Garcia *et al.* 2011) that predicts biological target sites of micro-RNAs present in the mRNAs of a gene, by searching for the presence of conserved six-, seven- and eight-mer sites in the mRNA matched to the seed region of micro-RNA.

2.7. Protein Multifunctionality:

Proteins perform almost all the cellular functions within an organism. However, they do not function alone. Instead, in most of the cases proteins perform their functions by interacting with other proteins, known as their interacting partners. Thus, a certain cellular function usually requires many proteins. In other words, besides being assigned to perform highly specialized functions, proteins usually perform many other functions and become ‘multitasking’. Protein multifunctionality indicates the number of functions to which a protein is associated. As the functions of a protein are mediated by its domains, the number of domains in the structure of a protein can be used as a proxy to measure the number of its functions. Additionally, protein multifunctionality can be measured directly from its functional annotation. This can be done by using the Gene Ontology (GO) terms assigned to the genes and/or proteins. For example, using the unique GO terms for the GO domain ‘biological function’ for a protein will reveal the unique biological processes in which the protein takes part. For our study, we obtained protein domains from Pfam (Finn *et al.* 2014) and Unique GO Biological Process terms from the Ensembl Genome Browser (Flicek *et al.*

2014; Yates *et al.* 2016). We calculated protein multifunctionality by using both the above-mentioned parameter.

2.8. Disease genes:

As we have seen earlier in Chapter 1, mutations on a gene may lead to changes in gene function. Such changes are often lethal to the organism and eliminated via purifying selection. However, there are consequences where a mutation on a gene does not lead to lethality; instead it may lead to disease formation. During the past few decades, researchers had attempted to identify such human genes which may cause disease progression upon mutations on them. For our analyses, we obtained the human disease genes from the Human Gene Mutation Database (HGMD) (Stenson *et al.* 2012), which contains both the Monogenic (Mendelian) and Polygenic (Complex) disease genes and represent a comprehensive collection of germline mutations in unclear genes associated with human inherited disease.

2.9. Developmental genes:

Genes associated with the embryonic or post-embryonic development of an organism are known as developmental genes. These genes are usually evolutionarily highly conserved in nature, shows variable expression patterns in different stages of development, and their misregulation leads to developmental defects. Examples include the homeobox (Hox) genes that govern the body-plan of the embryo. In our study, we obtained the developmental and non-developmental genes by using the gene-ontology annotation for the GO domain biological process. Here, a gene is considered 'developmental' if they are associated with one of the two GO terms: GO:0007275 (multicellular organismal development) and GO:0030154 (cell

differentiation) or their daughter terms, and the other genes were considered 'non-developmental' (Makino, Hokamp, McLysaght 2009).

Chapter 3

The complex association of gene duplication and gene essentiality: insights from human and mouse genome

Gene duplication is among one of the major driving forces shaping genome and organism evolution. Duplication creates multiple copies of a gene and is influenced by intrinsic properties of the gene. One such property is gene essentiality, depicting the functional importance of a gene and is measured by the fitness cost of the gene upon its deletion. Comparison of the fraction of essential genes among mouse and human revealed that the essential genes avoid duplication in mouse but not in humans. This study explores the reasons behind such discrimination in gene essentiality in the context of gene duplication by cross-species comparison of mouse and human genomes. In-depth functional analyses suggests that the essential human duplicated genes are functionally more redundant than that in mouse. The paralogs of mouse essential duplicates are more often pseudogenized than that of humans. Additionally, such functionally redundant duplicates are under stringent dosage regulation in humans. We also observed a higher evolutionary conservation in the paralogs of human essential genes than that in mouse. Together, our results demonstrate that the human essential genes are retained as duplicates to serve as backup copies that are under stringent dosage regulation and may shield themselves from loss-of-function mutations.

Keywords: Gene essentiality, Gene duplication, Functional divergence, Evolutionary conservation, dosage imbalance.

Adapted from Acharya et al., 2015, PLoS ONE 10(3): e0120784

3.1. Introduction

Gene duplication is one of the key factor regulating genome and organism evolution (Stephens 1951; Ohno, Wolf, Atkin 1968; Ohno 1970; Zhang 2003). Gene duplication supply raw genetic materials for structural and functional innovation and also conserves the parental function. Although duplication is not always advantageous, as after duplication most of the gene copies subsequently become nonfunctionalized or pseudogenized in the genome (Ohno 1970), it has many implications in the life of the organism. For example, the duplicates may be preserved in the genome for their immediate benefit like requirement of an increased gene dosage (Innan, Kondrashov 2010) or may serve as backup copies to restore the function of the parent copy upon the accumulation of loss-of-function mutation on the latter (Gu *et al.* 2003; Liang, Li 2009). The duplicated copy(s) may also undergo structural and functional modifications to take up new functions, an event known as neofunctionalization (Ohno 1970), or they may share their ancestral functions upon accumulating complementary degenerative mutations, a phenomenon known as subfunctionalization (Clark 1994; Lynch, Force 2000). The pattern of gene duplication varies between species as well as across different gene groups within the same species. Several factors contributing gene duplication have been observed till date in different organisms. Examples include protein connectivity and protein interaction network (Makino, Suzuki, Gojobori 2006; Liang, Li 2007; D'Antonio, Ciccarelli 2011), protein complexity (Yang, Lusk, Li 2003; Bhattacharya, Ghosh 2010), gene retention and sequence divergence (Waterhouse, Zdobnov, Kriventseva 2011), gene dosage balance (Makino, McLysaght 2010) and nevertheless, gene essentiality (He, Zhang 2006b; Liao, Zhang 2007; Makino, Hokamp, McLysaght 2009).

Essential genes are indispensable to an organism, and their deletion causes severe fitness reduction like sterility or lethality (Liao, Scott, Zhang 2006). Such genes are mostly associated with important biological or molecular functions. However, many genes performing such 'essential' functions are considered to be nonessential, as some other genes with similar or identical functions and expression compensate their deletion (Chen *et al.* 2012c). Gene duplication is an important source of such functional redundancy (Ohno 1970). Now, there are two possibilities for essential genes to prefer or avoid its duplication. First, duplication of essential genes provides backup copies that could shield themselves from harmful mutations. Secondly, from the evolutionary perspective, essential genes may avoid duplication since it relaxes the effect of purifying selection on gene copies that may increase the probability of accumulation of mutations in these duplicates. Such mutations are not acceptable for the essential genes since they are among the most vital and conserved gene-group within the genome (Jordan *et al.* 2002; Yang, Gu, Li 2003).

The accumulation of gene knockout and knockdown fitness data in model organisms lead to the identification and characterization of essential genes in different organisms. Such studies have depicted a complex relationship of gene essentiality with gene duplication (Makino, Hokamp, McLysaght 2009). It has been shown that the lower unicellular eukaryotes like yeast possess a higher fraction of essential genes as singletons (single-copy) than the duplicates (Gu *et al.* 2003). However, works in mammalian mouse model revealed an equal proportion of essential genes in singletons and duplicates (Liang, Li 2007; Liao, Zhang 2007). Arguably, more recent studies with mouse indicate that among the two gene groups, the fraction of essential genes is significantly higher in singletons (Su, Gu 2008; Chen *et al.* 2012c). However,

all studies regarding gene essentiality have been carried out in yeast and mouse, largely due to unavailability of human gene essentiality data. A previous study explored the properties of human orthologs of essential mouse genes, considering themselves as ‘human essential genes’ (Georgi, Voight, Bucan 2013). However, such estimation may not be accurate (Liao, Zhang 2008). Fitness data from gene knockdown experiments in human cell lines lead to the identification of human essential genes (Silva *et al.* 2008). Such experimentally validated and literature-curated human gene essentiality data were accumulated in Online Gene Essentiality (OGEE) database, which represents a valuable resource of essential genes in a large number of prokaryotic and eukaryotic organisms. In this study, we present a comprehensive analysis of essential human genes and their duplication pattern, by a genome-wide comparison of human and mouse. Our study suggests that mouse essential genes do not prefer duplication whereas human essential genes do. Such a trend is unexplored so far. To get a detailed insight into such observation, we tried to investigate underlying reasons of maintaining essential genes as duplicates in humans.

3.2. Materials and Methods:

3.2.1. Gene Essentiality and Gene Duplication:

We obtained the gene essentiality and gene duplication status of human (*Homo sapiens*) and mouse (*Mus musculus*) from the Online Gene Essentiality (OGEE) database (Chen *et al.* 2012b) (<http://ogeedb.embl.de>). For the duplicated genes, we obtained the duplicated pairs for mouse and human essential genes from authors of OGEE database (Chen *et al.* 2012b).

3.2.2. Developmental Genes:

We obtained the developmental genes of mouse and human from Online Gene Essentiality (OGEE) database (Chen *et al.* 2012b). In this database, developmental genes are determined by their association with one of the two GO terms: GO:0007275 (representing multicellular organism development) and GO:0030154 (representing cell differentiation) or their daughter terms. Genes that are not associated with any of the two terms or their daughter terms are considered non-developmental, based on Makino's classification of developmental genes (Makino, Hokamp, McLysaght 2009).

3.2.3. Phyletic Age and Overall Proportion of Essentiality :

Every gene has its own phyletic origin defined as the most distance group of organisms where the homologs (orthologs) of that gene exist. In the Online Gene Essentiality (OGEE) database (Chen *et al.* 2012b), the authors used the phyletic age prediction algorithm of Wolf *et al.* (Wolf *et al.* 2009) and divided all genes within a genome into seven classes according to their phyletic origin- 0 (not assigned), 1 (Mammalia), 2 (Chordata), 3 (Metazoa), 4 (Fungi/Metazoa), 5 (Eukaryote) and 6 (cellular organisms). For our analysis, we obtained the phyletic age mouse and human genes from OGEE database. We discarded the first group having unassigned phyletic age and selected the rest from mouse and human genomes. Our final data contains gene essentiality, gene duplication and phyletic age information of 5869 mouse genes and 18400 human genes, respectively. We divided these human and mouse genes into two groups depending on their phyletic age: the 'old duplicates' (containing the older three classes) and 'new duplicates' (containing the rest). From this, we computed the overall proportion of essential genes in singletons and duplicates for both species irrespective of their age-bias, as a weighted average using this formula (Chen *et al.* 2012c):

$$P_E = f_{old} \times P_E^{old} + f_{young} \times P_E^{young}$$

Here, f_{old} and f_{young} represent the fraction of old and young genes, respectively, within the gene group and P_E^{old} and P_E^{young} represents the proportion of essential genes in old and young classes, respectively. We used this formula to calculate the proportion of essential genes irrespective of their age bias in singleton and duplicates for both mouse and human.

3.2.4. Pseudogenization:

We obtained the mouse and human gene biotypes from ensemble biomart 71 (<http://www.ensembl.org/biomart/martview>) (Flicek *et al.* 2013). The genes annotated as pseudogene in their gene biotype were considered as pseudogenes. This comprise the following classes of pseudogenes: pseudogene, IG-C-pseudogene, IG-J-pseudogene, IG-V-pseudogene, TR-V-pseudogene, TR-J-pseudogene, polymorphic pseudogene and processed pseudogene. We computed the proportion of paralog pseudogenization by considering the duplicated essential mouse and human genes having at least one pseudogenized paralog.

$$Proportion\ of\ Paralog\ Pseudogenization = \frac{Number\ of\ Pseudogenized\ Paralogs}{Total\ Number\ of\ Paralogs}$$

3.2.5. Functional Distance:

We used the Gene Ontology (GO) annotations to compute the functional distance for mouse and human essential genes. We obtained the GO domain molecular function for essential genes and their paralogous copies for both species from the biomart interface of Ensembl Genome Browser (version-71) (<http://www.ensembl.org/biomart/martview>) (Flicek *et al.* 2013). For the estimation of functional divergence between mouse and human essential genes, we used the Czekanowski-Dice distance formula (Baudot, Jacq, Brun

2004) mentioned in Resources and Methodology section (Chapter 2, Section 2.4). Finally, we obtained the functional divergence for each human and mouse essential genes with their paralogous counterparts.

Although the Czekanowski-Dice formula is the most popularly used method for calculating functional distance, it is very sensitive to the count of associated GO terms per gene, which may vary between species. Thus, the estimation of functional distance can be erroneous for cross-species comparison, unless normalized by the number of GO terms associated with the essential genes of both species. To ensure that, we binned our functional distance data of the two species in three groups: Bin 1 (GO term count 1 to 4; $N_{\text{human}} = 367$, $N_{\text{mouse}} = 773$), Bin 2 (GO term count 5 to 8; $N_{\text{human}} = 343$, $N_{\text{mouse}} = 485$) and Bin 3 (GO term count > 8; $N_{\text{human}} = 244$, $N_{\text{mouse}} = 278$). We compared the functional distance of mouse and human essential genes within each bin.

3.2.6. Micro-RNA Target Sites:

We obtained the micro-RNA target sites for mouse and human genes from TargetScan Release 6.2 (<http://www.targetscan.org>) (Garcia *et al.* 2011). For our analysis, we pooled together all known paralogs of each essential gene individually, thus making various sets containing paralogs for each essential genes in both species. We calculated the mean micro-RNA target sites of these sets for the two species to acquire the micro-RNA target sites of essential genes' paralogs for each species. We considered the average value of all sets within a species for cross-species comparison.

3.2.7. Evolutionary Rate:

We obtained the nonsynonymous nucleotide substitution per nonsynonymous sites (dN) and synonymous nucleotide substitution per synonymous sites (dS) from the ensemble biomart (version 71)

(<http://www.ensembl.org/biomart/martview>), using one-to-one rat (*Rattus norvegicus*) orthologs. The evolutionary rates of human and mouse essential genes were calculated by the ratio of dN and dS (Flicek *et al.* 2013).

3.2.8. Statistical Analyses:

We performed the statistical analyses of the entire work using SPSS version 13 and in-house PERL script. We compared the mean values of different variables between two classes of genes using Mann-Whitney U test. In-house PERL script was used to perform two-sample Z-test to compare relative proportions of a variable between two groups of genes.

3.3. Results and Discussions:

3.3.1. Gene essentiality and gene duplication in human and mouse:

We examined the gene essentiality and gene duplication data of human and mouse and noticed that the proportion of essential genes among duplicated genes vary between human and mouse. Our observations show that in mouse, the proportion of essential genes is significantly higher in singletons, as 994 genes are essential among 2098 singleton genes (47.38%) and 1563 genes are essential among 3771 duplicated genes (41.45%) [$Z = 4.391$, confidence level 99%; $P < 0.0001$, two sample Z-test]. Whereas, in humans, the proportion of essential genes is higher among the duplicates, as 486 genes exist as essential (6.43%) among 7563 singleton genes, and 984 are essential among 10837 duplicated genes (9.08%) [$Z = -6.523$, confidence level 99%; $P < 0.0001$, two sample Z-test]. The observations clearly suggest an overall higher proportion of essential genes in mouse, largely due to the efficacy of the methods used to identify essential genes (Chen *et al.* 2012b) or the lack of a complete gene essentiality data. However, within the same species (where the same method is used for detection of essential genes), gene

essentiality should contribute equally, which is not the case here, as the comparison revealed a higher possibility of retaining essential genes as duplicates in humans, but not in mouse.

In a previous study, Makino *et al.* showed that genes involved in development are more likely to be essential than the non-developmental genes (Makino, Hokamp, McLysaght 2009) and hence, their abundance in a particular gene group may result in higher essentiality for that group. Thus, to find whether the overrepresentation of developmental genes affects our observations, we discarded the developmental genes and computed the proportion of essential genes in human and mouse, considering only the nondevelopmental genes (see Section 3.2.2). Our results revealed a similar trend (Table 3.1), suggesting that our results are not influenced by enrichment in developmental genes. Thus, we carried forward our analyses including both developmental and nondevelopmental mouse and human genes.

Another possible data bias in our analysis may arise from the differential age of the duplicates. Previous studies revealed that the older genes are more essential than younger genes (Chen *et al.* 2012c) and genes derived from older duplication events are more essential than singletons (Su, Gu 2008). Therefore, the gene age may influence gene essentiality, leading to a biased estimation of gene essentiality in our dataset. This age-bias was corrected by incorporating the phyletic age of the genes to calculate the overall proportion of essentiality (Chen *et al.* 2012c) (see materials and methods) in singleton and duplicated mouse and human genes. We have not considered the duplication age (the origin of most recent duplication event) as our dataset contains both singletons and duplicates, hence, phyletic age is a more suitable measure. After correcting the age bias, we still obtained similar

Table 3.1. Proportion of essential genes among singleton and duplicates of mouse and human non-developmental genes.

Species	Gene group	Total genes	Essential genes	Proportion of essential genes	Z-score and P value
Mouse (<i>Mus musculus</i>)	Singleton	1237	462	37.348	Z= 5.0323 (Confidence level 99%) P <0.0001
	Duplicates	2301	669	29.074	
Human (<i>Homo sapiens</i>)	Singleton	6347	332	5.231	Z= -3.7168 (Confidence level 99%) P= 0.0002
	Duplicates	8581	575	6.701	

trend in the proportion of essential genes in singletons and duplicates in both mouse and human (Table 3.2).

Our results contradicted the previous study of Liao and Zhang (Liao, Zhang 2007), which revealed an equal proportion of essential genes among mouse

Table 3.2. Proportion of essential genes as weighted average among singleton and duplicates of mouse and human.

Species	Gene group	Total genes	Proportion of essential genes as weighted average $P_E = f_{old} \times P_E^{old} + f_{young} \times P_E^{young}$	Z-score and P value
Mouse (<i>Mus musculus</i>)	Singleton	2098	47.379	Z= -4.392 (Confidence level 99%) P <0.0001
	Duplicate	3771	41.448	
Human (<i>Homo sapiens</i>)	Singleton	7563	6.426	Z= -6.535 (Confidence level 99%) P <0.0001
	Duplicate	10837	9.081	

singleton and duplicate genes. The reason behind such contradiction may be the difference in essential gene collection procedure followed in the older dataset (Mouse Genome Informatics or MGI) which they used and the newer dataset (Online Gene Essentiality or OGEE database) which we have used. However, our observation of a higher proportion of essential genes in mouse

singletons is consistent with two more recent studies (Su, Gu 2008; Chen *et al.* 2012c). Thus, in this study, we explored why essential genes prefer duplication in humans, but not in mouse. Here, we have done a detailed analysis using human and mouse essential genes. As we know, duplication leads to subsequent increase in gene dosage in the protein-protein interaction network, which is not acceptable for essential genes as they are highly connected (hub-like) in protein-protein interaction network (Jeong *et al.* 2001; Barabasi, Oltvai 2004; He, Zhang 2006a; Goh *et al.* 2007). Such a duplication leading to the dosage imbalance may not be favorable and the duplicates must either be diversified (Li, Yang, Gu 2005) or maintained under stringent dosage regulation (Makino, McLysaght 2010).

To investigate whether the functional divergence of the essential duplicates supports their fixation in the human genome, or they are maintained as backup copies under stringent dosage-regulation, we did a cross-species comparison of the essential genes and their paralogs in mouse and human genomes.

3.3.2. The proportions of paralog pseudogenization in human and mouse essential genes:

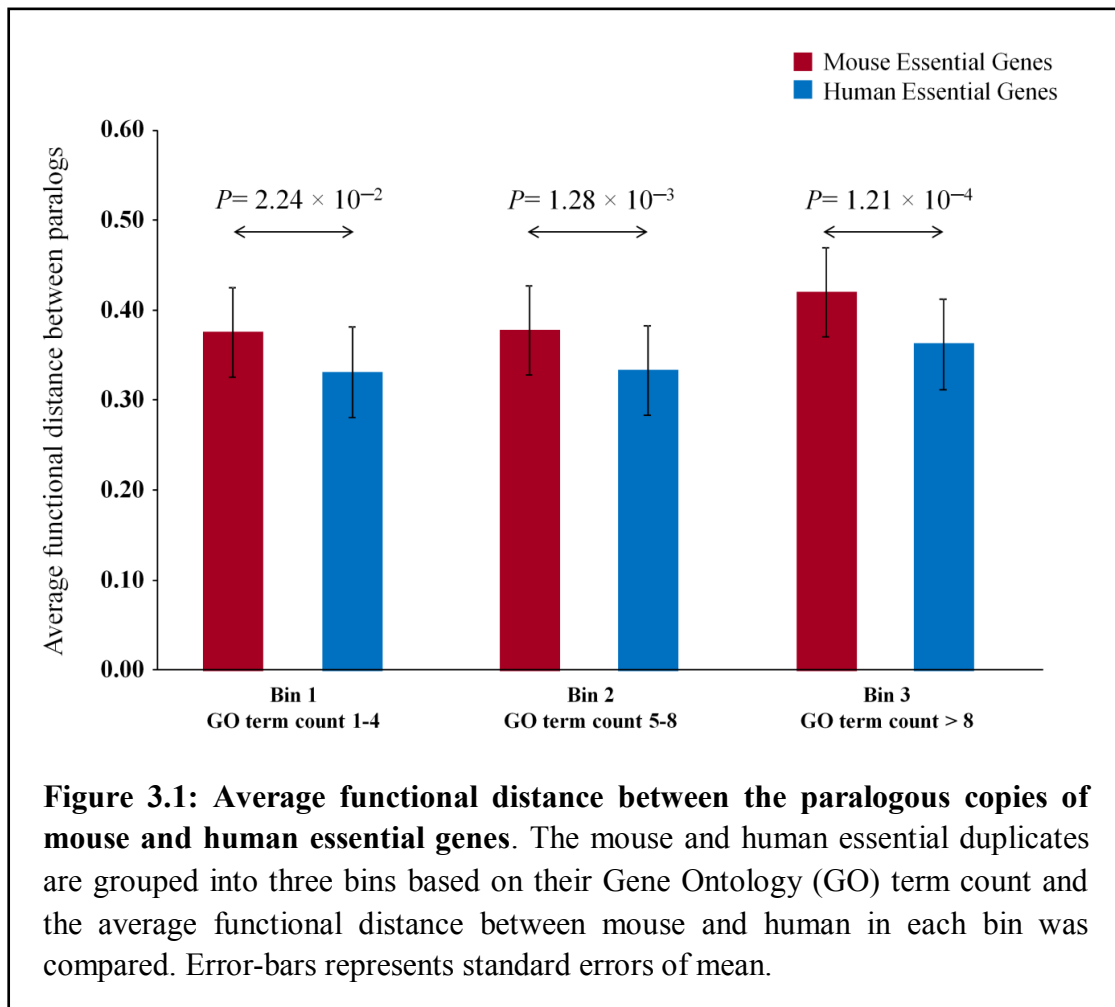
As gene duplication often generates ‘useless’ duplicates that become pseudogenized within the genome, we studied the occurrence of pseudogenized paralogs among human and mouse essential genes. As we are dealing with essential genes of the two species, no occurrence of pseudogene was observed in our dataset. However, a very small proportion of the paralogs of essential genes were found to remain as pseudogenes, with no significant difference between human (0.50%) and mouse (0.82%) ($Z = -1.584$, $P = 1.13 \times 10^{-1}$, two sample Z-test). Such low proportions of pseudogenes in

both the species is normal, as we are dealing with the paralogs of genes having crucial functions. Furthermore, we considered the proportion of paralog pseudogenization for each essential genes having at least one pseudogenized paralog (see Section 3.2.4). We obtained a lower proportion of paralog pseudogenization in human essential genes than their mouse counterparts (Proportion of paralog pseudogenization in human = 0.048, $N_{\text{human}} = 63$, Proportion of paralog pseudogenization in mouse = 0.178, $N_{\text{mouse}} = 17$; $P = 1.44 \times 10^{-7}$, Mann-Whitney U test). This result suggests that a very low fraction of paralogous copies of essential genes are nonfunctionalized (pseudogenized) and the paralogs of mouse essential genes become pseudogenized more easily than humans.

3.3.3. Functional distance of human and mouse essential genes:

As most of the paralogs of essential genes in both species are 'functional', We explored whether the human essential duplicates are functionally diversified to become fixed within the genome, we used the gene ontology (GO) annotations for human and mouse essential genes and their paralogous copies from biomart interface of Ensembl 71 (Flicek *et al.* 2013) for the GO domain Molecular function. The Czekanowski-Dice distance formula (see section 2.4) was used to compute functional distance of human and mouse essential genes (Baudot, Jacq, Brun 2004). We obtained a significantly lower functional distance in human essential duplicates (Average functional distance=0.340, N=954) than the mouse essential duplicates (Average functional distance=0.385, N=1536) ($P=3.73 \times 10^{-6}$, Mann-Whitney U test). However, the Czekanowski-Dice distance formula is sensitive to the number

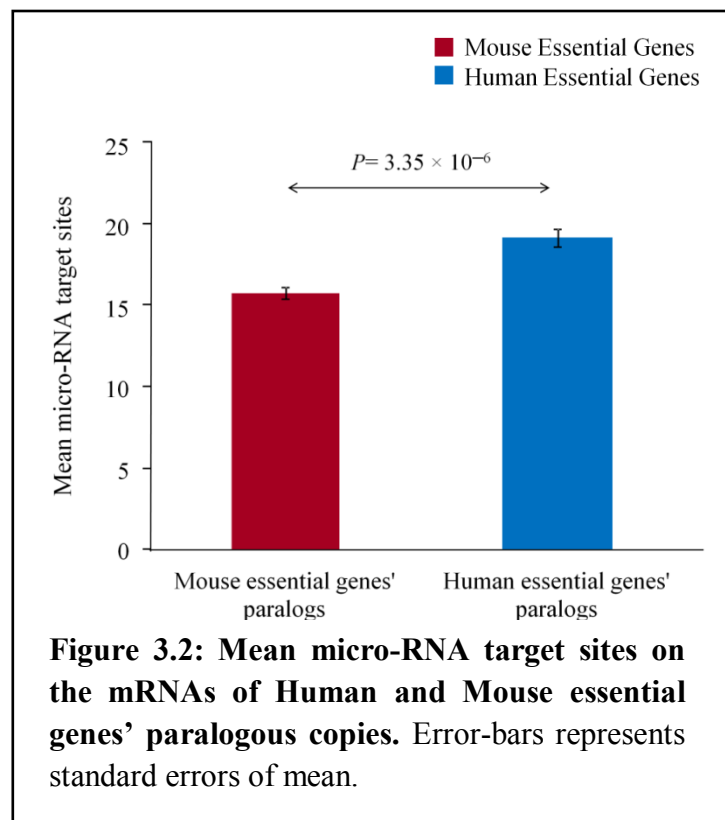
of gene ontology terms, which is species-specific and may vary between



human and mouse. Thus, an unbiased cross-species comparison of functional distance was performed by grouping the dataset into three bins according to their GO term count (see Section 3.2.5). We obtained a significantly lower functional distance in human essential duplicates than their mouse counterparts in all three bins (Figure 3.1), which suggests that the duplicated copies of human essential genes are functionally similar and have the potential to serve as backup copies.

3.3.4. Mean micro-RNA target sites in the paralogs of human and mouse essential genes:

However, the maintenance of these functionally redundant duplicates is very crucial, as it often leads to dosage imbalance in protein-protein interaction network. Thus, the stable maintenance of the duplicates require efficient dosage-regulatory mechanism, like the micro-RNA mediated regulation of gene expression acting at the post-transcriptional level, which maintain the backed up essential genes by reducing their expression (Li, Musso, Zhang 2008). We measured the average micro-RNA target sites of the longest mRNAs of paralogous copies of human and mouse essential genes to understand the ability to maintain the backed up duplicates in both species (see Section 3.2.6 for details). Comparing the average micro-RNA target sites between the two species, we observed a significantly higher micro-RNA target sites in



the mRNAs duplicated essential genes of human (Mean micro-RNA count 19.15, Number of sets=742) than in mouse (Mean micro-RNA count 15.82, Number of sets=1202) (Figure 3.2) ($P= 3.35 \times 10^{-6}$; Mann-Whitney U test). This suggests that the paralogous copies of human essential duplicates are under more robust regulation by micro-RNAs, which enables humans to maintain the redundant duplicates.

Therefore our study clearly suggests that the duplicated copies of human essential genes remain mostly functionally redundant and are maintained as backup copies, having the ability to escape the dosage imbalance.

3.3.5. The evolutionary rates of paralogs of human and mouse essential genes:

As essential duplicates of human are functionally more redundant and are under more stringently dosage regulated than mouse, their paralogs should be evolutionarily more conserved to serve as backup copies upon gene deletion. Comparing the evolutionary rates between paralogs of human and mouse essential in terms of the ratio of dN (Nonsynonymous substitution rates per nonsynonymous sites) and dS (Synonymous substitution rates per synonymous sites) (see Section 3.2.7), we obtained a significantly slower evolutionary rate in the paralogs of human essential genes ($dN/dS_{\text{human}} = 0.101$, $dN/dS_{\text{mouse}} = 0.128$, $P = 2.53 \times 10^{-5}$, Mann Whitney U test, $N_{\text{mouse}} = 2931$, $N_{\text{human}} = 1651$), as indicated by their lower dN/dS ratio. This suggests that the redundant paralogous copies of human essential duplicate genes possess higher evolutionary conservation, and therefore may serve as backup copies, increasing the robustness against gene deletion fitness.

3.4. Conclusion:

Gene duplication is a genetic mutation that generates multiple redundant copies of a gene. The retention of these redundant gene copies demands functional diversification or functional redundancy with regulated protein dosage. In this study, we observed that human duplicated genes have a higher proportion of essential genes, a trend which is different from mouse. We showed that the duplicated copies of human essential genes are functionally more redundant. These copies are also evolutionarily more

conserved than that in mouse. We also demonstrated that these functionally redundant gene copies could be maintained by a more efficient dosage-regulation in humans. This study sheds light on the importance of human duplicated essential genes that reduce the fitness effect of gene deletion, thereby increasing the robustness against deleterious mutations in humans.

Chapter 4

The importance of whole-genome duplication in human genome evolution

Gene duplication provides raw genetic materials required for structural and functional innovations that are essential elements for genome and organism evolution. The duplication of one or a few genes (Small scale duplication or SSD) is required for such functional innovations. However, major evolutionary transitions may require a vast number of new raw genetic materials, that are yielded by processes like the duplication of the whole genome (whole genome duplication or WGD) to generate new functions beneficial for such transitions. More recent studies with gene duplication consider these two group of duplicates separately, as studies with yeast revealed plenty of differences between the two classes of duplicates: Yeast WGD pairs were functionally more similar, more similar in subcellular localization and are depleted in essential genes. The two rounds of whole genome duplication occurring early in vertebrate evolution is the root of the whole-genome duplicates that are today present in vertebrates like fishes, amphibians, reptiles, birds and mammals. Here, we explored the evolutionary genomic attributes of human SSD and WGD genes, in a comparative analysis involving these two classes of duplicates in human, to investigate the contribution of whole-genome duplication in human evolution.

Keywords: Small-scale duplication, Whole-genome duplication, Functional divergence, Evolutionary rate, Protein multifunctionality, Gene essentiality, Disease genes.

Adapted from Acharya and Ghosh, 2016, BMC Genomics 17:71

4.1. Introduction:

Gene duplication generates new gene copies from pre-existing ones and is an important source of genome evolution (Stephens 1951; Ohno, Wolf, Atkin 1968; Ohno 1970). Initially after duplication, the newly-formed gene copies remain functionally redundant, leading to a relaxation of purifying selection on both the gene copies, often resulting in the adoption of new functions (Ohno 1970; Clark 1994; Teshima, Innan 2008; Innan, Kondrashov 2010).

Therefore, gene duplication is an essential process guiding organism evolution as it provides raw genetic elements for genome evolution (Ohno, Wolf, Atkin 1968; Ohno 1970; Taylor, Raes 2004). However, the retention of duplicated gene copies is not an easy process, as most of the duplicates nonfunctionalize and/or lost from the genome following duplication (Ohno 1970), whereas some of the duplicates are retained in the genome in the course of evolution. Such retention of duplicates may prove advantageous, as the newly formed redundant duplicates serve as backup copies providing functional compensation after gene deletion (Liang, Li 2009), thus providing increased genetic robustness against harmful and deleterious mutations (Gu *et al.* 2003). However, gene duplication leads to a change in gene-dosage in protein-protein interaction network and thus, the retention of duplicated copies require favorable circumstances like increased gene dosage advantage, where the increment in the gene product after duplication turn out to be advantageous to the organism (Kondrashov, Kondrashov 2006; Innan, Kondrashov 2010) or a stringent regulation in gene dosage after duplication (Li, Musso, Zhang 2008; Chang, Liao 2012) or regulation of the expression patterns of duplicated copies (Li, Yang, Gu 2005; Ganko, Meyers, Vision 2007;

Li *et al.* 2009; Qian *et al.* 2010). Additionally, after duplication, the duplicated copies may be diversified by adapting new functions or expression patterns (neofunctionalization) (Ohno 1970), or by sharing the ancestral function or expression after accumulating complementary degenerative mutations (subfunctionalization) (Force *et al.* 1999; Lynch, Force 2000). They may also diverge at the subcellular localization, where the proteins encoded by the duplicated pairs localize into different cellular compartments (Marques *et al.* 2008).

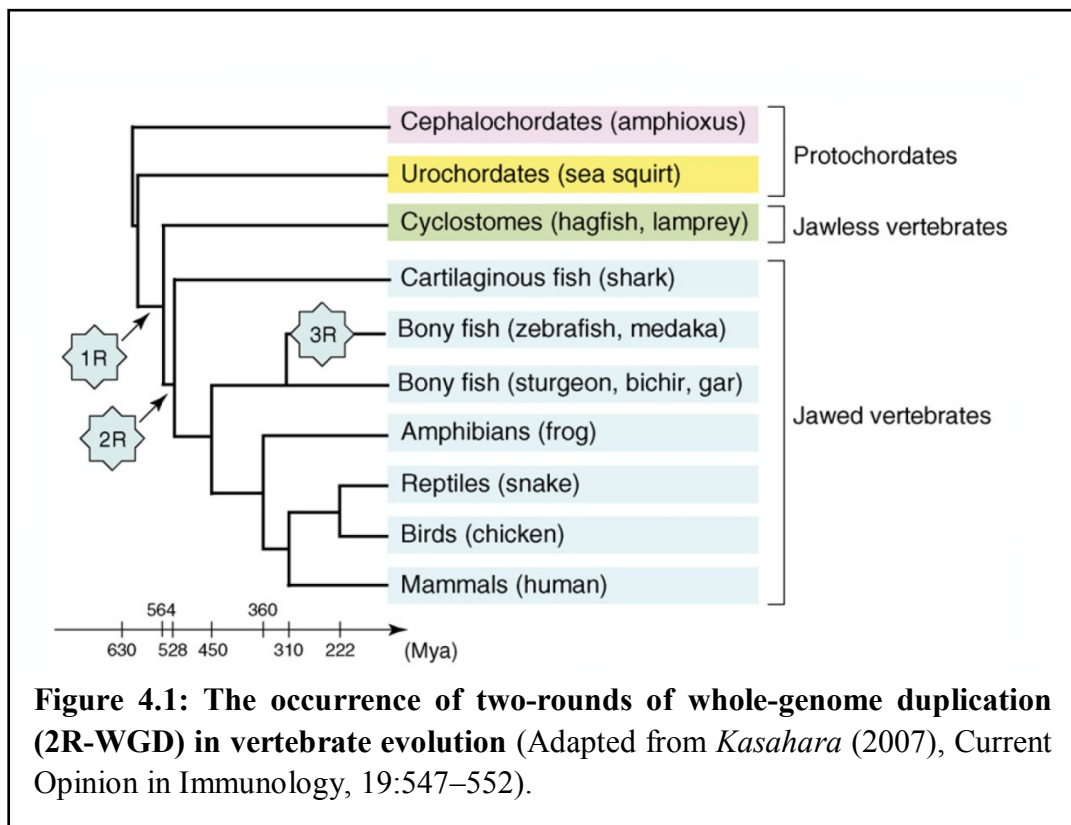
Therefore, it is quite clear that gene duplication leads to a modification in the protein-protein interaction network. However, there are subtle differences in the extent of gene duplication. Usually, duplication involves a single gene [known as small-scale duplication or SSD], whereas, duplications at a larger scale may comprise many genes, chromosomal segments. Duplication may even involve the whole genome, a phenomenon known as the whole-genome duplication (WGD) (Hakes *et al.* 2007). These two types of gene duplication also vary regarding their occurrence. While small-scale duplication may occur at any moment and may be retained in the course of evolution, whole genome duplication events are much rarer within eukaryotes, being most frequent and broadly studied in plant genome evolution (Stebbins 1971; Blanc *et al.* 2000; Wendel 2000; Adams, Wendel 2005b).

However, evidences of an ancient WGD in the yeast genome evolution (~150 Mya) (Wolfe, Shields 1997; Dujon *et al.* 2004; Kellis, Birren, Lander 2004) and two-rounds of whole-genome duplication (2R-WGD) in the early vertebrate evolution (~530 Mya) (Allendorf, Thorgaard 1984; Zhou, Cheng, Tiersch 2001; McLysaght, Hokamp, Wolfe 2002; Dehal, Boore 2005; Brunet *et al.* 2006b;

Nakatani *et al.* 2007) were also confirmed in earlier studies. Previous studies also claim that the 2R-WGD provides the raw materials for increasing genome and organism complexity and extensive species diversity (Zhou, Cheng, Tiersch 2001; Dehal, Boore 2005), being an important process in vertebrate evolution (Allendorf, Thorgaard 1984; Dehal, Boore 2005).

Moreover, the functions of a gene are mediated by the proteins encoded by them, which in turn function by interacting with other such proteins, thereby forming a protein-protein interaction network (PPIN) (Chakraborty, Ghosh 2013). Thus, the functional precision of any gene not only depends on itself but also on its interacting partners. The stoichiometric balance of the proteins within a PPIN, therefore, has an important contribution on gene function. The retention of duplicated genes creates a stoichiometric disparity in the PPIN, as the duplicated genes produces more proteins than the non-duplicated proteins within the same PPIN (Papp, Pal, Hurst 2003; He, Zhang 2006b; Birchler, Veitia 2007). Hence, the contribution of SSDs and WGDs to the stoichiometric balance of their associated PPIN are different (Lynch, Conery 2000; Freeling, Thomas 2006; Hakes *et al.* 2007; Makino, McLysaght 2010; Fares *et al.* 2013). Following WGD, the whole PPIN duplicate simultaneously, maintaining the stoichiometric balance of PPIN; but after SSD, the duplicated gene forms more protein relative to its non-duplicated interacting partners, thereby creating a protein dosage imbalance in the PPIN. Therefore, from the above perspective, whole-genome duplicates are thought to be preserved intact within the genome, keeping the gene dosage of the PPIN intact (Makino, McLysaght 2010). Thus, deviating from the conventional comparison between the evolutionary genomic properties of singletons and duplicates (Robinson-Rechavi, Laudet 2001; Gu *et al.* 2003; Jordan, Wolf, Koonin 2004), a better resolution can be achieved by a comparison between

duplicates originating from SSD and WGD events. With the availability of completely sequenced genome for many yeast species, researchers identified the duplicates originated from these two types of duplication (Kellis, Birren, Lander 2004). The comparison of these two distinct duplicate groups in yeast revealed noticeable differences (Guan, Dunham, Troyanskaya 2007; Hakes *et al.* 2007; Fares *et al.* 2013). The yeast duplicate pairs originating from WGD are functionally more similar than those originating from SSD, irrespective of their sequence similarity (Guan, Dunham, Troyanskaya 2007; Hakes *et al.* 2007). Additionally, yeast WGD-pairs are more often colocalized in the same cellular compartment (Fares *et al.* 2013). Also, genes undergoing small-scale duplication in yeast contain a higher proportion of essential genes than WGDs (Guan, Dunham, Troyanskaya 2007; Hakes *et al.* 2007).



The occurrence of 2R-WGD early in the vertebrate evolution (Zhou, Cheng, Tiersch 2001; McLysaght, Hokamp, Wolfe 2002; Dehal, Boore 2005; Brunet *et*

al. 2006b; Kasahara 2007; Singh, Arora, Isambert 2015b), and the subsequent detection of the existing whole-genome duplicates within human genome (McLysaght, Hokamp, Wolfe 2002; Makino, McLysaght 2010; Singh, Arora, Isambert 2015b) lead us to compare the evolutionary genomic attributes of human small-scale and whole-genome duplicates (SSDs and WGDs, respectively). As stated previously, the human WGDs are much more older than yeast in terms of their origin (~530 Mya vs ~150 Mya, respectively), we hypothesized that these human duplicates were subjected to more evolutionary pressure than yeast due to their longer evolutionary exposure.

Thus, our study will explore the significance and the long-term evolutionary fate of human whole-genome duplicates with those duplicates originating spontaneously at small-scale.

4.2. Materials and methods:

4.2.1. Classification of human duplicated genes:

We obtained the human protein-coding genes (N=22447) from the Ensembl biomart (version 77) (Flicek *et al.* 2014) (<http://www.ensembl.org/biomart/martview>). Human whole-genome duplicates (WGDs) were collected from two datasets: 1. The supplementary dataset from Makino and McLysaght (Makino, McLysaght 2010) and 2. OHNOLOGS database (<http://ohnologs.curie.fr/>) (Singh, Arora, Isambert 2015b) using the strict dataset [q-score (outgroup) < 0.01 and q-score (self comparison) < 0.01] to maintain the stringency of our data. Other human duplicated pairs not assigned as WGDs were obtained from the Ensembl biomart 77 and designated small-scale duplicates (SSDs). We used 50% sequence identity and high paralogy confidence to assign paralogs, to retain old and distant paralogs in our dataset. Finally, we obtained 34746

duplicated pairs among which 21446 pairs are SSD-pairs (comprising 4670 genes), and 13300 pairs are WGD pairs (containing 7070 genes).

As the origin of WGDs and SSDs are different in evolutionary time-scale, these two classes of duplicates may also differ in terms of sequence similarity between duplicated pairs. The WGDs are much older, being originated during the evolution of early vertebrates, whereas the SSDs are of both recent and ancient in origin, containing more recent duplicates than WGDs. Therefore, the SSDs are most likely to be less diverged in sequence level than the WGDs. Thus, the bias due to the differential sequence divergence of SSDs and WGDs should be removed for the comparison of the functional properties of these two classes of duplicates. Such bias was removed by binning our dataset according to the nonsynonymous nucleotide substitution per nonsynonymous sites (dN) values between each duplicated pairs, as dN brings changes at the protein level and older duplicate group (WGD) typically have higher dN than the newer one (SSD). We split both the SSD and WGD pairs into five bins based on dN ranges between the paralogs – $dN_{0.0-0.1}$, $dN_{0.1-0.2}$, $dN_{0.2-0.3}$, $dN_{0.3-0.4}$ and $dN_{>0.4}$. We did a pairwise comparison the evolutionary genomic features of SSD and WGD genes in each such dN bin.

For the comparison of individual gene properties of SSD and WGD genes, we made another dataset. From our data of SSD and WGD gene pairs, the genes having multiple paralogs of different origin, that is, at least one originated via small-scale duplication and at least one originated via whole-genome duplication were discarded. Finally, we obtained a nonredundant set of 9386 genes with only SSD or only WGD pairs, but not both. Considering these two groups of gene pairs, we prepared two distinct sets of genes: 1) Genes (and its paralogous copies) involved in Small-scale duplication only (SSD-only)

(containing 3478 genes), and 2) Genes involved in Whole-genome duplication only (WGD only) (containing 5908 genes).

4.2.2. Functional similarity:

The functional annotation of human protein-coding genes by considering their association with Gene Ontology terms were obtained from the Ensembl biomart (version 77)(Flicek *et al.* 2014). We considered the GO domains 'Biological Process' and 'Molecular function' separately for the estimation of functional similarity within paralogous pairs. The functional similarity within each of the SSD- and WGD-pairs were calculated by their GO annotations, using the using Bayesian data integration method (Guan, Dunham, Troyanskaya 2007; Podder, Ghosh 2011), that measures the functional similarity between any duplicated pairs '*i*' and '*j*' as -

$$\text{Functional Similarity}(i, j) = \frac{2 \times S(i, j)}{[\text{GO terms}(i) + \text{GO terms}(j)]}$$

Where '*S(i, j)*' represents the Gene Ontology terms shared between the duplicated gene pairs '*i*' and '*j*'.

4.2.3. Subcellular localization:

We obtained the protein subcellular localization by the association of respective genes' Gene Ontology terms for the GO domain 'Cellular component' from the Ensembl biomart (version 77) (Flicek *et al.* 2014). With the associated GO-terms of a gene and its paralogous copy(ies), we calculated the shared subcellular compartments for each SSD- and WGD-pairs. With the same formula used in section 4.2.2 for the calculation of functional similarity, we calculated the shared subcellular localization for each duplicated gene pairs. We compared the SSD- and WGD- pairs of similar dN bins (as mentioned in Section 4.2.1).

4.2.4. Gene expression:

We obtained the RNA-seq gene expression data of human duplicated genes from two databases- (1) **The Human Protein Atlas** (Release 9) (<http://www.proteinatlas.org/>): containing gene expression values of 9113 duplicated genes in 27 different tissues (namely, *adipose tissue, adrenal gland, appendix, bone marrow, cerebral cortex, colon, duodenum, oesophagus, gallbladder, heart muscle, kidney, liver, lung, lymph node, ovary, pancreas, placenta, prostate, salivary gland, skin, small intestine, spleen, stomach, testis, thyroid gland, urinary bladder, and uterus*) (Uhlen *et al.* 2005; Uhlén *et al.* 2015) and (2) **EMBL-EBI Expression Atlas** (<http://www.ebi.ac.uk/gxa/>): containing 9393 duplicate genes in 32 human tissues (namely *adipose tissue, adrenal gland, ovary, appendix, bladder, bone marrow, cerebral cortex, colon, duodenum, endometrium, oesophagus, fallopian tube, gall bladder, heart, kidney, liver, lung, lymph node, pancreas, placenta, prostate, rectum, salivary gland, skeletal muscle, skin, small intestine, smooth muscle, spleen, stomach, testis, thyroid, and tonsil*) (Kapushesky *et al.* 2012; Petryszak *et al.* 2014). These two repositories present high-throughput experimental RNA-seq gene expression data in human tissues. We obtained the expression profile similarity within each duplicate pairs by the Pearson correlation coefficient, which states that for a paralogous pair '*i*' and '*j*', the expression correlation is-

$$\text{Pearson correlation coefficient } (r) = \frac{N \sum ij - (\sum i)(\sum j)}{\sqrt{[N \sum i^2 - (\sum i)^2][N \sum j^2 - (\sum j)^2]}}$$

Where 'N' is the number of tissues, ' $\sum ij$ ' is the sum of the products of paired expression intensities, ' $\sum i$ ' sum of expression intensities for gene '*i*', ' $(\sum i^2)$ ' is sum of squared expression intensities of gene '*i*', ' $\sum j$ ' is the sum of

expression intensities for gene ' j ', ' $\sum j^2$ ' is sum of squared expression intensities of gene ' j '.

4.2.5. Evolutionary rate:

We calculated the evolutionary rate of human genes by the dN values (Begum, Ghosh 2010), as well as the $\frac{dN}{dS}$ ratio (Wall *et al.* 2005; Chen *et al.* 2012a), both being conventionally and widely used for the estimation of evolutionary rate, where dN refers to Nonsynonymous nucleotide substitution per nonsynonymous sites and dS denotes Synonymous nucleotide substitution per synonymous sites.

In this study, we took one-to-one Human-Mouse (*Homo sapiens*-*Mus musculus*) and Human-Chimpanzee (*Homo sapiens*-*Pan troglodytes*) orthologs to obtain the dN and dS values from Ensembl biomaRT (version 77) (Flicek *et al.* 2014). We controlled the mutation saturation by discarding all dS values ≥ 3.00 (Begum, Ghosh 2014). We compared the evolutionary rate differences between the SSD-only and WGD-only gene groups.

4.2.6. Multifunctionality:

The Multifunctionality of a gene and its encoded protein was measured by two approaches: (A) Using their Gene Ontology annotation (Gene Ontology 2004) for the GO domain 'biological process' from Ensembl Genome Browser (Flicek *et al.* 2014), we calculated the unique biological processes of which a gene and its encoded protein(s) take part and used as the measurement of multifunctionality (Podder, Mukhopadhyay, Ghosh 2009; Satake *et al.* 2012), (B) Additionally, we also considered the number of functional protein domains as proxy of Multifunctionality using Pfam protein families database. Finally, we compared the multifunctionality of SSD-only and WGD-only genes.

4.2.7. Gene essentiality:

We obtained human essential and nonessential genes from the Online GENE Essentiality (OGEE) database (<http://ogeedb.embl.de/#overview>) (Chen *et al.* 2012b). From this, we were able to match the gene essentiality information of 2692 SSD-only and 5730 WGD-only genes in our dataset. We calculated the proportion of essential genes within each of these duplicate sets.

4.2.8. Disease genes:

The genes which cause disease phenotypes upon mutation(s) is referred to as 'disease genes'. We obtained such disease genes in human genome from the 'Human Gene Mutation Database' (<http://www.hgmd.cf.ac.uk/ac/index.php>) (Stenson *et al.* 2012), which contains both the monogenic and polygenic disease genes and considered these together as human disease genes. From this database, we collected 9668 disease genes, among which 9299 genes were matched to our dataset. All the other genes were considered as non-disease genes (N= 13148). Here, we compared the proportion of disease genes between the SSD-only (N=3478) and WGD-only (N=5908) sets.

4.2.9. Software:

The SPSS package (version 13) (Nie, Bent, Hull 1970) and our in-house PERL-script was used for all statistical analyses. The R-package (Ihaka, Gentleman 1996) was used for representation of data.

4.3. Results:

4.3.1. Functional similarity of human SSD and WGD pairs:

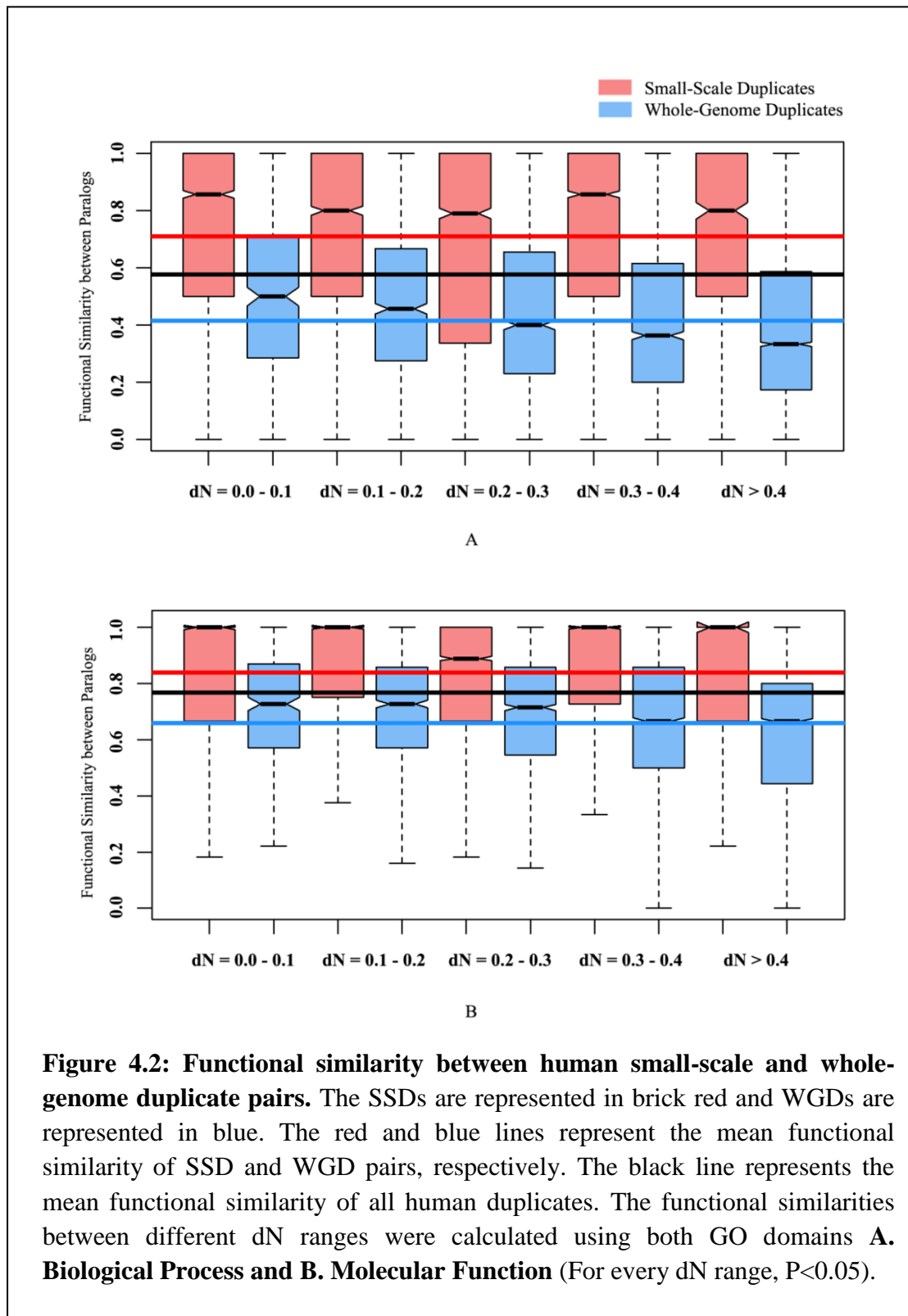
We compared the functional similarities between the human small-scale and whole-genome duplicate pairs using the Gene Ontology (GO) terms for both

GO domains 'biological process' and 'molecular function' from Ensembl biomart (version 77) (Flicek *et al.* 2014). We observed a higher functional similarity in small-scale duplicate (SSD) pairs than the whole-genome duplicates (WGD) (Table 4.1). As we know, the functional similarity between duplicated pairs usually decrease due to their nucleotide substitutions, which is also dependent on the age of the duplicates, the older duplicates being prone to more nucleotide substitutions. Thus, we compared the functional similarity between paralogs by binning our dataset according to different dN (nonsynonymous nucleotide substitution per nonsynonymous site) ranges (Section 4.2.1), as dN brings changes in amino acids in genes' encoded protein(s). The binning was adapted from Hakes *et al.* (Hakes *et al.* 2007), and we observed that WGDs have a higher dN value the SSDs, as they are evolutionarily more ancient. We found a higher functional similarity among SSD-pairs in contrast to the WGD-pairs in all the dN bins (Table 4.1) considering both the GO domains- biological processes and molecular function (Figure 4.2). In other terms, human WGD pairs are diverge more in their function, irrespective of their sequence divergence.

4.3.2. Subcellular localization of human SSD and WGD pairs:

As the function of genes are typically mediated by their encoded proteins, which becomes relocated to the desired cellular compartments after being synthesized. The function of a protein is usually limited to the cellular compartment to which it is localized and hence, the proteins encoded by paralogous genes may become localized to different cellular compartments. Thus, the subcellular protein compartmentalization neutralizes the functional redundancy of duplicates at the protein level (Byun-McKay, Geeta 2007; Marques *et al.* 2008; Qian, Zhang 2009). In this study, we considered

the protein subcellular localization as a fate of the duplicated genes.

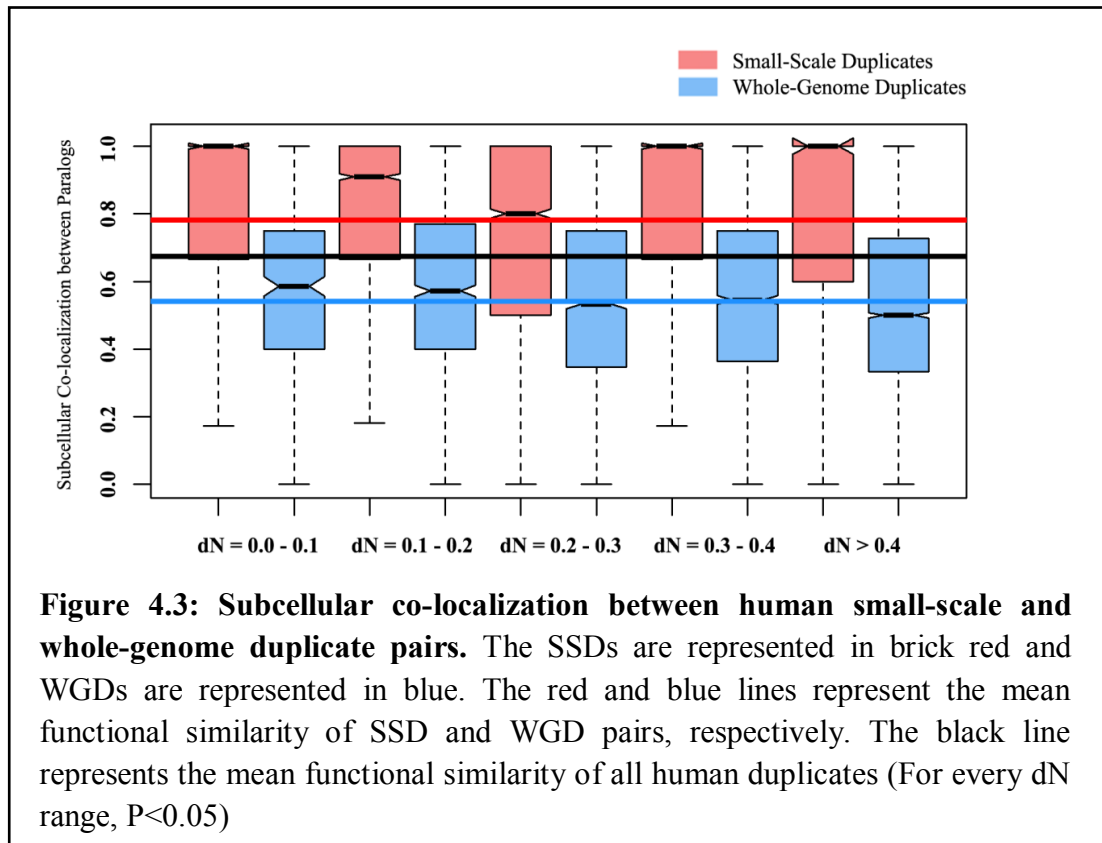


We obtained the protein subcellular localization of human genes using the Gene Ontology (GO) terms with the GO domain 'Cellular Component'.

Table 4.1. The differences between human small-scale duplicate(SSD) and whole-genome duplicate(WGD) pairs in the whole dataset and different dN bins. Pair wise Mann-Whitney U-test was used to compute the P-values.				
Parameter Measured	Database used	Overall	dN= 0.0-0.1	dN= 0.1-0.2
Functional Similarity between paralogs	Shared GO Terms for Biological Process	P-value	1.983×10 ⁻¹³⁵	5.072×10 ⁻¹⁰⁷
		WGD	\bar{x} = 0.391, N= 6274	\bar{x} = 0.628, N= 6468
		SSD	\bar{x} = 0.710, N= 756	\bar{x} = 0.826, N= 860
		P-value	1.397×10 ⁻²⁷⁴	4.976×10 ⁻²¹⁸
		WGD	\bar{x} = 0.413, N= 2192	\bar{x} = 0.677, N= 2250
		SSD	\bar{x} = 0.726, N= 4264	\bar{x} = 0.846, N= 5246
	Shared GO Terms for Molecular Function	P-value	2.892×10 ⁻¹²⁰	1.814×10 ⁻⁹⁶
		WGD	\bar{x} = 0.440, N= 2002	\bar{x} = 0.696, N= 2076
		SSD	\bar{x} = 0.657, N= 3328	\bar{x} = 0.810, N= 4300
		P-value	5.925×10 ⁻¹²³	1.075×10 ⁻¹²⁹
		WGD	\bar{x} = 0.476, N= 1140	\bar{x} = 0.706, N= 1188
		SSD	\bar{x} = 0.720, N= 2754	\bar{x} = 0.856, N= 3510
Shared Subcellular Compartment of paralogs	GO Cellular Component	P-value	2.325×10 ⁻⁵²	6.077×10 ⁻⁴⁷
		WGD	\bar{x} = 0.499, N= 414	\bar{x} = 0.724, N= 410
		SSD	\bar{x} = 0.734, N= 3640	\bar{x} = 0.850, N= 4668
		P-value	<1.00×10 ⁻⁶	<1.00×10 ⁻⁶
		WGD	\bar{x} = 0.415 N= 12022	\bar{x} = 0.659, N= 12392
		SSD	\bar{x} = 0.710, N= 14742	\bar{x} = 0.840, N= 18584

Table 4.1. (continued) The differences between human small-scale duplicate(SSD) and whole-genome duplicate(WGD) pairs in the whole dataset and different dN bins. Pair wise Mann-Whitney U-test was used to compute the P-values.				
Parameter Measured	Database used	Overall	dN = 0.0-0.1	
			dN = 0.1-0.2	dN = 0.2-0.3
Gene expression profile similarity between paralogs	Human Protein Atlas	P-value	1.558×10 ⁻⁶³	1.032×10 ⁻³⁰
		WGD	\bar{x} = 0.254, N= 426	\bar{x} = 0.284, N= 422
		SSD	\bar{x} = 0.615, N= 2588	\bar{x} = 0.508, N= 3628
		P-value	<1.00×10 ⁻⁶	<1.00×10 ⁻⁶
		WGD	\bar{x} = 0.193, N= 13060	\bar{x} = 0.216, N= 13072
		SSD	\bar{x} = 0.403, N= 11726	\bar{x} = 0.450, N= 15404
	Expression Atlas	P-value	1.774×10 ⁻⁴²	5.953×10 ⁻⁵³
		WGD	\bar{x} = 0.253, N= 1226	\bar{x} = 0.280, N= 1220
		SSD	\bar{x} = 0.414, N= 2758	\bar{x} = 0.457, N= 3458
		P-value	7.331×10 ⁻³²	1.377×10 ⁻¹⁰⁵
		WGD	\bar{x} = 0.191, N= 2158	\bar{x} = 0.216, N= 2166
		SSD	\bar{x} = 0.307, N= 2834	\bar{x} = 0.430, N= 3792
		P-value	1.131×10 ⁻³⁴	5.471×10 ⁻⁹⁶
		WGD	\bar{x} = 0.190, N= 2366	\bar{x} = 0.219, N= 2370
		SSD	\bar{x} = 0.316, N= 3042	\bar{x} = 0.420, N= 3922
		P-value	1.308×10 ⁻¹⁷	5.735×10 ⁻³⁶
		WGD	\bar{x} = 0.179, N= 6884	\bar{x} = 0.199, N= 6894
		SSD	\bar{x} = 0.322, N= 504	\bar{x} = 0.394, N= 604

We calculated the subcellular co-localization between the paralogous copies of human SSD- and WGD-pairs (see Materials and methods). The subcellular colocalization denotes the shared cellular compartments by a duplicated pair. We obtained a higher subcellular co-localization of SSD pairs than the WGD pairs (Table 4.1). After binning our dataset in different dN ranges as mentioned earlier, the trend remains unchanged in each dN range (Table 4.1, Figure 4.3), suggesting that the proteins encoded by SSD-pairs are more often co-localized and that in WGD-pairs are co-localized less often, irrespective of their sequence divergence.



4.3.3. Gene expression correlation between SSD and WGD pairs:

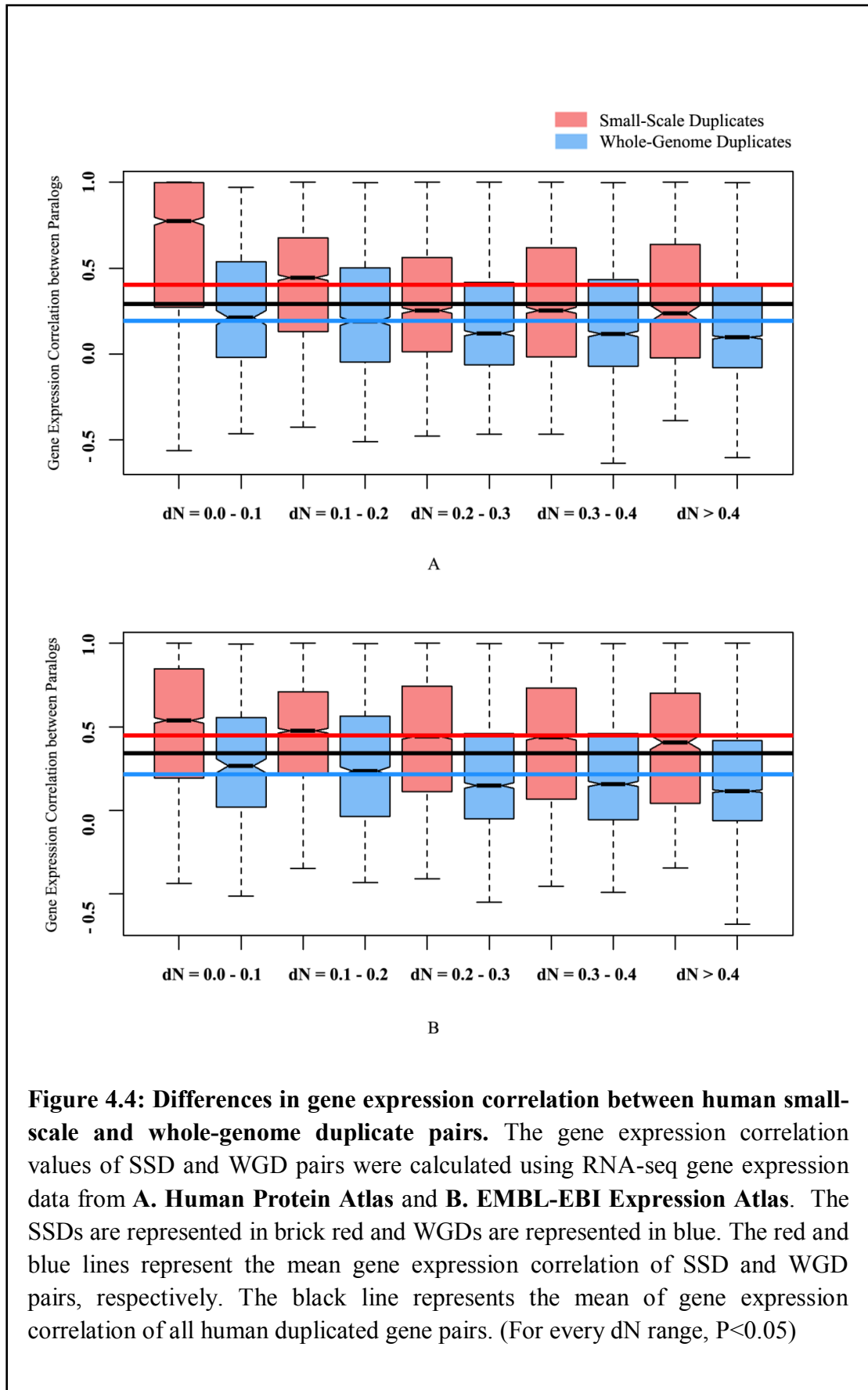
Gene expression is an important contributor in the maintenance of duplicated genes. Duplication leads to an increment of gene products, which necessitates their divergence. Previous studies suggested that gene expression patterns of duplicated pairs usually undergo spatial variation

[reviewed in Li *et al.* (Li, Yang, Gu 2005)], leading to their stable maintenance (Qian *et al.* 2010). Thus, the divergence of duplicated pairs also occurs at their gene expression profiles. To delve further in this context, we explored the co-expression of duplicate pairs among different tissues after gene duplication, using the gene expression profiles of human genes and their paralogous copies in a wide range of normal tissues (Li, Yang, Gu 2005; Ganko, Meyers, Vision 2007; Marques *et al.* 2008).

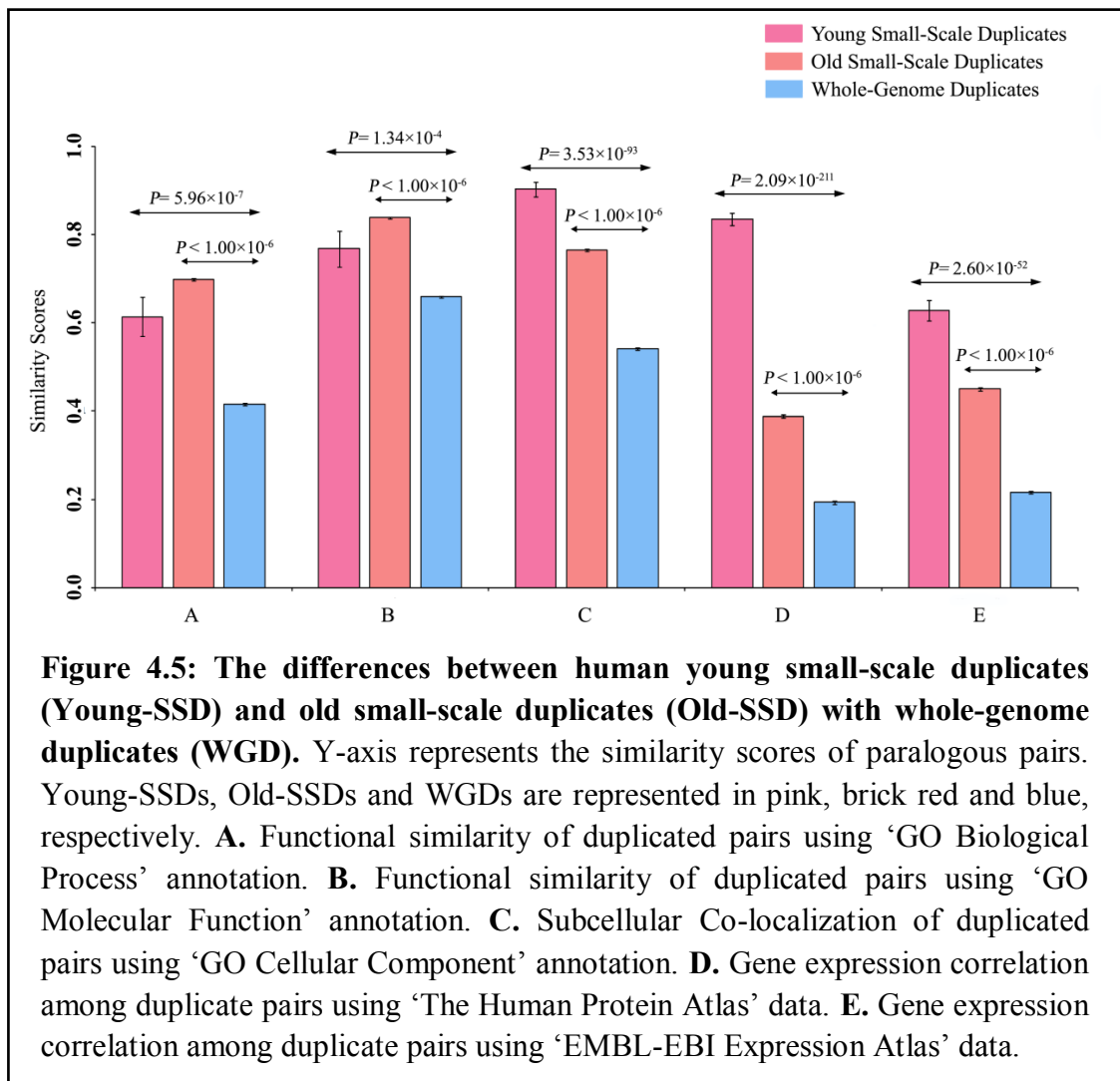
We obtained recent high-throughput RNA-seq gene expression data of a wide range of normal human tissues from the **Human Protein Atlas** (Uhlen *et al.* 2005; Uhlén *et al.* 2015) and **EMBL-EBI Expression Atlas** (Kapushesky *et al.* 2012; Petryszak *et al.* 2014) (see Section 4.2.4). We used the Pearson correlation coefficient to calculate the coexpression of paralogous pairs and compared the SSD- and WGD-pairs. We observed a higher co-expression in human SSD-pairs than the WGD-pairs. The result is also consistent when splitted in dN bins (Table 4.1, Figure 4.4). This suggests that the functionally redundant SSD-pairs are coexpressed in the same tissue(s) more often than the functionally divergent WGD-pairs.

4.3.4. Comparison of human whole-genome duplicates with young and old small-scale duplicates:

Our study with human small-scale and whole-genome duplicates clearly suggests that these two groups of duplicates are quite different in their evolutionary genomic properties. To explain this, we hypothesized that our results reflect the long-term evolutionary fates of vertebrate whole-genome duplication. However, as the timing of duplication and subsequently the age of WGD and SSD duplicates may be different, we were interested to observe the proportion of recent and ancient duplicates among the SSD duplicates in



our dataset. For this, we obtained the phylostratum gene age data from Neme and Tautz (2013)(Neme, Tautz 2013) where the human genes were ranked according to their earliest evolutionary origin. We classified all the SSD genes in our dataset into two groups- (A) Old SSD: representing all the genes before the emergence of eutherian mammals (having phylostratum rank 1-15) and (B) Young SSD: genes originated during eutherian lineage or later (having phylostratum rank 16-20). Mapping these two classes with our dataset of 4640 genes involved in small-scale duplication, we obtained 3888(95.29%) Old-SSD and 192(4.71%) Young-SSD genes. We mapped these Old- and New-SSD genes with our dataset of 21446 SSD pairs and obtained 14846 Old-SSD pairs and 642 Young-SSD pairs. We discarded the duplicated pairs where one



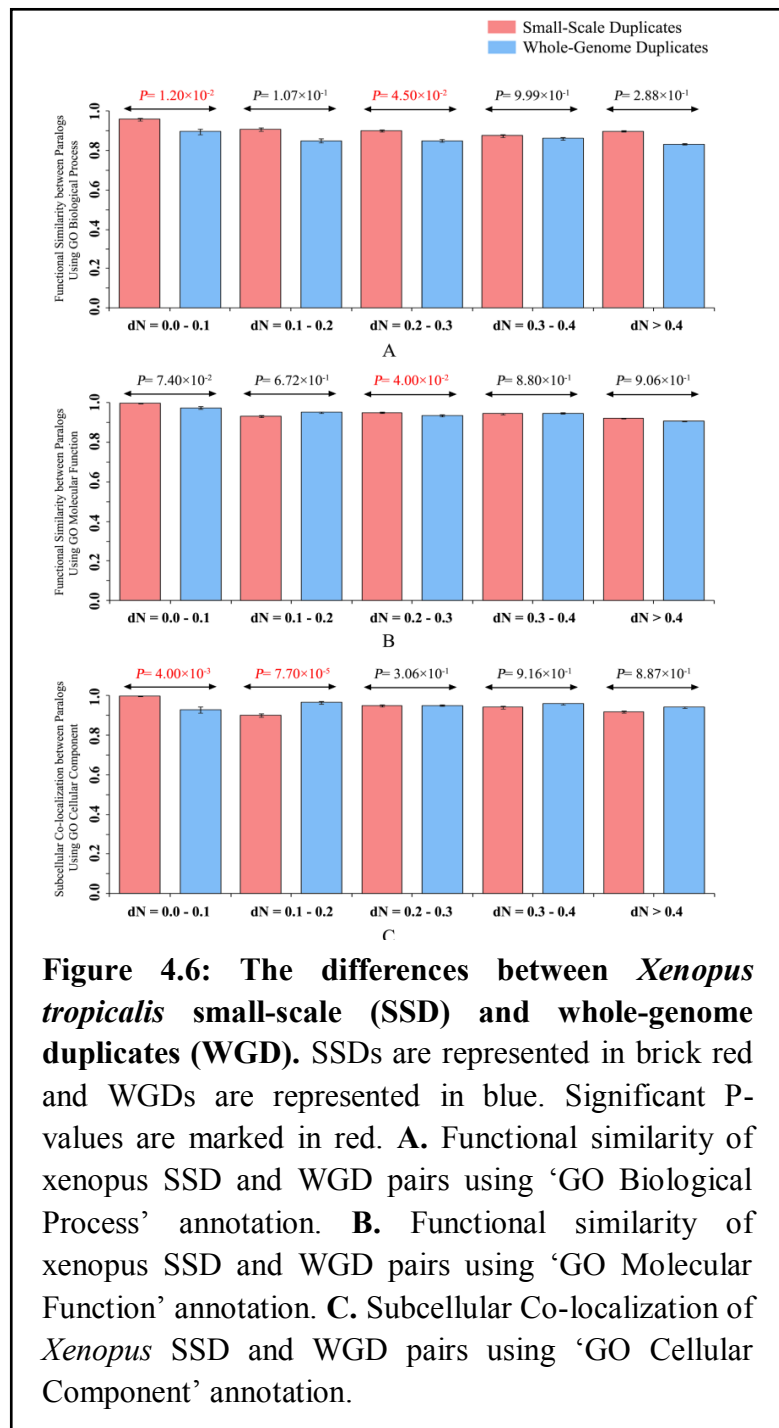
gene is Old-SSD, and other is Young-SSD for maintaining stringency and facilitate pairwise comparison. We observed that both the Old- and Young-SSD genes show significant differences with WGD genes (Figure 4.5).

In addition to this, the low proportion of Young-SSD genes in our dataset clearly indicates that indeed our data of SSD pairs are not significantly enriched in younger genes ($Z = 79.875$, confidence level 99%; $P < 1.00 \times 10^{-4}$, two sample Z-test). Therefore, the differences between human SSD and WGD duplicates really reflect the long-term fate of vertebrate whole-genome duplicates, in comparison with small-scale duplicates.

4.3.5. The difference between Small-scale and Whole-genome duplication in *Xenopus tropicalis* genome:

Our results suggest a higher functional divergence, less subcellular colocalization and lower gene expression correlation between WGD-pairs, in contrast to SSDs. This indicates a possibility of whole-genome duplicates to become diverged and adapted to new functions, reflecting the fate of vertebrate genome duplication in long evolutionary time-scale. However, as humans are very distantly related to the whole genome duplication event during early vertebrate evolution, we compared WGD and SSD in *Xenopus* genome to strengthen our conclusion. As *Xenopus* (Class: Amphibia, Order: Anura) is much more closely related with reference to whole-genome duplication than human, we compared the *Xenopus* SSD and WGD-pairs for a detailed insight into the fate of SSD and WGD-genes. We used our dataset of 34746 human duplicated gene pairs and matched with their one-to-one *Xenopus* (*Xenopus tropicalis*) orthologs from Ensembl biomart (version77)(Flicek *et al.* 2014). Finally, we obtained 1020 SSD and 8078 WGD pairs of *Xenopus*. Similar to our analysis in humans, the functional similarity

between *Xenopus* duplicated pairs were obtained using the Gene Ontology annotation for GO domains 'Biological Process' and 'Molecular Function'. The protein subcellular co-localization was measured using Gene Ontology annotation for GO domain 'Cellular Component'. We binned our dataset according to different dN ranges (as described in the Materials and Methods section in the manuscript) and compared SSD and WGD pairs within each dN range. We



observed that, unlike humans, most of the differences are insignificant. We summarized all results in Figure 4.6. For ease of understanding, the significant P-values were marked in red font.

4.3.6. Evolutionary rate of human SSD and WGD genes:

The differences in the evolutionary genomic attributes of human SSD- and WGD-pairs clearly suggest that the human WGDs diverge themselves to become expressed in new locations and serve new functions. However, the evolutionary rate differences between these two classes of duplicated genes are unclear. Thus, we compared the evolutionary rates of human SSD-only and WGD-only genes by their dN-values and the $\frac{dN}{dS}$ ratios (see Section 4.2.5 for details), using their one-to-one Mouse and Chimpanzee orthologs. We obtained a significantly slower evolutionary rate in WGD-only genes in all the cases (Figure 4.7). This indicates that the human duplicated genes originated via whole-genome duplication are evolutionarily more conserved, besides their higher functional divergence and lower coexpression and co-localization than the SSD genes. Our result is consistent with a previous study (Satake *et al.* 2012) and in agreement with the idea of slower evolutionary rate of duplicated genes after adapting to new functions and locations, as revealed by Jordan *et al.* (Jordan, Wolf, Koonin 2004).

4.3.7. Multifunctionality of human SSD and WGD genes:

The higher functional divergence along with the lower subcellular co-localization and gene expression correlation of human WGD genes and their higher evolutionary conservation indicate that they tend to adapt to miscellaneous functions, compared to the SSD counterparts. As one of our major aim of present study is to explore functional fates of human SSD and WGD genes, we were curious which group among these two is associated with more functions. For this, the unique GO biological process terms (Salathe, Ackermann, Bonhoeffer 2006; Podder, Mukhopadhyay, Ghosh 2009) and the Pfam domain count (Finn *et al.* 2014) were used as proxies of multifunctionality (see, Section 4.2.6). Comparing the SSD-only and WGD-only

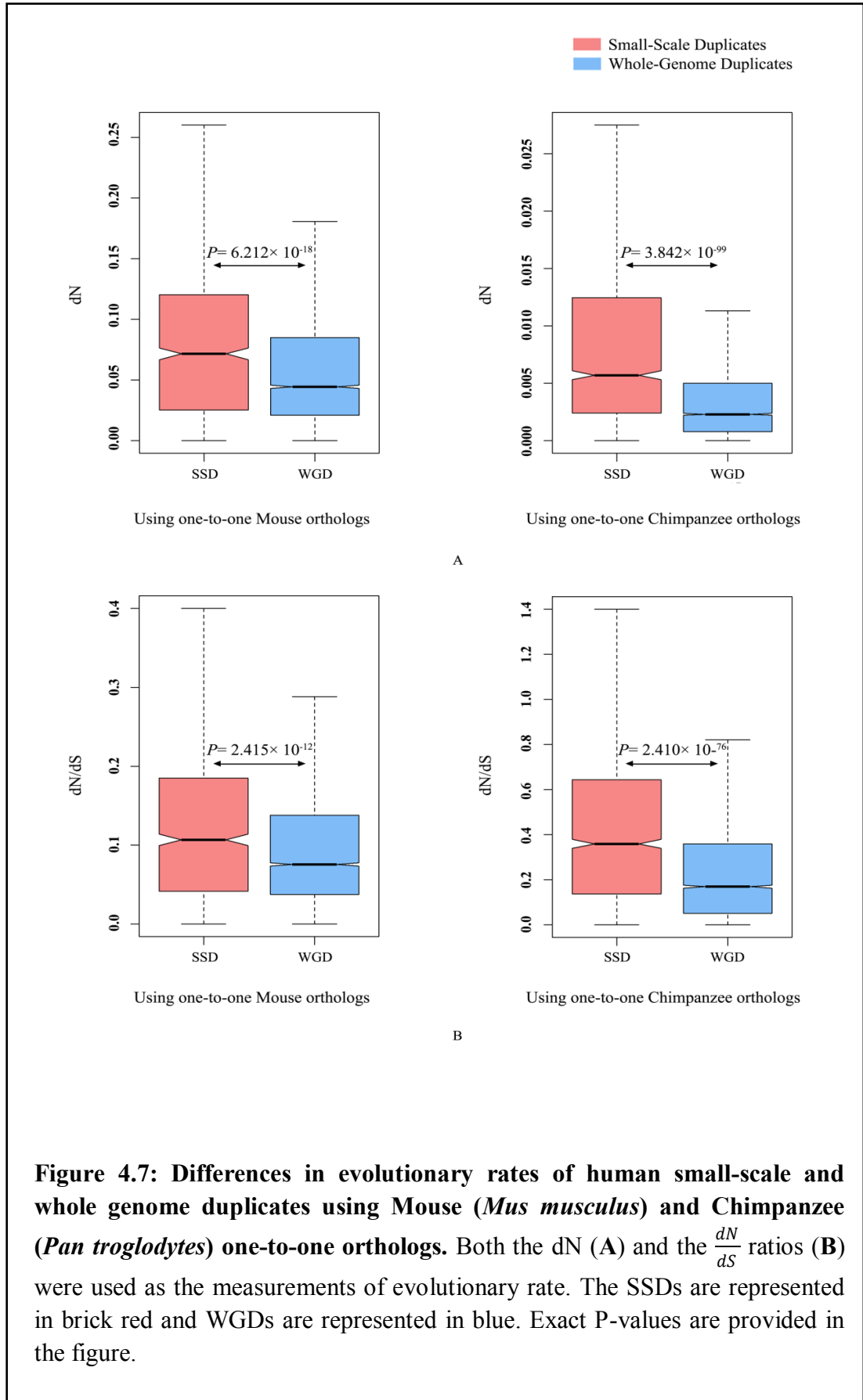
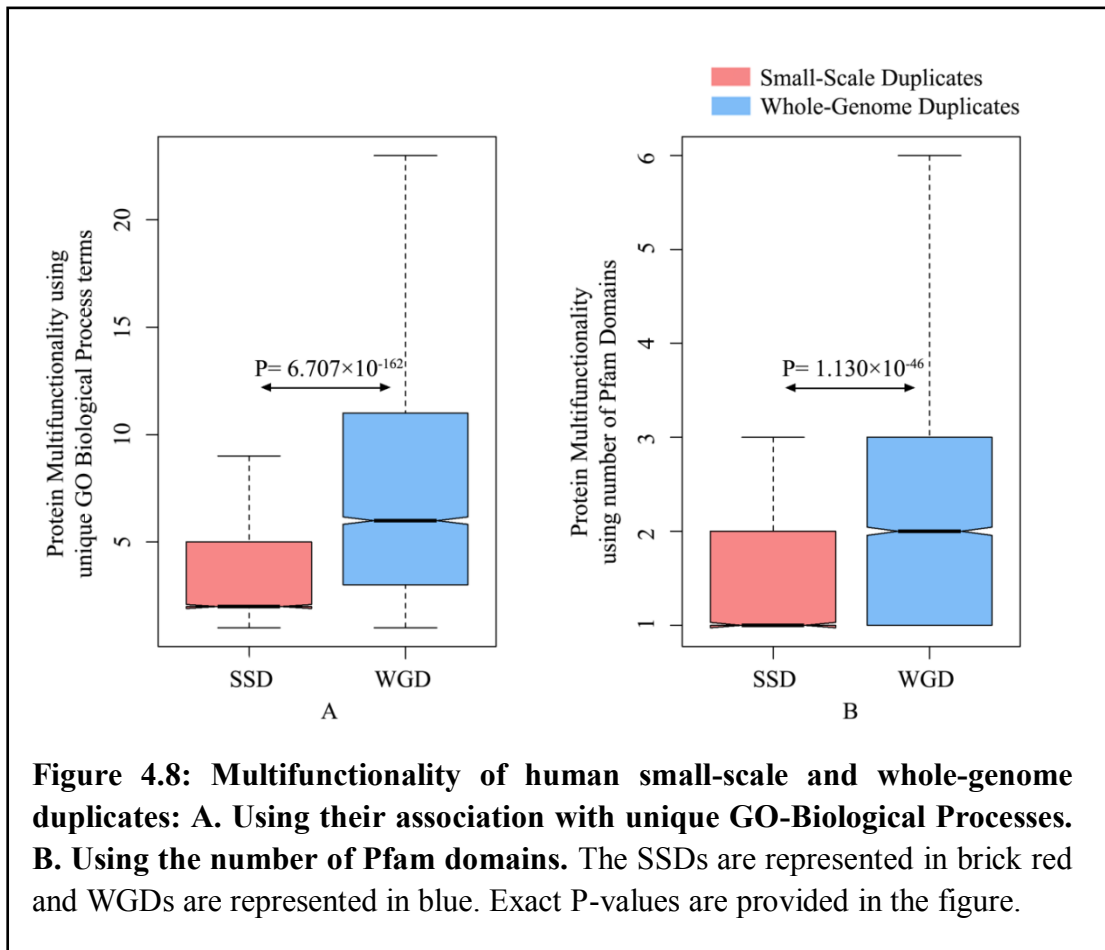


Figure 4.7: Differences in evolutionary rates of human small-scale and whole genome duplicates using Mouse (*Mus musculus*) and Chimpanzee (*Pan troglodytes*) one-to-one orthologs. Both the dN (A) and the $\frac{dN}{dS}$ ratios (B) were used as the measurements of evolutionary rate. The SSDs are represented in brick red and WGDs are represented in blue. Exact P-values are provided in the figure.

genes, we observed that WGD genes are associated with more GO biological process terms [Mean number of unique GO-BP terms in SSD-only genes ≈ 5 , Mean number of unique GO-BP terms in WGD-only genes ≈ 10 , $P = 6.707 \times 10^{-162}$, *Mann Whitney U test*, $N_{\text{SSD}} = 2569$, $N_{\text{WGD}} = 5437$] (Figure 4.8A). Additionally, the WGD group contains significantly more domains in their encoded proteins than the SSD group [Mean number of Pfam domains in SSD = 1.61, Mean number of Pfam domains in WGD = 2.02, $P = 1.130 \times 10^{-46}$, *Mann Whitney U test*, $N_{\text{SSD}} = 3060$, $N_{\text{WGD}} = 5607$] [Figure 4.8B]. Together, these results suggest that the human whole-genome duplicates are associated with more number of functions than the small-scale duplicates.



4.3.8. Gene essentiality of human SSD and WGD genes:

Our study suggests that the WGD genes are evolutionarily more conserved and adapted to more functions than the SSD genes. However, for a deeper understanding of the important roles played by WGD genes in humans, the importance of such functions from organismal perspective is also crucial.

The importance of a gene is usually measured in terms of its essentiality, as indicated by the fitness effect of gene-deletion. Essential genes, that leads to sterility or lethality in humans were obtained from the Online GENE Essentiality (OGEE) Database (Chen *et al.* 2012b). Comparing the proportion of essential genes in human SSD-only and WGD-only genes, we observed a significantly higher proportion of essential genes within the WGD class [Proportion of essential genes in SSD-only genes= 4.601%, Proportion of essential genes in WGD-only genes= 11.344%; $N_{SSD} = 2692$, $N_{WGD} = 5730$] [$Z = -9.99$, confidence level 99%; $P < 1.00 \times 10^{-4}$, two sample Z-test]. In other terms, a higher portion of WGD genes are required for the viability than SSD genes, which may be due to the absence of redundant paralogs in WGDs.

4.3.9. Disease association of human SSD and WGD genes:

Another contributing factor to the importance of a gene within an organism is its association with disease. Previous studies suggest that the gene duplication and subsequent increase in functional redundancy reduces the risk of disease formation by functional restoration upon deleterious mutations (Dean *et al.* 2008; Hsiao, Vitkup 2008; Wagner 2008). Therefore, the genes involved in disease should remain as singletons (Forslund *et al.* 2011). More recent studies hypothesise that the increased genetic redundancy after gene duplication prefers accumulation of disease-prone mutations on the duplicates. As a consequence, the duplicates may be more disease prone than singletons (Dickerson, Robertson 2012). Works with

Monogenic disease genes revealed their association with whole-genome duplicates (Makino, McLysaght 2010; Chen *et al.* 2013). In this study, we obtained the human disease associated genes from the Human Gene Mutation Database (HGMD) (Stenson *et al.* 2012), which contains both monogenic (Mendelian) and polygenic (complex) disease genes. We observed that theOur result revealed a significantly higher proportion of disease genes among the duplicates originating from whole-genome duplication [Proportion of disease genes in WGD= 61.46%, $N_{\text{WGD}} = 5908$]; than those originating from small-scale duplication [Proportion of disease genes in SSD genes= 27.89%, $N_{\text{SSD}} = 3478$] [$Z = -31.420$, confidence level 99%; $P < 1.00 \times 10^{-4}$, two sample Z-test]. Together, these results suggest that the functions to which the human WGD genes are adapted are more vital than that of SSD genes. Also, the reduced functional redundancy of WGD genes increases their susceptibility to cause disease, lethality, and sterility in contrast to the functionally more redundant SSD genes.

4.4. Discussions:

Gene duplication is the main source of new genetic materials, thus playing a major role in increasing genetic novelty and genome evolution. Gene duplication and subsequent accumulation of mutations lead to the generation of new genes from the older ones. Mutation on the duplicated gene copies creates structural changes within the DNA, subsequently leading to changes in protein structure and function.

Although after gene duplication the duplicated copies of a gene may retain their functional redundancy and maintained as backup copies, they may accumulate mutations to diverge and adopt new functions during evolution (Ohno 1970; Zhang 2003; Taylor, Raes 2004). However, duplication of a gene

creates a disparity in the protein-protein interaction network, as the duplicated gene produces more proteins than its singleton interacting partners. Such disparity, known as ‘dosage imbalance’, becomes even more pronounced after the duplication of a highly-connected (hub) gene (Jeong *et al.* 2001). However, based on its extent, gene duplication may have dissimilar effects on the gene-dosage in protein-protein interaction (PPI) network. Previous studies with yeast revealed that the duplicates originating from whole-genome duplication event maintain their stoichiometry within the protein-protein interaction network, as it increases the dosage of its every participant. Duplicates originating via small-scale duplication, in contrast, creates a stoichiometric imbalance within the PPI-network. They have also shown that yeast small-scale duplicates become functionally more divergent to maintain the stoichiometric balance of PPI network (Lynch, Conery 2000; Freeling, Thomas 2006; Hakes *et al.* 2007; Makino, McLysaght 2010; Fares *et al.* 2013).

However, as whole-genome duplication leads to the simultaneous generation of many genes, they are associated with major evolutionary transitions (Vandepoele *et al.* 2004; Dehal, Boore 2005; Singh *et al.* 2012; Singh, Arora, Isambert 2015b). Thus, we hypothesized that with increasing complexity and genetic robustness of organisms, the whole-genome duplicated genes may adapt to new functions, maintaining the resilience of the PPI network at the same time. In this study, we explored the long-term fates of vertebrate whole-genome duplication, by analyzing the human whole-genome duplicates (WGDs). The human WGDs have originated from the two rounds of vertebrate whole-genome duplication occurred long time ago in evolutionary scale (~530 Mya). Thus, they must be evolved during the evolution from fish to humans.

In this study, we compared the attributes of human duplicates originated via small-scale duplication to those via whole-genome duplication. As the SSDs and WGDs are not similar in their origin, and therefore there are differences in their sequence divergence, with the older WGD-pairs having a higher probability of accumulating sequence divergence. Therefore, we binned our datasets based on the non-synonymous nucleotide substitutions (dN) between duplicated pairs to compare the evolutionary genomic properties of SSD and WGD duplicates (Hakes *et al.* 2007). Using this approach, we were able to compare these duplicates independent of the changes in nucleotide sequence that bring changes in amino acids, and in turn encoded proteins (Hakes *et al.* 2007).

Our results suggest that the human SSDs and WGDs possess subtle differences in their evolutionary genomic properties. While SSD pairs are functionally more similar to each other than the WGD pairs, irrespective of their sequence divergence. The results are same using both the 'Biological Process' and 'Molecular Function' domains of Gene Ontology(GO) (Figure 4.2, Table 4.1). Thus, the whole-genome duplicates tend to diverge functionally more than the small-scale duplicates. Furthermore, the functions of a gene is mediated by its encoded proteins, which after their synthesis moves to the desired cellular compartment where they function (Emanuelsson, Heijne 2001). Therefore, the subcellular localization of proteins encoded by duplicated genes is also associated with their functional diversification (Byun-McKay, Geeta 2007; Marques *et al.* 2008). Our study revealed a higher subcellular protein colocalization in SSD pairs (Figure 4.3, Table 1), indicating that the human WGD-pairs also diverge more in their protein subcellular localization. Thus, our study with human SSDs and WGDs revealed an exactly opposite trend revealed by the earlier studies with yeast SSDs and WGDs,

where SSD pairs were more divergent in terms of both function and subcellular localization.

In contrast to the lower unicellular eukaryotes, higher eukaryotes possessing tissue-level organization can regulate the duplicated gene copies at their gene expression level among different tissues (Li, Yang, Gu 2005; Ganko, Meyers, Vision 2007; Leach *et al.* 2007; Ha, Kim, Chen 2009; Li *et al.* 2009; Qian *et al.* 2010). For example, after gene duplication, the functionally redundant paralogs may adapt themselves to express differentially in different tissues, so that the overall expression breadth (number of tissues where a gene is expressed) of the gene prior to duplication is maintained. The spatial variation of gene expression, therefore, can be treated as a possible mechanism associated to the maintenance of duplicated pairs in multicellular organisms. We used the high-throughput RNA-seq gene expression data of human to compare the spatial variation in gene expression patterns of SSD- and WGD-pairs.

We observed that the human SSD pairs are coexpressed in the same tissue more often than the WGD pairs, which tend to express in different tissues (Figure 4.4). This reveals that human whole-genome duplicate pairs are not only adapted to divergent functions or new locations, but also expressed in different tissues.

Therefore, from these results, it is quite clear that human WGD-pairs are more divergent in their function, subcellular localization and gene expression than the SSD pairs. However, as humans are very distantly related to vertebrate whole-genome duplication event (~530 Mya), our results may reflect the outcome of more than 500 million years of evolution of WGD genes. Thus, we hypothesise that our results demonstrate the long-term

evolutionary outcome of genes originating via WGD, with those generated via SSD. However, this may not be true as the results may be due to an enrichment of recent SSD-pairs in our dataset, which are usually functionally more similar. Thus, we classified the SSD-pairs in two groups- young-SSD pairs and old-SSD pairs based on their phylostratum rank (Neme, Tautz 2013). The proportion of young SSDs in our dataset was found to be very low, suggesting they have no significant effect on our results. For further confirmation, we compared the old-SSD and the young-SSD separately with the WGD genes and observed that both the old- and young-SSDs show differences with WGDs (Figure 4.5). This suggests that the age of SSD genes have no significant influence on the differences in evolutionary genomic attributes of human SSD- and WGD-genes. Also, among the old-SSD and the WGDs, WGD-pairs diverge themselves more than the old-SSDs, despite both being evolutionarily older counterparts of the genome.

To further strengthen our hypothesis, we used *Xenopus tropicalis* as a control and compared the evolutionary genomic features of *Xenopus* orthologs of human small-scale and whole-genome duplicates. Interestingly, both the SSD- and WGD-pairs shows high functional similarity value in *Xenopus*, with very little or no significant difference between the two classes of duplicates in their functional similarity and subcellular colocalization (Figure 4.6). We could not analyze the gene expression correlation between the paralogous pairs in *Xenopus*, due to unavailability of data, but our results are sufficient to reveal that although initially after the vertebrate two round of whole-genome duplication event both the SSD and WGD genes were similar in their attributes, but during the course of evolution, the WGD genes established themselves as more suitable candidates to diverge themselves to perform new functions.

However, as the functional redundancy decreases and the duplicated gene copies become diversified enough to separate functionally, their evolutionary rate reduces and they become evolutionarily conserved to maintain the function (Jordan, Wolf, Koonin 2004). Thus we were curious to observe the evolutionary rate differences between human SSD and WGD genes. Our results revealed a slower evolutionary rate and therefore, higher evolutionary conservation in human WGD genes compared to the SSD counterparts. Thus our study furnish a clear indication that the human WGDs have adapted to new locations, serves new functions and lost their redundancy, eventually became slow evolving to maintain themselves (Figure 4.7). Additionally, it also suggests that the functions adopted by the WGDs are also evolutionarily conserved. Furthermore, to obtain a detailed insight into the functional roles played by SSDs and WGDs in humans and the importance of their functions, we explored the multifunctionality, gene essentiality and the disease-association of genes within these two groups. We obtained a higher protein multifunctionality in WGDs, indicated by their association with more numbers of unique Gene Ontology biological process terms and more functional domains within their proteins' structure (Figure 4.8). In the next part, we compared the functional importance of these duplicates. We studied the proportion of essential genes, as they comprise the most important part of the genome. Although the human essential genes show a significantly higher enrichment in duplicates than in singletons (Acharya *et al.* 2015), but the difference in the enrichment of essential genes within SSD and WGD is still not clear. Additionally, considering the disease-associated genes, duplicates contains a higher proportion of such genes than the singletons (Dickerson, Robertson 2012).

When we compared the functional importance of human SSD and WGD genes, by studying the proportion of essential genes and disease-associated genes. Our results showed a significantly higher enrichment of essential genes among WGD genes. Additionally, considering both monogenic and polygenic disease genes together, we found a significant higher enrichment of disease-associated genes within WGDs, a result consistent with earlier studies with monogenic (Mendelian) disease genes only (Makino, McLysaght 2010). These results suggest a higher involvement of functionally important genes with the WGDs. Thus, our study provide a clear demonstration of the fate of the human whole genome duplicates originated during vertebrate whole genome duplication event representing the adaptation of these WGD genes to serve more functions, and functions that are crucial and vital for human survival and may cause disease, sterility, and lethality upon disruption.

4.5. Conclusions:

In this study, we compared the human small-scale and whole-genome duplicates based on their genomic and evolutionary attributes. Our results suggest that the human duplicates originated from whole-genome duplication (WGDs) during vertebrate evolution show various differences with those originating at a smaller-scale (SSDs). However, these differences between human SSDs and WGDs are exactly opposite to that in yeast. We hypothesised that such a trend reflects the preservation of WGD genes in long evolutionary time span, as the human WGDs as human WGDs have originated from two rounds of whole-genome duplication during early vertebrate evolution. However, the human SSDs are also enriched in ancient genes in our dataset. Therefore, both these duplicates have faced many

evolutionary challenges. But in course of evolution, WGDs diverge significantly more in function, location and expression, as suggested by our study. The human WGDs are associated with more functions and perform crucial roles than the SSD genes. Thus, the WGDs cause more profound effects upon mutations, for the inability of their paralogous genes to mask the fitness effect of gene-deletion.

Therefore, our study represents long-term evolutionary fates of whole-genome duplication, in contrast to their immediate effect on the organism, as shown by the early studies with yeast (Guan, Dunham, Troyanskaya 2007; Hakes *et al.* 2007; Fares *et al.* 2013).

Chapter 5

Summary and General Conclusion.

This thesis focus on the importance of human gene duplication and duplicated genes generated by such duplications in human genome evolution. Gene duplication is a genetic mutation that generates new gene copies capable of developing new function upon the act of mutations that modify the gene sequence. For this reason, gene duplication is considered as the major evolutionary force guiding genome and organism evolution. However, duplication of a gene often leads to the generation of ‘useless duplicates’ with that are nonfunctional and remain as the ‘debris’ within the genome, increasing genomic burden. Occasionally, duplicates are retained as backup copies, particularly when the increase in gene product is beneficial to the organism. Such duplicates are retained within the genome and may subsequently diversify themselves functionally, after acquiring mutations required for such functional changes. More precisely, duplication of a gene leads to the relaxation of purifying selection on that gene and as a consequence, both the duplicated copies start accumulating mutations. This impedes the backup capacity of duplicated copy, but in turn may lead to the generation of new functions. Also, such a relaxation of purifying selection after gene duplication may harm the genes associated with critical functions essential for survival.

In our study, we first focused on the duplication of human genes that are essential for human survival and/or reproduction. These genes represent the most vital portion of the human genome. Previous studies with the essential genes of model organisms revealed that these genes usually ‘avoid’ duplication, and such a trend is consistent across diverse group of organisms like Fungi (Yeast), Plants (Arabidopsis), Nematodes (Caenorhabditis) and Rodent mammals (Mouse). The essential genes usually remain highly connected in protein-protein interaction network. Thus, their duplication creates a dosage imbalance in the network, as the duplicated gene produces more protein products relative to its interacting partners in the PPI network. However, duplication of such genes associated with essential functions provides increased robustness against deleterious mutations and thus, maintaining essential genes as duplicates lead to insignificant fitness reduction of the organism upon the accumulation of deleterious mutation(s) on those genes. Thus, for the organisms that maintain essential genes as duplicates will have the upper hand against gene deletion, but the PPI network should kept dosage-balanced.

Our study reveals a higher proportion of essential genes in human duplicated genes, a trend that is different from a wide-range of organisms. The in-depth functional analysis reveals that these human essential duplicate pairs are functionally more similar to each other than those in the most popular mammalian mouse model. Such duplicates are associated with a higher number of potential micro-RNA target sites, indicating they may regulate their dosage-balance in the protein-protein interaction network. Furthermore, the human essential duplicates are evolutionarily more conserved than that in mouse, revealing that they have a higher efficiency to

act as backup copies, leading to increased robustness against deleterious mutation in humans.

We are the first to compare the functional genomic attributes of human small-scale and whole-genome duplicates (SSDs and WGDs, respectively). The WGDs present within the human genome had their origin long time ago in evolutionary scale, during the evolution of early vertebrates (~530 Mya). However, such WGDs have also occurred in different lineages, being most predominant in the evolution of land plants. Duplicated genes originated during the WGD event in the yeast genome (~150 Mya) were characterized and compared to the yeast SSDs. However, yeast WGD-event has occurred quite recently in evolutionary time scale compared to that in the vertebrates. Thus, we hypothesized that vertebrate WGD event must have played an important role in their diversification into such a variety of organisms. Our comparative analysis of human SSDs and WGDs revealed a lot of differences in their evolutionary genomic properties. The human WGDs originated during vertebrate WGD event are adapted to new functions and new locations, a trend being different from that in yeast. Such human WGDs performs more functions and are engaged in the most vital functions than the SSDs. They are also evolutionarily conserved, indicating their functional significance in humans. Together, our results establish human WGDs as the most important counterparts, that plays crucial roles within the human genome. Thus, our studies focus on the significance of human gene duplication and reveal the roles of duplicated copies generated by this process in human evolution, with a detailed account of their functional association and the importance of such functions.

References

- Acharya, D, D Mukherjee, S Podder, TC Ghosh. 2015. Investigating different duplication pattern of essential genes in mouse and human. PLoS One 10:e0120784-e0120784.
- Adams, KL, JF Wendel. 2005a. Polyploidy and genome evolution in plants. Curr Opin Plant Biol. 8:135-141.
- Allendorf, FW, GH Thorgaard. 1984. Tetraploidy and the evolution of salmonid fishes. In: BJ Turner, editor. Evolutionary Genetics of Fishes: Plenum Press, New York. p. 1-53.
- Barabasi, AL, ZN Oltvai. 2004. Network biology: Understanding the cell's functional organization. Nat Rev Genet. 5:101-U115.
- Baudot, A, B Jacq, C Brun. 2004. A scale of functional divergence for yeast duplicated genes revealed from analysis of the protein-protein interaction network. Genome Biol. 5.
- Begum, T, TC Ghosh. 2010. Understanding the Effect of Secondary Structures and Aggregation on Human Protein Folding Class Evolution. J Mol Evol. 71:60-69.

- Begum, T, TC Ghosh. 2014. Elucidating the genotype-phenotype relationships and network perturbations of human shared & specific disease genes from an evolutionary perspective. *Genome Biol Evol.* 6(10), 2741-2753.
- Bhattacharya, T, TC Ghosh. 2010. Protein Connectivity and Protein Complexity Promotes Human Gene Duplicability in a Mutually Exclusive Manner. *DNA Res.* 17:261-270.
- Birchler, JA, RA Veitia. 2007. The gene balance hypothesis: From classical genetics to modern genomics. *Plant Cell.* 19:395-402.
- Blanc, G, A Barakat, R Guyot, R Cooke, I Delseny. 2000. Extensive duplication and reshuffling in the arabidopsis genome. *Plant Cell.* 12:1093-1101.
- Blomen, VA, P Majek, LT Jae, *et al.* 2015. Gene essentiality and synthetic lethality in haploid human cells. *Science.* 350:1092-1096.
- Bowring, SA, IS Williams. 1999. Priscoan (4.00–4.03 Ga) orthogneisses from northwestern Canada. *Contrib Mineral Petrol.* 134:3-16.
- Brunet, FG, HR Crollius, M Paris, J-M Aury, P Gibert, O Jaillon, V Laudet, M Robinson-Rechavi. 2006a. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol.* 23:1808-1816.
- Brunet, FG, HR Crollius, M Paris, J-M Aury, P Gibert, O Jaillon, V Laudet, M Robinson-Rechavi. 2006b. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol.* 23:1808-1816.
- Byun-McKay, SA, R Geeta. 2007. Protein subcellular relocalization: a new perspective on the origin of novel genes. *Trends Ecol Evol.* 22:338-344.
- Chakraborty, S, TC Ghosh. 2013. Evolutionary Rate Heterogeneity of Core and Attachment Proteins in Yeast Protein Complexes. *Genome Biol Evol.* 5:1366-1375.

- Chang, AY-F, B-Y Liao. 2012. DNA Methylation Rebalances Gene Dosage after Mammalian Gene Duplications. *Mol Biol Evol.* 29:133-144.
- Chargaff, E, R Lipshitz, C Green. 1952. Composition of the desoxypentose nucleic acids of four genera of sea-urchin. *J Biol Chem.* 195:155-160.
- Chen, FC, BY Liao, CL Pan, HY Lin, AY Chang. 2012a. Assessing determinants of exonic evolutionary rates in mammals. *Mol Biol Evol.* 29:3121-3129.
- Chen, S, YE Zhang, M Long. 2010. New Genes in *Drosophila* Quickly Become Essential. *Science.* 330:1682-1685.
- Chen, W-H, P Minguez, MJ Lercher, P Bork. 2012b. OGEE: an online gene essentiality database. *Nucleic Acids Res.* 40:D901-D906.
- Chen, W-H, K Trachana, MJ Lercher, P Bork. 2012c. Younger Genes Are Less Likely to Be Essential than Older Genes, and Duplicates Are Less Likely to Be Essential than Singletons of the Same Age. *Mol Biol Evol.* 29:1703-1706.
- Chen, W-H, X-M Zhao, V van Noort, P Bork. 2013. Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs. *PLoS Comput Biol.* 9(5): e1003073.
- Clark, AG. 1994. INVASION AND MAINTENANCE OF A GENE DUPLICATION. *Proc Natl Acad Sci USA.* 91:2950-2954.
- Colbourne, JK, ME Pfrender, D Gilbert, WK Thomas, A Tucker, TH Oakley, S Tokishita, A Aerts, GJ Arnold, MK Basu. 2011. The ecoresponsive genome of *Daphnia pulex*. *Science.* 331:555-561.
- Crick, F. 1970. Central dogma of molecular biology. *Nature.* 227:561-563.
- Cullen, LM, GM Arndt. 2005. Genome-wide screening for gene function using RNAi in mammalian cells. *Immunol Cell Biol.* 83:217-223.
- Cunningham, F, MR Amode, D Barrell, *et al.* 2014. Ensembl 2015. *Nucleic Acids Res.* 43 (Database issue): D662-669.

- D'Antonio, M, FD Ciccarelli. 2011. Modification of Gene Duplicability during the Evolution of Protein Interaction Network. *PLoS Comput Biol.* 7(4):e1002029.
- Dalrymple, GB. 2001. The age of the Earth in the twentieth century: a problem (mostly) solved. *Geological Society, London, Special Publications* 190:205-221.
- Dean, EJ, JC Davis, RW Davis, DA Petrov. 2008. Pervasive and Persistent Redundancy among Duplicated Genes in Yeast. *PLoS Genet.* 4(7):e1000113.
- Dehal, P, JL Boore. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* 3:1700-1708.
- Dickerson, JE, DL Robertson. 2012. On the origins of Mendelian disease genes in man: the impact of gene duplication. *Mol Biol Evol.* 29:61-69.
- Dujon, B, D Sherman, G Fischer, *et al.* 2004. Genome evolution in yeasts. *Nature.* 430:35-44.
- Emanuelsson, O, G Heijne. 2001. Prediction of organellar targeting signals. *BBA-Mol Cell Res.* 1541(1-2):114-119.
- Fares, MA, OM Keane, C Toft, L Carretero-Paulet, GW Jones. 2013. The Roles of Whole-Genome and Small-Scale Duplications in the Functional Specialization of *Saccharomyces cerevisiae* Genes. *PLoS Genet.* 9(1):e1003176.
- Finn, RD, A Bateman, J Clements, *et al.* 2014. Pfam: the protein families database. *Nucleic Acids Res.* 42:D222-D230.
- Flicek, P, I Ahmed, MR Amode, *et al.* 2013. Ensembl 2013. *Nucleic Acids Res.* 41:D48-D55.
- Flicek, P, MR Amode, D Barrell, *et al.* 2014. Ensembl 2014. *Nucleic Acids Res.* 42:D749-D755.

- Force, A, M Lynch, FB Pickett, A Amores, YL Yan, J Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*. 151:1531-1545.
- Forslund, K, F Schreiber, N Thanintorn, ELL Sonnhammer. 2011. OrthoDisease: tracking disease gene orthologs across 100 species. *Brief Bioinform*. 12:463-473.
- Fraser, CM, JD Gocayne, O White, MD Adams, RA Clayton, RD Fleischmann, CJ Bult, AR Kerlavage, G Sutton, JM Kelley. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science*. 270:397-404.
- Freeling, M, BC Thomas. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res*. 16:805-814.
- Ganko, EW, BC Meyers, TJ Vision. 2007. Divergence in expression between duplicated genes in *Arabidopsis*. *Mol Biol Evol*. 24:2298-2309.
- Garcia, DM, D Baek, C Shin, GW Bell, A Grimson, DP Bartel. 2011. Weak seed-pairing stability and high target-site abundance decrease the proficiency of *Isy-6* and other microRNAs. *Nat Struct Mol Biol*. 18:1139-U1175.
- Gene Ontology, C. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. 32:D258-D261.
- Georgi, B, BF Voight, M Bucan. 2013. From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. *PLoS Genet*. 9(5):e1003484.
- Giaever, G, AM Chu, L Ni, *et al*. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*. 418:387-391.
- Goh, K-I, ME Cusick, D Valle, B Childs, M Vidal, A-L Barabasi. 2007. The human disease network. *Proc Natl Acad Sci USA*. 104:8685-8690.

- Gu, ZL, LM Steinmetz, X Gu, C Scharfe, RW Davis, WH Li. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature*. 421:63-66.
- Guan, Y, MJ Dunham, OG Troyanskaya. 2007. Functional analysis of gene duplications in *Saccharomyces cerevisiae*. *Genetics*. 175:933-943.
- Ha, M, E-D Kim, ZJ Chen. 2009. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proc Natl Acad Sci USA*. 106:2295-2300.
- Hakes, L, JW Pinney, SC Lovell, SG Oliver, DL Robertson. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol*. 8(10):R209.
- He, X, J Zhang. 2006a. Why do hubs tend to be essential in protein networks? *PLoS Genet*. 2:826-834.
- He, XL, JZ Zhang. 2006b. Higher duplicability of less important genes in yeast genomes. *Mol Biol Evol*. 23:144-151.
- Hsiao, T-L, D Vitkup. 2008. Role of duplicate genes in robustness against deleterious human mutations. *PLoS Genet*. 4:e1000014-e1000014.
- Hwang, Y-C, C-C Lin, J-Y Chang, H Mori, H-F Juan, H-C Huang. 2009. Predicting essential genes based on network and sequence analysis. *Mol. BioSyst*. 5:1672-1678.
- Ihaka, R, R Gentleman. 1996. R: a language for data analysis and graphics. *J Comput Graph Stat*. 5:299-314.
- Innan, H, F Kondrashov. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet*. 11:97-108.
- Jeong, H, SP Mason, AL Barabasi, ZN Oltvai. 2001. Lethality and centrality in protein networks. *Nature*. 411:41-42.

- Jordan, IK, IB Rogozin, YI Wolf, EV Koonin. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* 12:962-968.
- Jordan, IK, YI Wolf, EV Koonin. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. *Bmc Evolutionary Biology* 4.
- Juhas, M, L Eberl, JI Glass. 2011. Essence of life: essential genes of minimal genomes. *Trends cell Biol.* 21:562-568.
- Kamath, RS, AG Fraser, Y Dong, *et al.* 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature.* 421:231-237.
- Kapushesky, M, T Adamusiak, T Burdett, *et al.* 2012. Gene Expression Atlas update—a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 40:D1077-D1081.
- Kasahara, M. 2007. The 2R hypothesis: an update. *Curr Opin Immunol.* 19:547-552.
- Kellis, M, BW Birren, ES Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature.* 428:617-624.
- Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature.* 217:624-626.
- Kondrashov, FA, AS Kondrashov. 2006. Role of selection in fixation of gene duplications. *J Theor Biol.* 239:141-151.
- Leach, LJ, Z Zhang, C Lu, MJ Kearsey, Z Luo. 2007. The role of Cis-Regulatory motifs and genetical control of expression in the divergence of yeast duplicate genes. *Mol Biol Evol.* 24:2556-2565.
- Li, J, G Musso, Z Zhang. 2008. Preferential regulation of duplicated genes by microRNAs in mammals. *Genome Biol.* 9.

- Li, WH, J Yang, X Gu. 2005. Expression divergence between duplicate genes. *Trends Genet.* 21:602-607.
- Li, Z, H Zhang, S Ge, X Gu, G Gao, J Luo. 2009. Expression pattern divergence of duplicated genes in rice. *BMC Bioinformatics.* 10(6):S8.
- Liang, H, W-H Li. 2007. Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet.* 23:375-378.
- Liang, H, W-H Li. 2009. Functional compensation by duplicated genes in mouse. *Trends Genet.* 25:441-442.
- Liao, B-Y, NM Scott, J Zhang. 2006. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol Biol Evol.* 23:2072-2080.
- Liao, B-Y, J Zhang. 2007. Mouse duplicate genes are as essential as singletons. *Trends Genet.* 23:378-381.
- Liao, B-Y, J Zhang. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci USA.* 105:6987-6992.
- Lynch, M, JS Conery. 2000. The evolutionary fate and consequences of duplicate genes. *Science.* 290:1151-1155.
- Lynch, M, A Force. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics.* 154:459-473.
- Makino, T, K Hokamp, A McLysaght. 2009. The complex relationship of gene duplication and essentiality. *Trends Genet.* 25:152-155.
- Makino, T, A McLysaght. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci USA.* 107:9270-9274.
- Makino, T, Y Suzuki, T Gojobori. 2006. Differential evolutionary rates of duplicated genes in protein interaction network. *Gene.* 385:57-63.

- Malaguti, G, PP Singh, H Isambert. 2014. On the retention of gene duplicates prone to dominant deleterious mutations. *Theor Popul Biol.* 93:38-51.
- Marques, AC, N Vinckenbosh, D Brawand, H Kaessmann. 2008. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. *Genome Biol.* 9:R54.
- McLysaght, A, K Hokamp, KH Wolfe. 2002. Extensive genomic duplication during early chordate evolution. *Nat Genet.* 31:200-204.
- McLysaght, A, T Makino, HM Grayton, M Tropeano, KJ Mitchell, E Vassos, Da Collier. 2014. Ohnologs are overrepresented in pathogenic copy number mutations. *Proc Natl Acad Sci USA.* 111:361-366.
- Nakatani, Y, H Takeda, Y Kohara, S Morishita. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* 17:1254-1265.
- Neme, R, D Tautz. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics.* 14.
- Nie, N, D Bent, C Hull. 1970. SPSS: statistical package for the social sciences. New York: McGraw-Hill.
- Ohno, S. 1970. *Evolution by Gene Duplication*: Springer.
- Ohno, S, U Wolf, NB Atkin. 1968. Evolution from fish to mammals by gene duplication. *Hereditas.* 59:169-187.
- Papp, B, C Pal, LD Hurst. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature.* 424:194-197.
- Petryszak, R, T Burdett, B Fiorelli, *et al.* 2014. Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 42:D926-D932.

- Pierce, BA. 2012. Genetics: A Conceptual Approach. WH Freeman and Company.
- Podder, S, TC Ghosh. 2011. Insights into the molecular correlates modulating functional compensation between monogenic and polygenic disease gene duplicates in human. *Genomics*. 97:200-204.
- Podder, S, P Mukhopadhyay, TC Ghosh. 2009. Multifunctionality dominantly determines the rate of human housekeeping and tissue specific interacting protein evolution. *Gene*. 439:11-16.
- Qian, W, B-Y Liao, AY-F Chang, J Zhang. 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet*. 26:425-430.
- Qian, W, J Zhang. 2009. Protein Subcellular Relocalization in the Evolution of Yeast Singleton and Duplicate Genes. *Genome Biol Evol*. 1:198-204.
- Robinson-Rechavi, M, V Laudet. 2001. Evolutionary rates of duplicate genes in fish and mammals. *Mol Biol Evol*. 18:681-683.
- Roemer, T, B Jiang, J Davison, *et al*. 2003. Large-scale essential gene identification in *Candida albicans* and applications to antifungal drug discovery. *Mol Microbiol*. 50:167-181.
- Salathe, M, M Ackermann, S Bonhoeffer. 2006. The effect of multifunctionality on the rate of evolution in yeast. *Mol Biol Evol*. 23:721-722.
- Satake, M, M Kawata, A McLysaght, T Makino. 2012. Evolution of vertebrate tissues driven by differential modes of gene duplication. *DNA Res*. 19(4):305-316.
- Seringhaus, M, A Paccanaro, A Borneman, M Snyder, M Gerstein. 2006. Predicting essential genes in fungal genomes. *Genome Res*. 16:1126-1135.

- Silva, JM, K Marran, JS Parker, J Silva, M Golding, MR Schlabach, SJ Elledge, GJ Hannon, K Chang. 2008. Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science*. 319:617-620.
- Singh, PP, S Affeldt, I Cascone, R Selimoglu, J Camonis, H Isambert. 2012. On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. *Cell Rep*. 2:1387-1398.
- Singh, PP, J Arora, H Isambert. 2015a. Identification of ohnolog genes originating from whole genome duplication in early vertebrates, based on synteny comparison across multiple genomes. *PLoS Comput Biol*. 11:e1004394.
- Springer, M, JS Weissman, MW Kirschner. 2010. A general lack of compensation for gene dosage in yeast. *Mol Syst Biol*. 6:368.
- Stebbins, GL. 1971. *Chromosomal Evolution in Higher Plants*: Addison-Wesley.
- Stenson, PD, EV Ball, M Mort, AD Phillips, K Shaw, DN Cooper. 2012. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Current protocols in bioinformatics / editorial board, Andreas D. Baxeavanis ... [et al.] Chapter 1:Unit1.13-Unit11.13*.
- Stephens, SG. 1951. Possible significances of duplication in evolution. *Adv Genet*. 4:247-265.
- Su, Z, X Gu. 2008. Predicting the Proportion of Essential Genes in Mouse Duplicates Based on Biased Mouse Knockout Genes. *J Mol Evol*. 67:705-709.
- Taylor, JS, J Raes. 2004. Duplication and divergence: The evolution of new genes and old ideas. *Annu Rev Genet*. 38:615-643.

- Teshima, KM, H Innan. 2008. Neofunctionalization of duplicated genes under the pressure of gene conversion. *Genetics*. 178:1385-1398.
- Uhlen, M, E Bjorling, C Agaton, *et al*. 2005. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics*. 4:1920-1932.
- Uhlén, M, L Fagerberg, BM Hallström, *et al*. 2015. Tissue-based map of the human proteome. *Science*. 347:1260419.
- Vandepoele, K, W De Vos, JS Taylor, A Meyer, Y Van de Peer. 2004. Major events in the genome evolution of vertebrates: Paraneome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci USA*. 101:1638-1643.
- Veitia, Ra, S Bottani, Ja Birchler. 2013. Gene dosage effects: Nonlinearities, genetic interactions, and dosage compensation. *Trends Genet*. 29:385-393.
- Via, S, R Lande. 1985. Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution*. 39:505-522.
- Wagner, A. 2008. Gene duplications, robustness and evolutionary innovations. *Bioessays*. 30:367-373.
- Wall, DP, AE Hirsh, HB Fraser, J Kumm, G Giaever, MB Eisen. 2005. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci USA*. 102.
- Wang, J, W Peng, F-X Wu. 2013. Computational approaches to predicting essential proteins: A survey. *Proteomics Clin Appl*. 7:181-192.
- Waterhouse, RM, EM Zdobnov, EV Kriventseva. 2011. Correlating Traits of Gene Retention, Sequence Divergence, Duplicability and Essentiality in Vertebrates, Arthropods, and Fungi. *Genome Biol Evol*. 3:75-86.

- Watson, JD, TA Baker, SP Bell, A Gann, M Levine, R Losick. 2014. Molecular Biology of the Gene: Pearson
- Wendel, JF. 2000. Genome evolution in polyploids. *Plant Mol Biol.* 42:225-249.
- Wilde, SA, JW Valley, WH Peck, CM Graham. 2001. Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago. *Nature.* 409:175.
- Wolf, YI, PS Novichkov, GP Karev, EV Koonin, DJ Lipman. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci USA.* 106:7273-7280.
- Wolfe, KH, DC Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature.* 387:708-713.
- Yang, J, ZL Gu, WH Li. 2003. Rate of protein evolution versus fitness effect of gene deletion. *Mol Biol Evol.* 20:772-774.
- Yang, J, R Lusk, WH Li. 2003. Organismal complexity, protein complexity, and gene duplicability. *Proc Natl Acad Sci USA.* 100:15661-15665.
- Yang, L, J Wang, H Wang, Y Lv, Y Zuo, X Li, W Jiang. 2014. Analysis and identification of essential genes in humans using topological properties and biological information. *Gene.* 551:138-151.
- Yates, A, W Akanni, MR Amode, *et al.* 2016. Ensembl 2016. *Nucleic Acids Res.* 44:D710-D716.
- Zhang, JZ. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol.* 18:292-298.
- Zhou, RJ, HH Cheng, TR Tiersch. 2001. Differential genome duplication and fish diversity. *Rev Fish Biol Fisher.* 11:331-337.
- Zuckerkandl, E, L Pauling. 1965. Evolutionary divergence and convergence in proteins. *Evolving genes and proteins* 97:97-166.

Publications

Debarun Acharya, Dola Mukherjee, Soumita Podder, Tapash C Ghosh: *Investigating Different Duplication Pattern of Essential Genes in Mouse and Human*. PLoS ONE 03/2015; 10(3):e0120784., DOI:10.1371/journal.pone.0120784

Debarun Acharya, Tapash C. Ghosh: *Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution*. BMC Genomics 12/2016; 17(1), DOI:10.1186/s12864-016-2392-0

Unrelated to thesis work:

Arup Panda, Debarun Acharya, Tapash Chandra Ghosh: *Insights into human intrinsically disordered proteins from their gene expression profile*. Molecular BioSystems 10/2017; 13(12), DOI:10.1039/C7MB00311K

Kakali Biswas, Debarun Acharya, Soumita Podder, Tapash Chandra Ghosh: *Evolutionary rate heterogeneity between multi- and single-interface hubs across human housekeeping and tissue-specific protein interaction network: Insights from proteins' and its partners' properties*. Genomics (in press) 12/2017;, DOI:10.1016/j.ygeno.2017.11.006

Dola Mukherjee, Deeya Saha, Debarun Acharya, Ashutosh Mukherjee, Sandip Chakraborty, Tapash Chandra Ghosh: *The role of introns in the conservation of the metabolic genes of Arabidopsis thaliana*. Genomics (in press) 12/2017;, DOI:10.1016/j.ygeno.2017.12.003

Reprints

RESEARCH ARTICLE

Investigating Different Duplication Pattern of Essential Genes in Mouse and Human

Debarun Acharya, Dola Mukherjee, Soumita Podder, Tapash C. Ghosh*

Bioinformatics Centre, Bose Institute, Kolkata, West Bengal, India

* tapash@jcbose.ac.in



OPEN ACCESS

Citation: Acharya D, Mukherjee D, Podder S, Ghosh TC (2015) Investigating Different Duplication Pattern of Essential Genes in Mouse and Human. PLoS ONE 10(3): e0120784. doi:10.1371/journal.pone.0120784

Received: September 18, 2014

Accepted: January 27, 2015

Published: March 9, 2015

Copyright: © 2015 Acharya et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All the data used in the experiments are freely available in the paper, the supplemental files, and in well-recognized public repositories. All "gene essentiality, gene duplication, developmental genes and phyletic age data of mouse and human" are available from the Online Gene Essentiality Database (URL- <http://ogeedb.embl.de>). The dataset is also provided as a supplemental file S1_Dataset.xlsx. The duplicate pairs of mouse and human genes under study is provided in supplemental file S2_Dataset.xlsx. All Gene Ontology Annotation for mouse and human are available from the Ensembl biomart interface (Release 71) (URL- <http://www.ensembl.org/biomart/martview>). Gene biotype data for Pseudogenization for mouse and human are available from the Ensembl biomart interface (Release 71) (URL- <http://www.ensembl.org/>)

Abstract

Gene duplication is one of the major driving forces shaping genome and organism evolution and thought to be itself regulated by some intrinsic properties of the gene. Comparing the essential genes among mouse and human, we observed that the essential genes avoid duplication in mouse while prefer to remain duplicated in humans. In this study, we wanted to explore the reasons behind such differences in gene essentiality by cross-species comparison of human and mouse. Moreover, we examined essential genes that are duplicated in humans are functionally more redundant than that in mouse. The proportion of paralog pseudogenization of essential genes is higher in mouse than that of humans. These duplicates of essential genes are under stringent dosage regulation in human than in mouse. We also observed slower evolutionary rate in the paralogs of human essential genes than the mouse counterpart. Together, these results clearly indicate that human essential genes are retained as duplicates to serve as backed up copies that may shield themselves from harmful mutations.

Introduction

Gene duplication was thought to be one of the major driving factors stimulating genome and organism evolution [1–4], as it provides raw genetic materials for structural and functional modification and at the same time conserves the parental function. Although, gene duplication is not always beneficial, and most duplicates become subsequently inactivated or pseudogenized in the genome [4], it may have many implications in an organism's life. For example, the duplicates may be maintained in the genome for its immediate benefit to the organism, like increased gene dosage [5] or serve as backup copies to restore the function if the original one becomes deleted [6,7]. Apart from this, the duplicates may undergo modifications to take up novel functions, i.e. neofunctionalization [4], or they may share their function after complementary degenerative mutations, i.e. subfunctionalization [8,9]. The pattern of gene duplication may vary between species and also across different groups of genes within the same species. Several factors contributing gene duplication has been observed till date in diverse organisms like protein connectivity and protein interaction network [10–12], protein complexity [13,14], gene retention and sequence divergence [15], dosage balance [16] and nevertheless, gene essentiality [17–19].

biomart/martview). Nonsynonymous nucleotide substitution per nonsynonymous sites (dN) and synonymous nucleotide substitution per synonymous sites (dS) for mouse and human with corresponding one-to-one rat orthologs are available from the Ensembl biomart interface (Release 71) (URL-<http://www.ensembl.org/biomart/martview>). All micro-RNA target sites for mouse and human were obtained from TargetScan Release 6.2 (<http://www.targetscan.org>). In the case of any query, the readers may contact Mr. Debarun Acharya (e-mail: debarun@jcbosc.ac.in).

Funding: Funding from University Grants Commission (UGC) Sanction Letter No.F.2-8/2002 (SA-I) dated 04.10.2012, received by DA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Essential genes are indispensable to an organism and cause severe reduction in its fitness like sterility or lethality upon deletion [20]. These genes are mainly associated with important biological functions. However, many expressed genes performing such functions are considered to be nonessential, as their deletion can be compensated by other genes having similar or identical functions and expression [21]. Gene duplication is an important mechanism for such functional redundancy to occur [4]. Now, there may be two kinds of possibilities for essential genes to prefer or avoid the course of gene duplication. First, essential genes are required to become duplicated for providing backup copies that could shield themselves from any harmful mutations; secondly from evolutionary standpoint, essential genes may prefer to stay away from gene duplication since ectopic recombination and replication driven gene duplication may increase the chances of mutational load which is not at all acceptable for essential genes for being the most conserved gene-group [22,23].

Gene essentiality was widely studied across model organisms and shown to bear a complex relationship with gene duplication [19]. In lower eukaryotes like yeast, a higher proportion of essential genes were observed in singletons than in duplicates [7]. However, studies with mouse showed that the proportion of essential genes in duplicates are comparable to that in singletons [10,18]. Additionally, two follow-up studies with mouse also report that the proportion of essential genes is higher in singletons than in duplicates [21,24].

Till date, all the studies regarding essential genes were carried in yeast and mouse due to unavailability of human gene essentiality data. In a previous study, researchers attempted to explore the properties of human orthologs of mouse essential genes [25]. However, considering such human orthologs as essential may not be accurate [26]. Taking advantage of the Online Gene Essentiality (OGEE) database that represents a valuable resource of human and mouse essential genes, we performed a comprehensive analysis comparing duplication pattern of essential genes in human and mouse. We noticed that in mouse, the essential genes prefer to remain as singleton whereas the trend is reverse for human, which is unexplored so far. We have also explored the underlying reasons and the benefits of maintaining essential genes as duplicates in humans.

Materials and Methods

Gene Essentiality and Gene Duplication

Gene essentiality and duplication of human (*Homo sapiens*) and mouse (*Mus musculus*) were obtained from the Online Gene Essentiality (OGEE) database (<http://ogeedb.embl.de>) [27] (S1 Dataset). The paralog lists for human and mouse essential genes were provided by the authors of OGEE database [27] (S2 Dataset).

Developmental Genes

The developmental genes for mouse and human were obtained from Online Gene Essentiality (OGEE) database [27] (S1 Dataset). Here, a gene is considered as developmental if they are associated with one of the two GO terms: GO:0007275 (multicellular organismal development) and GO:0030154 (cell differentiation) or their daughter terms, and others as non-developmental, a method adapted by Makino et al. 2009 [19].

Phyletic Age and Overall Proportion of Essentiality

Phyletic origin of a gene can be defined as the most distance group of organisms where the homologs (orthologs) of that gene are present. The phyletic age of human and mouse genes was obtained from the Online Gene Essentiality (OGEE) database [27], where the authors used the

phyletic age prediction algorithm described by Wolf et al. [28]. The genes were divided in seven classes according to their evolutionary origin, namely 0 (not assigned), 1 (Mammalia), 2 (Chordata), 3 (Metazoa), 4 (Fungi/Metazoa group), 5 (Eukaryota) and 6 (cellular organisms). We discarded the first group in which the phyletic age was not assigned and selected the rest from mouse and human OGEE genes. We obtained the final mouse and human data with gene essentiality, gene duplication and phyletic age information containing 5869 and 18400 genes, respectively. We divided the human and mouse OGEE genes into two groups depending on their phyletic age: the ‘old duplicates’ (containing three older classes) and ‘new duplicates’ (containing the rest three classes) in both human and mouse (S1 Dataset). From this data, we calculated the overall proportion of essential genes in singletons and duplicates for both species as a weighted average using this formula [21]:

$$P_E = f_{old} \times P_E^{old} + f_{young} \times P_E^{young}$$

Where, f_{old} and f_{young} are the fraction of old and young genes contained in the gene group and the P_E^{old} and P_E^{young} are proportion of essential genes in old and young counterparts. Using this formula, we calculated the proportion of essential genes in singleton and duplicates for both species irrespective of their age bias.

Functional Distance

The functional distance for the human and mouse essential genes carried by the Gene Ontology (GO) annotations was calculated using the GO domain molecular function for essential genes and their paralogous copies of corresponding species from Ensembl 71 biomart interface (<http://www.ensembl.org/biomart/martview>) [29]. The GO terms for each human and mouse essential gene and the corresponding paralogous genes were calculated separately. Using the Czekanowski—Dice distance formula [30] mentioned below, we calculated the functional divergence for each human and mouse essential genes with their paralogous counterparts.

$$\text{Functional distance } (i, j) = \frac{\text{Number of Terms}(i) \Delta \text{Terms}(j)}{[\text{Number of } (\text{Terms}(i) \cup \text{Terms}(j)) + \text{Number of } (\text{Terms}(i) \cap \text{Terms}(j))]}$$

In which, i and j denote a gene and its paralogous gene within a species. Terms (i) and Terms (j) are the lists of the GO terms for individual genes. ‘ \cup ’ and ‘ \cap ’ denotes the nonredundant and common GO id count, respectively, of the two genes. ‘ Δ ’ is the symmetrical difference between the GO term sets of two genes, i.e. ‘ $(\cup - \cap)$ ’.

Although the Czekanowski-Dice distance formula is the most commonly used method for calculation of functional distance, it is sensitive to the number of GO terms per gene and therefore may be erroneous for cross-species comparison. Therefore, to compare the functional distance between mouse and human essential genes using the Czekanowski-Dice formula, we must consider the number of GO terms associated with the genes. To ensure that, we binned our functional distance data of the two species in three groups: Group A (with GO terms 1 to 4; $N_{\text{human}} = 367$, $N_{\text{mouse}} = 773$), Group B (with GO terms 5 to 8; $N_{\text{human}} = 343$, $N_{\text{mouse}} = 485$) and Group C (with GO terms > 8 ; $N_{\text{human}} = 244$, $N_{\text{mouse}} = 278$) and compared the functional distance of human and mouse essential genes within each group.

Pseudogenization

Mouse and human pseudogenes were obtained from the biomart interface of ensemble 71 (<http://www.ensembl.org/biomart/martview>) [29]. For both the species, we searched for the

gene IDs for which the gene biotype contains the term ‘pseudogene’. This includes pseudogene, IG-V-pseudogene, TR-V-pseudogene, polymorphic pseudogene, TR-J-pseudogene, IG-C-pseudogene, IG-J-pseudogene and processed pseudogene. We calculated the proportion of paralog pseudogenization by considering only the duplicated essential genes with at least one pseudogenized paralog. The proportion of paralog pseudogenization was calculated by the ratio of the number of pseudogenized paralogs and the total number of paralogs. The mouse and human essential genes with the biotype of the paralog are provided in [S3 Dataset](#).

Micro-RNA Target Sites

Average micro-RNA target sites for human and mouse were obtained from TargetScan Release 6.2 (<http://www.targetscan.org>) [31]. For each of the human and mouse essential genes having known paralogs, we made individual sets comprising the gene and all of its paralogs. We calculated the mean micro-RNA target sites of each of such sets for the two species. We considered the mean value of all sets within a species to obtain the mean micro-RNA target sites for that species.

Evolutionary Rate

Evolutionary rates of the human and mouse genes were calculated as the ratio of nonsynonymous nucleotide substitution per nonsynonymous sites (dN) and synonymous nucleotide substitution per synonymous sites (dS), from the biomart interface of ensemble 71 (<http://www.ensembl.org/biomart/martview>) [29], using rat (*Rattus norvegicus*) as an outgroup. We obtained the dN and dS of human and mouse genes from their corresponding one-to-one rat orthologs. We compared the dN/dS ratios of nonredundant sets of human and mouse essential genes’ paralogs.

Statistical Analyses

Statistical analyses of the entire work were performed using SPSS v.13 and in house PERL Script. Mann-Whitney U test was used in SPSS to compare the mean values of different variables between two classes of genes. We used our in house PERL Script to perform two-sample Z-test for comparing relative proportions of a variable between two gene groups.

Results and Discussions

We compared the duplication of human and mouse essential genes and noticed that the tendency of essential genes to remain as duplicate copy varies between human and mouse. In human, the proportion of essential genes is higher among the duplicated subsets compared to the singleton genes; whereas in mouse, the reverse was observed. We observed that in mouse among 2098 singleton genes, 994 genes are essential (47.38%) and among 3771 duplicated genes, 1563 genes are essential (41.45%) [$Z = 4.391$, confidence level 99%; $P < 0.0001$, two sample Z-test] whereas, in humans, among 7563 singleton genes, 486 genes exist as essential (6.43%) and among 10837 duplicated genes, 984 are essential (9.08%) [$Z = -6.523$, confidence level 99%; $P < 0.0001$, two sample Z-test]. The overall proportion of essentiality is higher in mouse, which may be due to the fidelity of the methods applied to detect essential genes [27] or the unavailability of the complete essentiality data, but within species (where the same method is used to detect essentiality), gene essentiality should contribute equally among singletons and duplicates, which is however, not the case, as our observations indicate a higher probability of retaining the essential genes as duplicates in humans but not in mouse. A previous study reported that developmental genes are more essential than non-developmental ones [19] and

their abundance may result higher essentiality for a particular gene group relative to other, which led us to hypothesise that the overrepresentation of developmental genes in a particular gene group may influence the overall trend. To explore if this is the case in our experiment, we discarded the developmental genes and calculated the proportion of essential genes in singleton and duplicate for human and mouse non-developmental genes only (see [materials and methods](#) for details). Here also, we obtained a similar trend ([Table 1](#)), which indicates that the results are not influenced by developmental genes. Therefore, we continued our study including both the developmental and nondevelopmental mouse and human genes.

Another possible bias in our dataset may arise due to the age of the duplicates. Previous studies showed that the genes originated from old duplications are more likely to be essential than singletons [24]. Therefore, the age of genes have an influence in gene essentiality, which may lead to overestimation of human essential genes as duplicates in our dataset as we have considered duplicates as the genes having at least one paralogous copy, no matter how ancient it is. This bias was corrected by considering the phyletic age of the genes to calculate the overall proportion of essentiality [21] (see [materials and methods](#)) in singleton and duplicated mouse and human genes. We did not consider the duplication age (the origin of most recent duplication event) as our dataset also contains singletons and hence, phyletic age will be a more suitable measure. After correcting the age bias, we still obtained the same trend in proportion of essential genes in singletons and duplicates in both species ([Table 2](#)).

Our study contradicted the previous study of Liao and Zhang [18] which entails that mouse singleton and duplicate genes have an equal proportion of essential genes. This may result from the difference in essential gene collection procedure followed in Mouse Genome Informatics (MGI) which they used and OGEE databases which we have used. However, our result of mouse genes essentiality is consistent with that shown by two more recent studies [21,24]. Thus, with no further controversy, we wanted to comprehend why essential genes prefer to remain as duplicates in humans. Firstly, we contemplated that human genes may be maintained to keep an extra copy for functional compensation. However, the higher connectivity (Hub like nature) of essential genes which was revealed in many previous studies [32–35] demands a stringent regulation, in order to maintain the whole protein interaction network dosage-balanced. Moreover, duplication leading to the increase in dosage may not be favourable and, as a result, duplicates must either be diversified [36] or kept silent (dosage-balanced) [16].

To investigate whether the diversification supports the fixation of duplicate copies of essential genes in the human genome, or the duplicates are maintained as a backup system under stringent dosage-regulatory mechanism, we compared the essential genes and their paralogs between mouse and humans.

Firstly, we wanted to explore if the essential genes are duplicated for becoming functionally diversified and fixed in the genome. For this, we considered GO annotations for each human and mouse essential genes and their corresponding paralogous copies from Ensembl 71 bio-mart interface [29] for the GO domain Molecular function. Using the Czekanowski—Dice distance formula [30] (see [materials and methods](#)), we have obtained a significantly lower ($P = 3.73 \times 10^{-6}$, Mann-Whitney U test) functional distance value in human duplicated essential genes (Average functional distance = 0.340, $N = 954$) than in mouse duplicated essential genes (Average functional distance = 0.385, $N = 1536$). However, the Czekanowski—Dice distance formula we used here is sensitive to the number of go terms associated with a gene, which may vary from species to species. Therefore, for an unbiased cross-species comparison of functional distance, we binned our dataset into three groups containing according to their go id count (see [materials and methods](#)). We observed a significantly lower functional distance in human essential genes than the mouse counterparts in all three groups [[Fig. 1](#)], suggesting a tendency of retaining the human duplicated copies of essential genes *per se* as backup copies.

Table 1. Proportion of essential genes among singleton and duplicates of mouse and human non-developmental genes.

Species	Gene group	Total genes	Essential genes	Proportion of essential genes	Z-score and P value
Mouse (<i>Mus musculus</i>)	Singleton	1237	462	37.348	Z = 5.0323 (Confidence level 99%) P < 0.0001
	Duplicates	2301	669	29.074	
Human (<i>Homo sapiens</i>)	Singleton	6347	332	5.231	Z = -3.7168 (Confidence level 99%) P = 0.0002
	Duplicates	8581	575	6.701	

doi:10.1371/journal.pone.0120784.t001

Although we observed that human essential duplicates are functionally less diverged than mouse, we were curious to understand the occurrence of pseudogenized paralogs among essential genes of both species. As our main dataset contains essential genes of human and mouse, no occurrence of pseudogene was observed. However, among the paralogs, we did not find any significant difference between mouse (0.82%) and human (0.50%) ($Z = -1.584$, $P = 1.13 \times 10^{-1}$, two sample Z-test), which may be due to the low proportion of pseudogene occurrence in both species ([S3 Dataset](#)). The low proportions of pseudogenes in our mouse and human essential genes' paralogs are normal as we are considering paralogs of the genes with crucial functions. However, when we considered the proportion of paralog pseudogenization for each human and mouse essential duplicate genes having at least one pseudogenized paralog (see [materials and methods](#)), the proportion of paralog pseudogenization were found to be lower in human essential genes than in the mouse counterpart (Proportion of paralog pseudogenization in mouse = 0.178, Proportion of paralog pseudogenization in human = 0.048; $P = 1.44 \times 10^{-7}$, Mann-Whitney U test, $N_{\text{mouse}} = 17$, $N_{\text{human}} = 63$). This result suggests that mouse essential genes' paralogs can become pseudogenized more easily. In other words, human essential genes retain their functionality more readily, which in turn can help them to serve as functional back-up copies, as we have previously shown that they are functionally more similar to their ancestral genes.

The human essential genes in our study were observed to show lower functional divergence. Thus, we hypothesize that the essential gene duplicates are functionally redundant and they may be maintained as backup copies. However, the maintenance of newly synthesized duplicates is very crucial and often performed by micro-RNA mediated post-transcriptional regulation, which may give support to the backed up essential genes by reducing their expression [[37](#)]. Therefore, to measure the ability to maintain the backed up duplicates, we measured the average micro-RNA target sites for mouse and human essential genes and their duplicates (see [materials and methods](#) for details). Consistent with our expectation, we observed a significantly higher ($P = 3.35 \times 10^{-6}$; Mann-Whitney U test) micro-RNA target sites in duplicated essential genes of human (Mean micro-RNA count 19.15, Number of sets = 742) than in mouse (Mean micro-RNA count 15.82, Number of sets = 1202), suggesting the robust regulation by micro-

Table 2. Proportion of essential genes as weighted average among singleton and duplicates of mouse and human.

Species	Gene group	Total genes	Proportion of essential genes as weighted average ($P_E = f_{\text{old}} \times P_E^{\text{old}} + f_{\text{young}} \times P_E^{\text{young}}$)	Z-score and P value
Mouse (<i>Mus musculus</i>)	Singleton	2098	47.379	Z = -4.392 (Confidence level 99%) P < 0.0001
	Duplicate	3771	41.448	
Human (<i>Homo sapiens</i>)	Singleton	7563	6.426	Z = -6.535 (Confidence level 99%) P < 0.0001
	Duplicate	10837	9.081	

doi:10.1371/journal.pone.0120784.t002

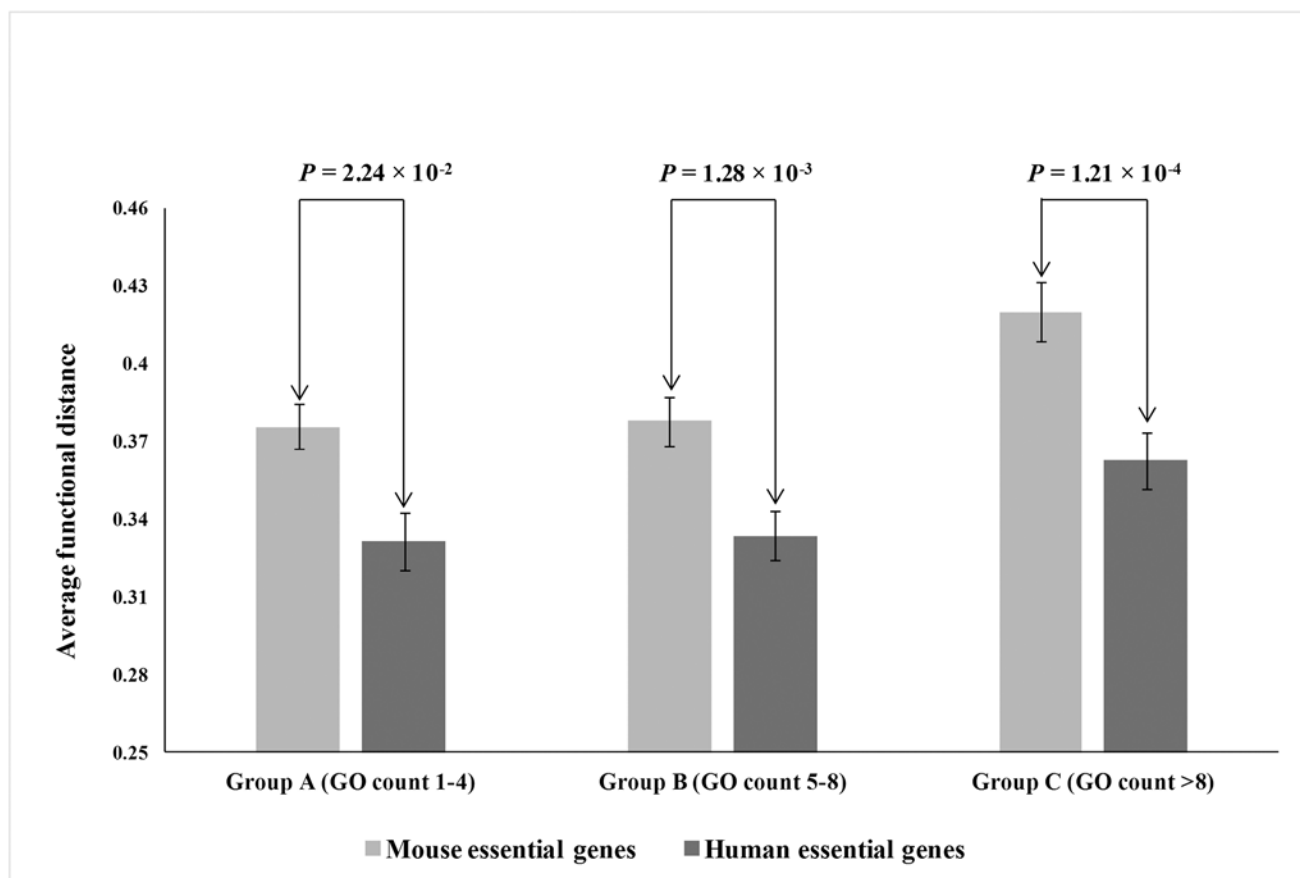


Fig 1. Average functional distance between mouse and human essential genes among three groups according to their count of GO terms. Group A with GO count 1–4, Group B with GO count 5–8 and Group C with GO count >8 (error bars indicate standard errors).

doi:10.1371/journal.pone.0120784.g001

RNAs after the duplication of essential genes enables humans to maintain the redundant copies.

We observed the human essential duplicate genes mostly prefer to remain functionally redundant and can be maintained as backup copies, being able to escape the dosage imbalance. However, as the gene duplication is the mean of providing raw materials for genome evolution [4], we were interested in understanding the selection pressure on these backed up copies. Now, as the essential duplicates are functionally less divergent and dosage-balanced, their paralogs must be evolutionarily more conserved, in order to serve as backup copies upon future needs. We measured the evolutionary rates of human and mouse duplicated essential genes' paralogs, in terms of the ratio of nonsynonymous substitution rates per nonsynonymous sites (dN) and synonymous substitution rates per synonymous sites (dS) [see [materials and methods](#)] and obtained a significantly lower evolutionary rate of human counterpart ($dN/dS_{human} = 0.101$, $dN/dS_{mouse} = 0.128$, $P = 2.53 \times 10^{-5}$, Mann Whitney U test, $N_{mouse} = 2931$, $N_{human} = 1651$), indicated by their lower dN/dS ratio [Fig. 2]. This indicates that the redundant copies of human essential duplicates are evolutionarily conserved and may serve as backup copies upon future requirement, having the potential to increase the gene deletion fitness effect.

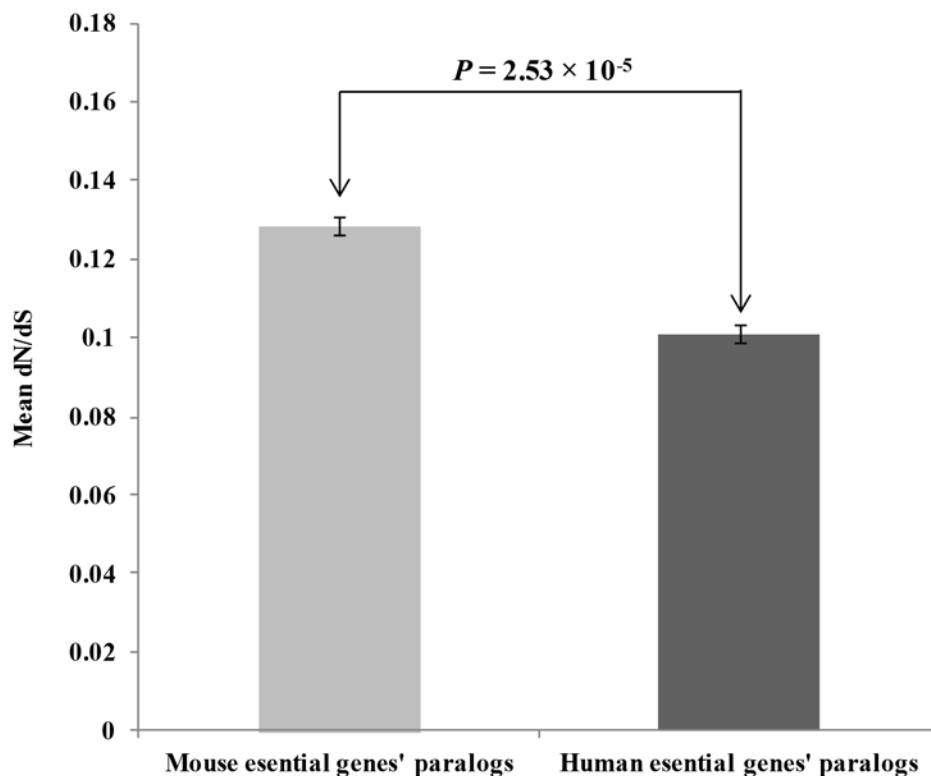


Fig 2. Mean dN/dS value of mouse and human essential genes' paralogs (error bars indicate standard errors).

doi:10.1371/journal.pone.0120784.g002

Conclusion

Gene duplication generates multiple copies of a gene that are initially functionally redundant, and their retention demands either functional diversification or regulation of the protein dosage. In this study we showed that human essential genes are mostly retained as duplicates, a trend which is different from mouse, with the duplicated copies being functionally more redundant in humans. Consistent with this, the evolutionary rate of these redundant human paralogs of essential genes is lower than that in mouse. We showed that these redundant human duplicates can be maintained due to the presence of more efficient dosage-regulation. Our study sheds light on the importance of the backup copies to restore the fitness effect of gene deletion, thereby increasing the fitness in humans. This study opens the future direction for in depth analysis of duplicated essential genes and their role in the human protein evolution.

Supporting Information

S1 Dataset. Mouse and Human genes used in this study. This dataset contains the essentiality, duplicability, involvement in development and phyletic age data of mouse and human genes.

(XLSX)

S2 Dataset. The duplicated pairs of Mouse and Human genes. This dataset contains the duplicate pairs for mouse and human genes used for functional distance measurement.

(XLSX)

S3 Dataset. The pseudogenization status of the paralogs of Mouse and Human essential duplicate genes. This dataset contains the pseudogene annotation for all mouse and human genes under study.
(XLSX)

Acknowledgments

We are thankful to the editor and the anonymous reviewer for their helpful suggestions in improving our manuscript. We acknowledge Mr. Sandip Chakraborty for his valuable comments and help.

Author Contributions

Conceived and designed the experiments: DA DM SP TCG. Performed the experiments: DA. Analyzed the data: DA DM. Contributed reagents/materials/analysis tools: DA SP TCG. Wrote the paper: DA SP TCG.

References

1. Ohno S, Wolf U, Atkin NB. Evolution from fish to mammals by gene duplication. *Hereditas*. 1968; 59: 169–187. PMID: [5662632](#)
2. Stephens SG. Possible significances of duplication in evolution. *Adv Genet*. 1951; 4: 247–265. PMID: [14943679](#)
3. Zhang JZ. Evolution by gene duplication: an update. *Trends Ecol Evol*. 2003; 18: 292–298.
4. Ohno S. *Evolution by Gene Duplication*. 1970; Springer.
5. Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet*. 2010; 11: 97–108. doi: [10.1038/nrg2689](#) PMID: [20051986](#)
6. Liang H, Li W-H. Functional compensation by duplicated genes in mouse. *Trends Genet*. 2009; 25: 441–442. doi: [10.1016/j.tig.2009.08.001](#) PMID: [19783063](#)
7. Gu ZL, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. Role of duplicate genes in genetic robustness against null mutations. *Nature*. 2003; 421: 63–66. PMID: [12511954](#)
8. Clark AG. INVASION AND MAINTENANCE OF A GENE DUPLICATION. *P Natl Acad Sci USA*. 1994; 91: 2950–2954. PMID: [8159686](#)
9. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. *Genetics*. 2000; 154: 459–473. PMID: [10629003](#)
10. Liang H, Li W-H. Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet*. 2007; 23: 375–378. PMID: [17512629](#)
11. Makino T, Suzuki Y, Gojobori T. Differential evolutionary rates of duplicated genes in protein interaction network. *Gene*. 2006; 385: 57–63. PMID: [16979849](#)
12. D'Antonio M, Ciccarelli FD. Modification of Gene Duplicability during the Evolution of Protein Interaction Network. *PLoS Comput Biol*. 2011; 7(4). doi: [10.1371/journal.pcbi.1002028](#) PMID: [21556131](#)
13. Bhattacharya T, Ghosh TC. Protein Connectivity and Protein Complexity Promotes Human Gene Duplicability in a Mutually Exclusive Manner. *DNA Res*. 2010; 17: 261–270. doi: [10.1093/dnares/dsq019](#) PMID: [20829394](#)
14. Yang J, Lusk R, Li WH. Organismal complexity, protein complexity, and gene duplicability. *P Natl Acad Sci USA*. 2003; 100: 15661–15665. PMID: [14660792](#)
15. Waterhouse RM, Zdobnov EM, Kriventseva EV. Correlating Traits of Gene Retention, Sequence Divergence, Duplicability and Essentiality in Vertebrates, Arthropods, and Fungi. *Genome Biol Evol*. 2011; 3: 75–86. doi: [10.1093/gbe/evq083](#) PMID: [21148284](#)
16. Makino T, McLysaght A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *P Natl Acad Sci USA*. 2010; 107: 9270–9274. doi: [10.1073/pnas.0914697107](#) PMID: [20439718](#)
17. He XL, Zhang JZ. Higher duplicability of less important genes in yeast genomes. *Mol Biol Evol*. 2006; 23: 144–151. PMID: [16151181](#)
18. Liao B-Y, Zhang J. Mouse duplicate genes are as essential as singletons. *Trends in Genetics*. 2007; 23: 378–381. PMID: [17559966](#)

19. Makino T, Hokamp K, McLysaght A. The complex relationship of gene duplication and essentiality. *Trends Genet.* 2009; 25: 152–155. doi: [10.1016/j.tig.2009.03.001](https://doi.org/10.1016/j.tig.2009.03.001) PMID: [19285746](https://pubmed.ncbi.nlm.nih.gov/19285746/)
20. Liao B-Y, Scott NM, Zhang J. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol Biol Evol.* 2006; 23: 2072–2080. PMID: [16887903](https://pubmed.ncbi.nlm.nih.gov/16887903/)
21. Chen W-H, Trachana K, Lercher MJ, Bork P. Younger Genes Are Less Likely to Be Essential than Older Genes, and Duplicates Are Less Likely to Be Essential than Singletons of the Same Age. *Mol Biol Evol.* 2012; 29: 1703–1706. doi: [10.1093/molbev/mss014](https://doi.org/10.1093/molbev/mss014) PMID: [22319151](https://pubmed.ncbi.nlm.nih.gov/22319151/)
22. Yang J, Gu ZL, Li WH. Rate of protein evolution versus fitness effect of gene deletion. *Mol Biol Evol.* 2003; 20: 772–774. PMID: [12679525](https://pubmed.ncbi.nlm.nih.gov/12679525/)
23. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* 2002; 12: 962–968. PMID: [12045149](https://pubmed.ncbi.nlm.nih.gov/12045149/)
24. Su Z, Gu X. Predicting the Proportion of Essential Genes in Mouse Duplicates Based on Biased Mouse Knockout Genes. *J Mol Evol.* 2008; 67: 705–709. doi: [10.1007/s00239-008-9170-9](https://doi.org/10.1007/s00239-008-9170-9) PMID: [19005716](https://pubmed.ncbi.nlm.nih.gov/19005716/)
25. Georgi B, Voight BF, Bucan M. From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. *PLoS Genet.* 2013; 9(5).
26. Liao B-Y, Zhang J. Null mutations in human and mouse orthologs frequently result in different phenotypes. *P Natl Acad Sci USA.* 2008; 105: 6987–6992. doi: [10.1073/pnas.0800387105](https://doi.org/10.1073/pnas.0800387105) PMID: [18458337](https://pubmed.ncbi.nlm.nih.gov/18458337/)
27. Chen W-H, Minguez P, Lercher MJ, Bork P. OGEE: an online gene essentiality database. *Nucleic Acids Res.* 2012; 40: D901–D906. doi: [10.1093/nar/gkr986](https://doi.org/10.1093/nar/gkr986) PMID: [22075992](https://pubmed.ncbi.nlm.nih.gov/22075992/)
28. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *P Natl Acad Sci USA.* 2009; 106: 7273–7280. doi: [10.1073/pnas.0901808106](https://doi.org/10.1073/pnas.0901808106) PMID: [19351897](https://pubmed.ncbi.nlm.nih.gov/19351897/)
29. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, et al. Ensembl 2013. *Nucleic Acids Res.* 2013; 41: D48–D55. doi: [10.1093/nar/gks1236](https://doi.org/10.1093/nar/gks1236) PMID: [23203987](https://pubmed.ncbi.nlm.nih.gov/23203987/)
30. Baudot A, Jacq B, Brun C. A scale of functional divergence for yeast duplicated genes revealed from analysis of the protein-protein interaction network. *Genome Biol.* 2004; 5(10:).
31. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of Isy-6 and other microRNAs. *Nat Struct Mol Biol.* 2011; 18: 1139–U1175. doi: [10.1038/nsmb.2115](https://doi.org/10.1038/nsmb.2115) PMID: [21909094](https://pubmed.ncbi.nlm.nih.gov/21909094/)
32. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature.* 2001; 411: 41–42. PMID: [11333967](https://pubmed.ncbi.nlm.nih.gov/11333967/)
33. Barabasi AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. *Nat Rev Genet* 2004; 5: 101–U115. PMID: [14735121](https://pubmed.ncbi.nlm.nih.gov/14735121/)
34. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *P Natl Acad Sci USA.* 2007; 104: 8685–8690. PMID: [17502601](https://pubmed.ncbi.nlm.nih.gov/17502601/)
35. He X, Zhang J. Why do hubs tend to be essential in protein networks? *PLoS Genet.* 2006; 2: 826–834.
36. Li WH, Yang J, Gu X. Expression divergence between duplicate genes. *Trends Genet.* 2005; 21: 602–607. PMID: [16140417](https://pubmed.ncbi.nlm.nih.gov/16140417/)
37. Li J, Musso G, Zhang Z. Preferential regulation of duplicated genes by microRNAs in mammals. *Genome Biol.* 2008; 9(8): R132. doi: [10.1186/gb-2008-9-8-r132](https://doi.org/10.1186/gb-2008-9-8-r132) PMID: [18727826](https://pubmed.ncbi.nlm.nih.gov/18727826/)

RESEARCH ARTICLE

Open Access



Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution

Debarun Acharya and Tapash C. Ghosh*

Abstract

Background: Gene duplication is a genetic mutation that creates functionally redundant gene copies that are initially relieved from selective pressures and may adapt themselves to new functions with time. The levels of gene duplication may vary from small-scale duplication (SSD) to whole genome duplication (WGD). Studies with yeast revealed ample differences between these duplicates: Yeast WGD pairs were functionally more similar, less divergent in subcellular localization and contained a lesser proportion of essential genes. In this study, we explored the differences in evolutionary genomic properties of human SSD and WGD genes, with the identifiable human duplicates coming from the two rounds of whole genome duplication occurred early in vertebrate evolution.

Results: We observed that these two groups of duplicates were also dissimilar in terms of their evolutionary and genomic properties. But interestingly, this is not like the same observed in yeast. The human WGDs were found to be functionally less similar, diverge more in subcellular level and contain a higher proportion of essential genes than the SSDs, all of which are opposite from yeast. Additionally, we explored that human WGDs were more divergent in their gene expression profile, have higher multifunctionality and are more often associated with disease, and are evolutionarily more conserved than human SSDs.

Conclusions: Our study suggests that human WGD duplicates are more divergent and entails the adaptation of WGDs to novel and important functions that consequently lead to their evolutionary conservation in the course of evolution.

Keywords: Small-scale duplication, Whole-genome duplication, Functional divergence, Gene essentiality, Disease genes, Protein multifunctionality, Evolutionary rate

Background

Gene duplication is a key source for generating new gene copies from pre-existing ones [1–3]. These newly-made gene copies are initially functionally redundant and relieved from selective pressure, and may adapt themselves to new functions [2, 4–6]. Thus, many of the previous studies concluded gene duplication as the primary guiding force of organism evolution for providing raw genetic materials for genome evolution [1, 2, 7]. Although, the retention of

duplicated genes is not a trouble-free process and most of the duplicates become nonfunctionalized and/or lost from the genome [2], whereas others become fixed within the genome in course of evolution. The retention of duplicates might be initially favourable due to circumstances like increased gene dosage advantage, where the duplication and subsequent increase in the gene product may be advantageous to the organism [5, 8]. Additionally, gene duplicates may serve as backup copies capable of functional compensation upon gene deletion [9] and provide increased genetic robustness against deleterious mutations [10], but their maintenance requires stringent regulation in

* Correspondence: tapash@jcbosc.ac.in
Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata
700054 West Bengal, India

gene dosage [11, 12] or expression patterns [13–16]. That apart, the duplicates may either diverge at the subcellular protein localization [17] or share the ancestral function [18] after complementary degenerative mutations (subfunctionalization) [19] or adapt to new functions (neofunctionalization) [2]. Furthermore, there are also subtle differences in the extent of gene duplication. In most of the cases, duplication involves a single gene and termed as small-scale duplication (SSD), whereas, large-scale duplications may involve many genes, chromosomal segments or even the entire genome, with the latter being known as whole-genome duplication (WGD) [20]. Although small-scale duplication can occur at any time and may be retained in course of evolution, there are a few evidences of whole genome duplication in eukaryotic organisms, being most common and widely studied in the evolution of plant genome [21–24]. Many previous studies highlighted the evidence of an ancient WGD in the yeast genome [25–27]. Additionally, evidence of two rounds of whole-genome duplication was also prominent in the early vertebrate evolution [28–33], which provides the raw materials for increasing genome and organism complexity and extensive species diversity [29, 31] and hence, is an important process in vertebrate evolution [30, 31].

However, as genes' functions are mainly mediated by their encoded proteins, which primarily function with the association of other such proteins [34], the proper functioning of a gene depends on the stoichiometric balance of the proteins participants. The retention of duplicated genes creates a stoichiometric disparity in the protein-protein interaction network, with the duplicated genes producing more proteins than the non-duplicated ones [35–37]. The two extent of duplication affect their associated protein-interaction network differentially [20, 38–41]. In WGD, the whole PPI network becomes simultaneously duplicated, and the stoichiometric balance of the participant proteins remains the same; whereas in SSD, the duplicated gene tends to form more protein in contrast to the non-duplicated interacting partners, thereby creating an imbalance in the whole PPI network. Therefore, in general, whole-genome duplicates are expected to be retained intact within the genome [39].

Most of the studies highlighting gene duplication compared the attributes of duplicated genes with that of singletons [10, 42, 43]. This raised an important question – are all duplicates equal in their genomic and evolutionary characteristics? With the well-established gene duplication data in yeast, it became possible to identify the duplicates originated from whole-genome duplication as well as those from small-scale duplication [25]. Comparing these

two distinct duplicate groups, researchers observed quantifiable differences in yeast [20, 41, 44]. They found that the yeast WGDs are functionally more similar than SSD genes, which is independent of their sequence similarity [20, 44]. Additionally, yeast SSDs also diverge more at their subcellular localization than the WGDs [41]. Also, yeast SSD genes were found to contain a higher proportion of essential genes than WGD genes [20, 44].

The occurrence of two rounds of whole-genome duplication in early vertebrate lineage [28–33] and the subsequent detection of traces of these whole-genome duplicates in human [32, 39, 45] lead us to differentiate the genomic and evolutionary attributes of human small-scale and whole-genome duplicates. As the human WGDs stem from the ancient two rounds of genome duplication that had occurred in early vertebrates, it can be stated that these human duplicates became subjected to more evolutionary pressure due to their long term evolutionary exposure than that in yeast. Therefore, our study will explore the relative importance and the long-term fate of these whole-genome duplicates that had originated during the early vertebrate evolution in contrast to the duplicates originating spontaneously at small-scale.

Results

Functional similarity of human SSD and WGD genes

The functional similarities between each pair of human small-scale and whole-genome duplicates were calculated using the Gene Ontology (GO) annotation from the biomart interface of Ensembl (version 77) [46], using GO domains 'biological process' as well as 'molecular function'. We obtained a higher functional similarity in small-scale duplicates than the whole-genome duplicated group (Table 1). However, the functional diversification of paralogs is dependent on their nonsynonymous nucleotide substitution per nonsynonymous site (dN), and the whole-genome duplicates tend to have a higher dN value than the small-scale duplicates, for being evolutionarily more ancient. Therefore, we binned our dataset according to different dN ranges (nonsynonymous nucleotide substitution per nonsynonymous site) (see Materials and methods) and compared the functional similarity between SSD and WGD duplicate pairs. This approach is similar to that adopted by Hakes et al. [20]. We found that SSD duplicate pairs are functionally more similar than the WGD pairs in each dN range (Table 1) considering both their involvement in biological processes and molecular function (Fig. 1). In other terms, human WGD pairs were found to be functionally more divergent, independent of their sequence divergence.

Table 1 Differences between the properties of human small-scale and whole-genome duplicate pairs in different dN ranges. Pair wise two-tailed *Mann–Whitney U test* were used to compare the means of SSD and WGD pairs within each group

Parameter Measured	Database used	Overall			dN 0.0–0.1			dN 0.1–0.2		
		SSD	WGD	P-value	SSD	WGD	P-value	SSD	WGD	P-value
Functional Similarity between paralogs	Shared GO Terms for Biological Process	$\bar{x} = 0.710$ $N = 14742$	$\bar{x} = 0.415$ $N = 12022$	$<1.00 \times 10^{-6}$	$\bar{x} = 0.734$ $N = 3640$	$\bar{x} = 0.499$ $N = 414$	2.325×10^{-52}	$\bar{x} = 0.720$ $N = 2754$	$\bar{x} = 0.476$ $N = 1140$	5.925×10^{-123}
	Shared GO Terms for Molecular Function	$\bar{x} = 0.840$ $N = 18584$	$\bar{x} = 0.659$ $N = 12392$	$<1.00 \times 10^{-6}$	$\bar{x} = 0.850$ $N = 4668$	$\bar{x} = 0.724$ $N = 410$	6.077×10^{-47}	$\bar{x} = 0.856$ $N = 3510$	$\bar{x} = 0.706$ $N = 1188$	1.075×10^{-129}
Shared Subcellular Compartment of paralogs	GO Cellular Component	$\bar{x} = 0.782$ $N = 15248$	$\bar{x} = 0.541$ $N = 12198$	$<1.00 \times 10^{-6}$	$\bar{x} = 0.816$ $N = 3790$	$\bar{x} = 0.579$ $N = 380$	5.341×10^{-76}	$\bar{x} = 0.788$ $N = 2914$	$\bar{x} = 0.581$ $N = 1162$	5.652×10^{-119}
Gene expression profile similarity between paralogs	Human Protein Atlas	$\bar{x} = 0.403$ $N = 11726$	$\bar{x} = 0.193$ $N = 13060$	$<1.00 \times 10^{-6}$	$\bar{x} = 0.615$ $N = 2588$	$\bar{x} = 0.254$ $N = 426$	1.558×10^{-63}	$\bar{x} = 0.414$ $N = 2758$	$\bar{x} = 0.253$ $N = 1226$	1.774×10^{-42}
	Expression Atlas	$\bar{x} = 0.450$ $N = 15404$	$\bar{x} = 0.216$ $N = 13072$	$<1.00 \times 10^{-6}$	$\bar{x} = 0.508$ $N = 3628$	$\bar{x} = 0.284$ $N = 422$	1.032×10^{-30}	$\bar{x} = 0.457$ $N = 3458$	$\bar{x} = 0.280$ $N = 1220$	5.953×10^{-53}

Table 1 Differences between the properties of human small-scale and whole-genome duplicate pairs in different dN ranges. Pair wise two-tailed *Mann–Whitney U test* were used to compare the means of SSD and WGD pairs within each group (*Continued*)

Parameter Measured	Database used	dN 0.2–0.3			dN 0.3–0.4			dN > 0.4		
		SSD	WGD	P-value	SSD	WGD	P-value	SSD	WGD	P-value
Functional Similarity between paralogs	Shared GO Terms for Biological Process	$\bar{x} = 0.657$ N = 3328	$\bar{x} = 0.440$ N = 2002	2.892×10^{-120}	$\bar{x} = 0.726$ N = 4264	$\bar{x} = 0.413$ N = 2192	1.397×10^{-274}	$\bar{x} = 0.710$ N = 756	$\bar{x} = 0.391$ N = 6274	1.983×10^{-135}
	Shared GO Terms for Molecular Function	$\bar{x} = 0.810$ N = 4300	$\bar{x} = 0.696$ N = 2076	1.814×10^{-96}	$\bar{x} = 0.846$ N = 5246	$\bar{x} = 0.677$ N = 2250	4.976×10^{-218}	$\bar{x} = 0.826$ N = 860	$\bar{x} = 0.628$ N = 6468	5.072×10^{-107}
Shared Subcellular Compartment of paralogs	GO Cellular Component	$\bar{x} = 0.740$ N = 3444	$\bar{x} = 0.541$ N = 2036	3.344×10^{-139}	$\bar{x} = 0.781$ N = 4356	$\bar{x} = 0.555$ N = 2228	1.421×10^{-215}	$\bar{x} = 0.777$ N = 744	$\bar{x} = 0.527$ N = 6392	1.156×10^{-100}
Gene expression profile similarity between paralogs	Human Protein Atlas	$\bar{x} = 0.307$ N = 2834	$\bar{x} = 0.191$ N = 2158	7.331×10^{-32}	$\bar{x} = 0.316$ N = 3042	$\bar{x} = 0.190$ N = 2366	1.131×10^{-34}	$\bar{x} = 0.322$ N = 504	$\bar{x} = 0.179$ N = 6884	1.308×10^{-17}
	Expression Atlas	$\bar{x} = 0.430$ N = 3792	$\bar{x} = 0.216$ N = 2166	1.377×10^{-105}	$\bar{x} = 0.420$ N = 3922	$\bar{x} = 0.219$ N = 2370	5.471×10^{-96}	$\bar{x} = 0.394$ N = 604	$\bar{x} = 0.199$ N = 6894	5.735×10^{-36}

Subcellular localization of SSD and WGD pairs

In addition to the functional divergence, insight into the function of a gene is associated with the location of its encoded protein within the cell at the subcellular level. Many previous studies reported that gene duplication and the functional redundancy of duplicates can often be neutralized at the protein level by the subcellular protein compartmentalization [17, 47, 48]. Therefore, we also considered the subcellular localization of their encoded proteins as an alternative and/or associated mechanism beside functional divergence of the duplicated genes. The localization of the protein can be obtained by using the Gene Ontology (GO) terms under the GO domain ‘Cellular Component’ against its gene identifier. The shared cellular component between the paralogous copies of all SSD and WGD genes were calculated (see Materials and methods). We observed an overall higher subcellular compartment sharing of SSD pairs than that of WGD pairs (Table 1). When we binned our dataset according to different dN ranges as mentioned previously, the trend remains the same for each dN range (Table 1, Fig. 2), which indicates that the SSD genes are more often co-localized, and WGD genes are significantly more diverged in their subcellular localization, irrespective of their sequence divergence.

Gene expression correlation between SSD and WGD pairs

The divergence of duplicated genes and can also occur at the gene expression levels. Earlier studies suggested that the gene expression patterns of duplicated pairs often undergo a spatial variation [reviewed in Li et al. [15]], and this can be considered as a mechanism for their stable maintenance [13]. Therefore, it is essential to understand the co-expression of

the paralogs in different tissues after gene duplication, which is measured using the gene expression profiles of the paralogous copies in a wide range of normal tissues [14–16]. We used the high-throughput recent RNA-seq gene expression data of a wide range of normal human tissues from the Human Protein Atlas [49] and Expression Atlas [50] (see Materials and methods for more details). However, we observed that human SSD pairs have higher expression profile similarity than the WGD counterparts as a whole, and in each dN range (Table 1, Fig. 3), suggesting that the functionally redundant human SSD genes also have more correlated expression profiles, and WGDs tend to diverge more in gene expression patterns.

Evolutionary rate of human SSD and WGD genes

The differences of human SSD and WGD pairs in their evolutionary genomic attributes clearly suggest that the human WGDs may tend to adapt themselves to new functions and locations. To investigate this, we used the one-to-one Mouse as well as Chimpanzee orthologs (see Materials and methods for details) to compare the evolutionary rates of human SSD and WGD genes by the Nonsynonymous nucleotide substitution per nonsynonymous sites (dN) and the $\frac{dN}{dS}$ ratio, where ‘dS’ denotes synonymous nucleotide substitution per synonymous sites. We obtained a significantly slower evolutionary rate in WGD genes than the SSD genes for both the cases (Table 2, Fig. 4), indicating that the human WGD genes are evolutionarily more conserved, besides being functionally more diverged than the SSD genes, which is also supported by a previous study [51] and is consistent with the idea of the slower evolutionary rate of duplicated genes following their adaptation to new circumstances as described in Jordan et al.[43].

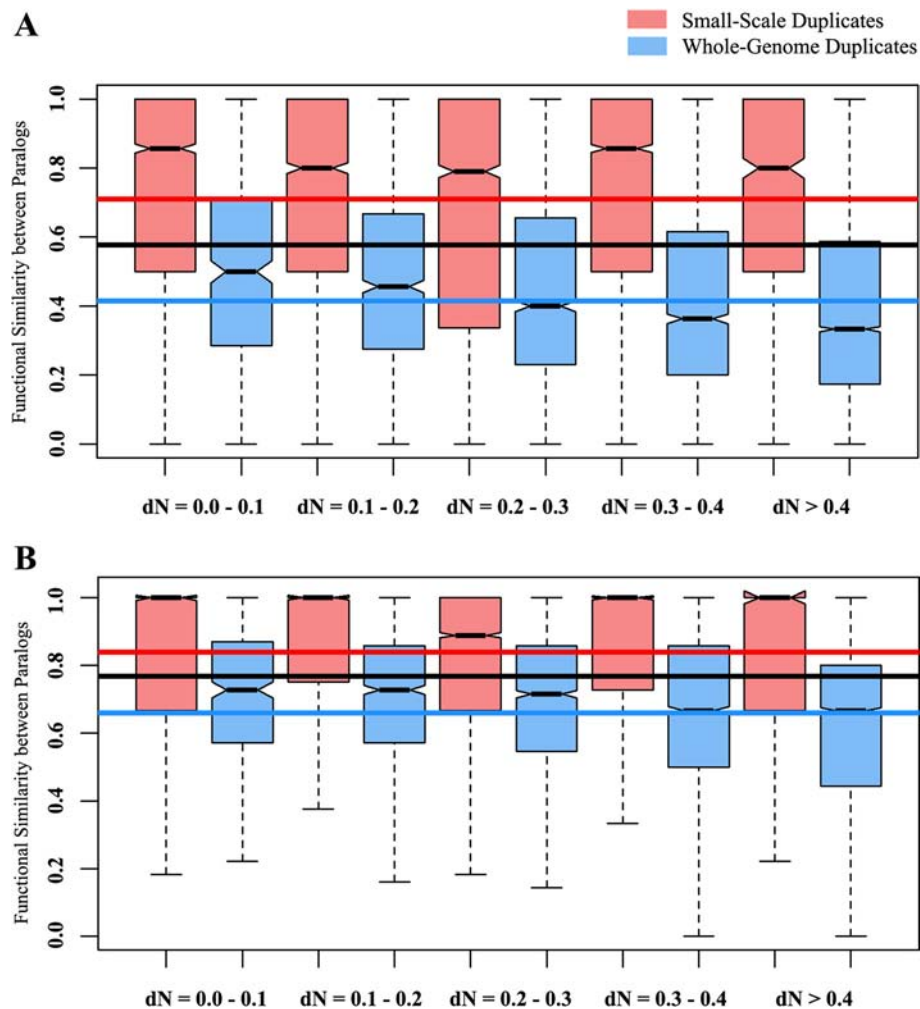


Fig. 1 Functional similarity between human small-scale and whole-genome duplicate pairs. The SSDs are represented in brick red and WGDs are represented in blue. The red and blue lines represent the mean functional similarity of SSD and WGD pairs, respectively. The black line represents the mean functional similarity of all human duplicates. The functional similarities between different dN ranges were calculated using both GO domains. **a.** Biological Process and **b.** Molecular Function (For every dN range, $P < 0.05$). For exact P-values, refer Table 1

Multifunctionality of human SSD and WGD genes

The higher probability of functional, sub-cellular localization and gene expression divergence of human WGD genes and their evolutionary conservation suggests that they may be associated with miscellaneous functions in contrast to the SSD counterparts. As our study is based on the functional fates of SSD and WGD genes, we were interested to observe which group is associated with more numerous functions. We used the unique GO biological process terms [52, 53] and the Pfam domain count [54] as proxies of multifunctionality (see Materials and methods). We observed that WGD-only genes are associated with more numerous Gene Ontology terms [Mean number of unique GO terms in SSD ≈ 5 , Mean number of unique GO terms in WGD ≈ 10 , $P = 6.707 \times 10^{-162}$, *Mann Whitney U test*, $N_{SSD} = 2569$, $N_{WGD} = 5437$] [Fig. 5a] and contain significantly more

domains in their encoded proteins [Mean number of Pfam domains in SSD = 1.61, Mean number of Pfam domains in WGD = 2.02, $P = 1.130 \times 10^{-46}$, *Mann Whitney U test*, $N_{SSD} = 3060$, $N_{WGD} = 5607$] [Fig. 5b] than SSD-only genes. This suggests that human whole-genome duplicates are associated with more variety of functions than human SSD genes.

Gene essentiality between human SSD and WGD genes

So far, the comparison between the human SSD and the WGD genes showed that the SSD genes tend to diverge less in their function, subcellular localization, as well as in gene expression levels in different tissues. Additionally, WGD genes were also found to be evolutionarily more conserved and were adapted to new functions. But the importance of such functions from organismal perspective also plays a crucial role to get the whole

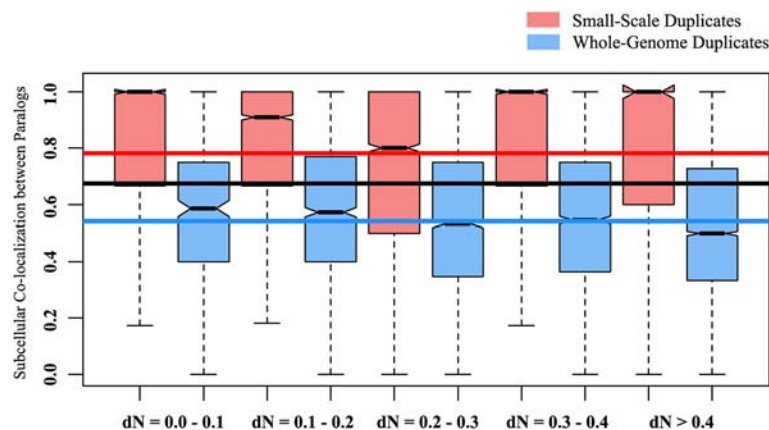


Fig. 2 Subcellular co-localization between human small-scale and whole-genome duplicate pairs. The SSDs are represented in brick red and WGDs are represented in blue. The red and blue lines represent the mean functional similarity of SSD and WGD pairs, respectively. The black line represents the mean functional similarity of all human duplicates (For every dN range, $P < 0.05$). For exact P-values, see Table 1

picture. The importance of a gene can be measured in terms of gene essentiality. We used the Online GENE Essentiality (OGEE) Database [55] to obtain human essential genes [56] and observed a significantly higher proportion of essential genes among the WGD genes [Proportion of essential genes in SSD genes = 4.601 %, $N_{SSD} = 2692$; Proportion of essential genes in WGD genes = 11.344 %, $N_{WGD} = 5730$] [$Z = -9.99$, confidence level 99 %; $P < 1.00 \times 10^{-4}$, two sample Z-test]. In other words, a greater portion of WGD genes shows lethality or sterility upon deletion than SSD genes, due to the absence of redundant paralogs in the former group.

Disease association of human SSD and WGD genes

Like gene essentiality, another important factor contributing to the importance of a gene in the organism is its disease association. It was previously hypothesised that gene duplication creates additional gene copies, and the increased functional redundancy can reduce the probability of disease formation by functional restoration upon gene deletion [57–59]. Therefore, the disease genes should remain as singletons [60]. More recently, studies linking gene duplication with disease hypothesise that duplication increase genetic redundancy, which in turn prefers accumulation of disease-associated mutations on the duplicates and hence, the duplicates may be more disease prone than the singletons [61]. Works with Mendelian disease genes reported their association with WGD genes [39, 62]. For our study, we considered all human disease associated genes from the Human Gene Mutation Database (HGMD) [63], which contains both Mendelian (monogenic) and complex (polygenic) disease genes. We observed that the proportion of disease genes is much higher among genes originating from whole-genome duplication [Proportion of disease genes in WGD genes = 61.46 %, $N_{WGD} = 5908$]; than the small-scale duplicates [Proportion of disease genes in

SSD genes = 27.89 %, $N_{SSD} = 3478$] [$Z = -31.420$, confidence level 99 %; $P < 1.00 \times 10^{-4}$, two sample Z-test]. This suggests that the reduction of functional redundancy in WGD genes increases disease susceptibility, and the increased ability of functional restoration reduce disease association of SSD genes.

Discussions

Gene duplication is the major source of genetic novelty that brings about genomic evolution. The term ‘genetic novelty’ comprise the generation of new genes from the pre-existing ones by mutation. Genetic mutation creates structural changes within the DNA which may lead to changes in the protein structure as well as its function. Although initially the duplicates are functionally redundant, they may either diverge or be maintained as backup copies during the course of evolution [2, 7, 64]. Recent studies with yeast confirmed that the whole-genome duplication maintains the stoichiometry of protein interaction network by increasing the dosage of its every participant, and small-scale duplication creates a stoichiometric imbalance within the network and hence, become functionally more divergent to maintain this balance [20, 38–41]. However, with the increasing organismal complexity and the genetic robustness, the whole-genome duplicates may also adapt to new functions, besides maintaining the resilience of protein interaction network. It will therefore be very interesting to explore the long-term fates of whole-genome duplication by observing human whole-genome duplicates (WGD), as the identifiable WGDs in human are traced from long back in the evolutionary scale i.e. from the two rounds of whole-genome duplication that had occurred in early vertebrate evolution. Therefore they must be evolved during the course of evolution from early vertebrates (like fish) to humans. In this study, we explored the

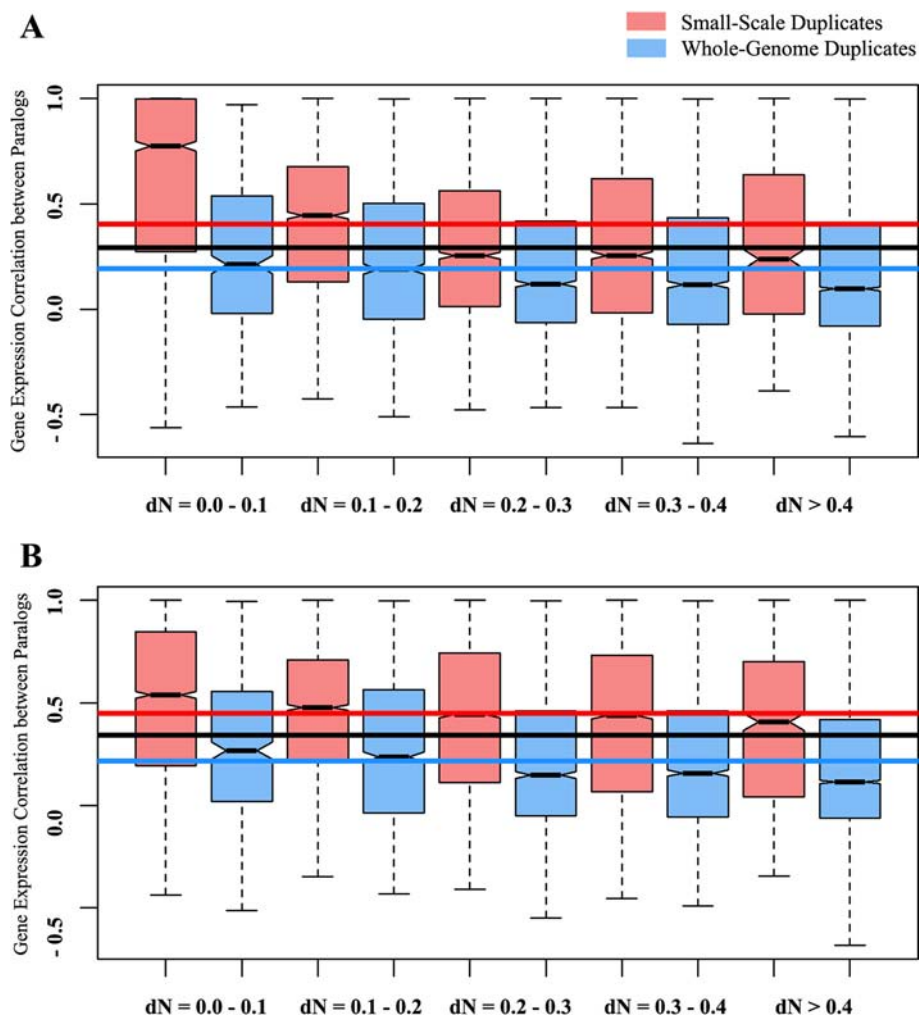


Fig. 3 Differences in gene expression correlation between human small-scale and whole-genome duplicate pairs. The gene expression correlation values of SSD and WGD pairs were calculated using RNA-seq gene expression data from **a.** Human Protein Atlas and **b.** Expression Atlas. The SSDs are represented in brick red and WGDs are represented in blue. The red and blue lines represent the mean gene expression correlation of SSD and WGD pairs, respectively. The black line represents the mean of gene expression correlation of all human duplicated gene pairs. (For every dN range, $P < 0.05$). Exact P-values are provided in Table 1

distinguishable differences between human small-scale and whole-genome duplicates. As we mentioned, the small-scale (SSD) and whole genome duplicates (WGD) are not similar in terms of their origin, and therefore in sequence divergence. So, we binned our datasets according to the non-synonymous nucleotide substitutions (dN) to compare the similarities in evolutionary genomic

properties between SSD and WGD duplicates independent of sequence changes that bring changes in amino acids, and in turn encoded proteins [20]. We observed that the human SSD and WGD duplicates were not similar in terms of their evolutionary and genomic properties. Based on their gene ontology terms, we found that WGD pairs share less functional similarity

Table 2 The evolutionary rate differences between human small-scale and whole-genome duplication using mouse (*Mus musculus*) and chimpanzee (*Pan troglodytes*) as outgroups. Two-tailed *Mann–Whitney U-Test* was used for comparisons between groups

Outgroup Used	Gene Group	Number of genes	Mean dN	P-value	Mean $\frac{dN}{dS}$	P-value
Mouse (<i>Mus musculus</i>)	Human Small-Scale Duplicates	958	0.089	6.212×10^{-18}	0.135	2.415×10^{-12}
	Human Whole-Genome Duplicates	5689	0.062		0.101	
Chimpanzee (<i>Pan troglodytes</i>)	Human Small-Scale Duplicates	1611	0.012	3.842×10^{-99}	0.480	2.410×10^{-76}
	Human Whole-Genome Duplicates	5309	0.006		0.257	

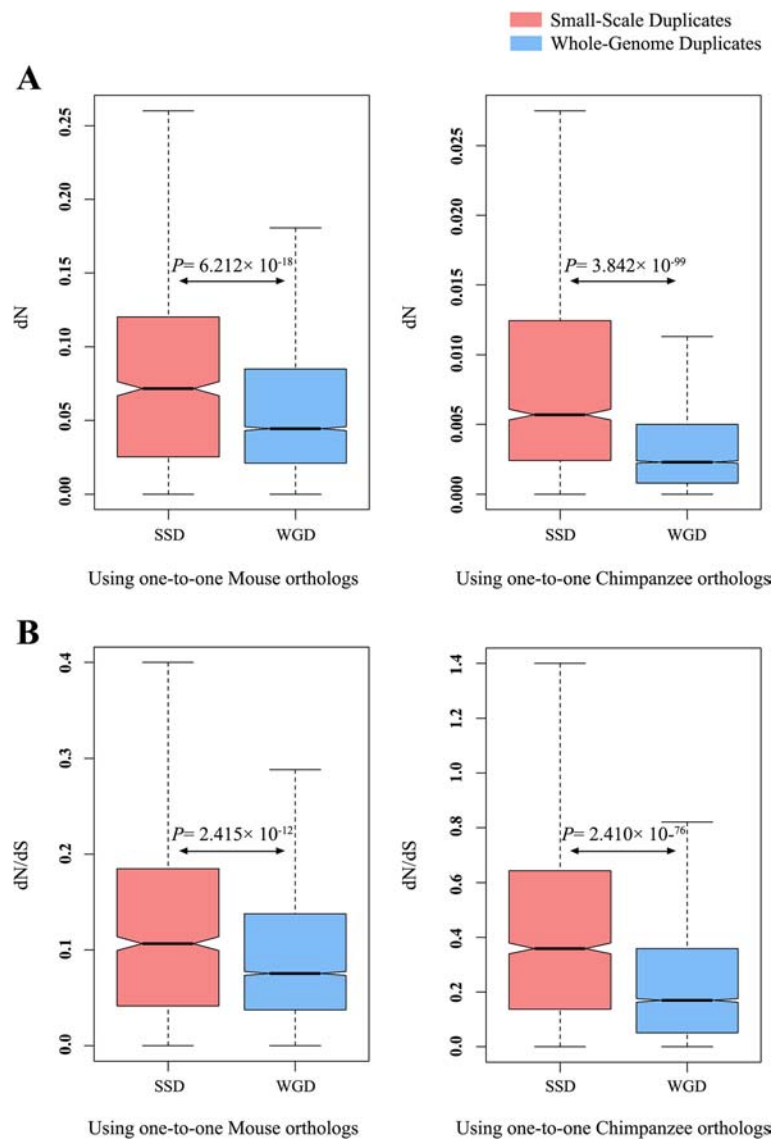


Fig. 4 Differences in evolutionary rates of human small-scale and whole genome duplicates using Mouse (*Mus musculus*) and Chimpanzee (*Pan troglodytes*) one-to-one orthologs. Both the dN (**a**) and the $\frac{dN}{dS}$ ratios (**b**) were used as the measurements of evolutionary rate. The SSDs are represented in brick red and WGDs are represented in blue. Exact P-values are provided in the figure and in Table 2

than the SSD pairs, irrespective of their sequence divergence for both the ‘GO Biological Process’ and ‘GO Molecular Function’ domains (Fig. 1, Table 1). We observed that these results are not influenced by duplicates having a large family size by conducting the same experiments using the closest duplicate pair only for both SSD and WGD duplicates (Additional file 1: Figure S1). We also observed that this difference is not due to the percentage identity based on which the SSD pairs are obtained, as using more stringent thresholds for determining SSD pairs also shows the similar trend (Additional file 1: Figure S2).

As the function of a protein is dependent on its localization in subcellular compartments [65], another possible mode of channelizing duplicated genes is in the subcellular localization of their encoded proteins [17]. Previous reports highlighted that the subcellular adaptation of duplicated proteins is also associated with the functional diversification [17, 47]. Consistent with this finding, we also observed a higher subcellular colocalization of the proteins encoded by SSD pairs (Table 1, Fig. 2; Additional file 1: Figure S3). This pattern is also opposite to that of yeast, where SSD pairs were more divergent in their subcellular localization, suggesting the human

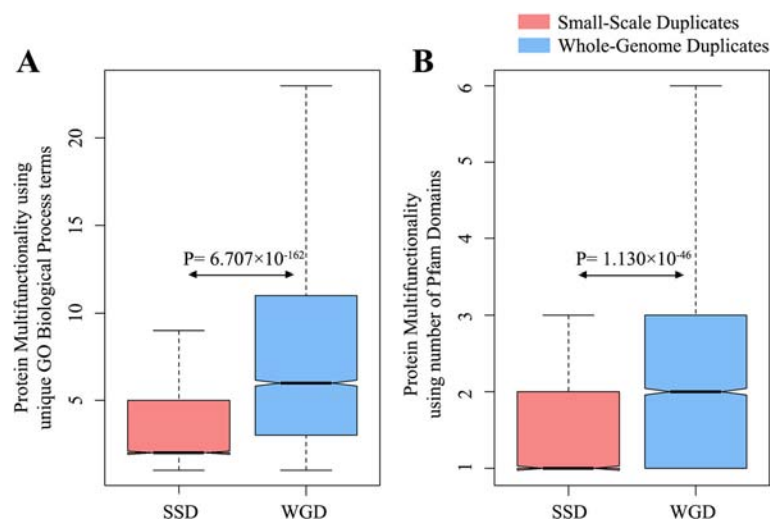


Fig. 5 Multifunctionality of human small-scale and whole-genome duplicates: **a.** Using their association with unique GO-Biological Processes. **b.** Using the number of Pfam domains. The SSDs are represented in brick red and WGDs are represented in blue. Exact P-values are provided in the figure

whole-genome duplicates have a higher probability of adapting themselves to new locations than the SSD counterparts. However, in higher eukaryotes having a tissue-level organization, gene duplication and the subsequent functional redundancy between the paralogs are often regulated by patterning their gene expression in different tissues [13–16, 66–68]. For example, the paralogs may express differentially in different tissues so that the amount of the produced protein remain at a steady level. Therefore, the spatial variation of gene expression can be treated as a possible candidate for the maintenance of duplicated pairs in humans. But the differences in gene expression patterns of SSD and WGD duplicated pairs were still not clear. Using high-throughput RNA-seq gene expression data of human for at least 27 normal tissues, we observed that the SSD pairs are more often coexpressed in the same tissue, whereas, WGD pairs tend to express differentially, *i.e.* in different tissues. This explains the idea that these WGD duplicates have not only adapted themselves to divergent functions or new locations, but also in divergent tissues. This also suggests a higher probability of specialization of expression patterns of human WGD pairs than the SSDs having the same level of sequence divergence (Fig. 3, Table 1). Using more stringent sequence identity for identification of SSDs also shows the similar trend (Additional file 1: Figure S4). Additionally, using closest paralogs to normalize the influence of duplicates with large gene families also shows that the differences between human SSD and WGD pairs hold true (Additional file 1: Figure S1). However, as humans are very distantly related with reference to the vertebrate whole-

genome duplication event, we hypothesised that our results reflect the long-term evolutionary fates of genes originating from whole-genome duplication, with those originating from small-scale duplications. To prove our hypothesis, we firstly explored the influence of recent small-scale duplications in our dataset using phylostratum rank as the age of SSD genes [69]. We differentiated the SSD pairs in two groups- young-SSD pairs and old-SSD pairs (see Additional file 2 for more details) and reperformed our overall analysis. We observed that the proportion of young SSDs are very low in our dataset ($Z = 79.875$, confidence level 99 %; $P < 1.00 \times 10^{-4}$, two sample Z-test) and differences between old-SSD and WGD genes still persist (Additional file 2: Figure S5). From another perspective, we used *Xenopus tropicalis* as a control and compare the attributes of small-scale and whole-genome duplicates in xenopus genome. Interestingly, both the SSD and WGD pairs shows high functional similarity in xenopus, with very little or no difference (Additional file 3: Figure S6). This also indicates that in course of vertebrate evolution, although initially both the SSD and WGD duplicates were similar in their attributes, the WGD genes were found to be more suitable candidates to diverge themselves to perform novel functions.

The higher functional divergence of human WGD genes and divergence in their subcellular and tissue-specific gene expression patterns lead us to investigate the differences in evolutionary conservation between SSD and WGD genes. In general, the duplicated genes tend to evolve faster than singletons just after duplication due to the increased functional redundancy, and

subsequently upon its functional specialization, these duplicates evolve at a slower rate to maintain the functions to which it became adopted [43]. However, the human WGD genes are identified as the genes originated at the early vertebrate lineage, where the two rounds of genome duplication had happened. We observed a slower evolutionary rate in human WGD genes in contrast to SSD counterparts, which clearly demonstrates that the WGD genes have become adapted to new functions and lost its redundancy, and became slow evolving to maintain these functions (Fig. 4). The slower evolutionary rates and higher functional divergence of WGD genes indicate that the functions, to which they are adopted, are also evolutionarily conserved.

Our hypothesis that human WGDs have adapted to divergent functions and became evolutionarily conserved is further strengthened by the analysis of protein multifunctionality. The WGD genes and their encoded proteins tend to have higher multifunctionality than the SSD genes (Fig. 5), which strengthen our idea of higher adaptation of human WGD genes to new functions in contrast to SSD counterparts. However, besides the functional fate of duplicated genes, we were interested to comprehend the importance of such functions to the organism's life. Therefore, we also considered the importance of such functions to human. We used the gene essentiality along with the disease association as measurements of the vitality of a gene. Firstly we studied the effect of gene deletion to understand the functional restoration by the paralogous copy(ies). A recent study showed that the proportion of essential human gene is significantly higher in duplicates than in singletons [56]. Additionally, disease-associated genes were found to be enriched in duplicates [61]. Considering Mendelian disease genes, researchers also found WGDs to be more frequently disease-associated [39]. As our data contains two groups of duplicates which are quite different in their evolutionary genomic properties, we were curious to observe the proportion of essential genes and disease-associated genes (considering both Mendelian and Complex disease genes) among human SSD and WGD gene sets. We obtained a higher proportion of essential genes, as well as disease genes in the whole-genome duplicate set. Taken together, these results prove that the WGD genes have adapted themselves to serve more functions, which are more often crucial for humans, and may cause disease, sterility or even lethality upon disruption.

Conclusions

In summary, our results suggest that the human duplicates originated from WGD event in early vertebrate evolution are quite different from those originating spontaneously at a smaller-scale (SSD), but these

differences are exactly opposite to that of yeast. The possible explanation for this scenario is that the human WGDs have been traced back from long time ago on the evolutionary scale, as in humans, we obtained the WGDs from two rounds of whole-genome duplication occurred in early vertebrate lineage. Additionally, the SSDs in our dataset are also enriched in ancient genes. Clearly, it suggests that both the human SSD and WGD genes have faced many evolutionary challenges than that in yeast. However, we found that in long evolutionary timespan, WGDs are more prone to diverge themselves in structure, function as well as expression to perform new and beneficial roles within the organism than the SSD genes. This also increases the chance to cause disease or lethality upon mutation on the WGD genes, due to the inability of their paralogous copies to restore the gene-deletion fitness. However, why the ancient SSD and WGD genes show differences in their functional divergence, being evolutionarily similar in origin, is a matter of future investigation. In conclusion, our study provide an insight into the long-term evolutionary fate of duplicates originated from whole-genome duplication, rather than their immediate impact on the organism, to which the early studies with yeast [20, 41, 44] were focussed.

Methods

Identification of human small-scale and whole genome duplicates

We obtained 22,447 human protein coding genes from the biomart interface of Ensembl version 77 [46] (<http://www.ensembl.org/biomart/martview>). The whole-genome duplicate (WGD) pairs were obtained and compiled from two datasets: 1. Makino and McLysaght [39] and 2. OHNOLOGS (<http://ohnologs.curie.fr/>) [45]. We used the strict [q-score (outgroup) < 0.01 and q-score] (self comparison) < 0.01] dataset of OHNOLOGS database to discard false positives and maintain stringency of our data. All other duplicates were obtained from the biomart interface of Ensembl 77 [45] and termed as small-scale duplicates (SSD). We used 50 % sequence identity with high paralogy confidence to assign paralogs, in order to retain old and/or distant paralogs. Finally, we obtained 34,746 duplicated pairs with 21,446 SSD and 13,300 WGD pairs comprising 4670 and 7070 genes, respectively (Additional file 4: Table S1).

As our dataset contains two groups of duplicates originated differentially in evolutionary time-scale, they are also different in terms of sequence divergence between duplicated pairs. The whole genome duplicates have originated during the evolution of early vertebrates and the small-scale duplicates have originated spontaneously at different times, thus, the latter may contain more recent duplicates with a possibility of being less diverged in sequence level. Therefore, it is necessary to remove

the bias due to the differential sequence divergence of SSDs and WGDs for calculating their differences in their functional properties. For this, we binned our dataset according to the nonsynonymous nucleotide substitution per nonsynonymous sites (dN) between each duplicated pairs, as the nonsynonymous substitutions bring change at protein level and older duplicate group (WGD) will tend to have higher dN than the newer one (SSD). Finally, both SSD and WGD duplicate pairs, we obtained five groups based on dN ranges between the paralogs – dN $_{0.0-0.1}$, dN $_{0.1-0.2}$, dN $_{0.2-0.3}$, dN $_{0.3-0.4}$ and dN $_{>0.4}$ and differentiated the evolutionary features between SSD and WGD genes in each dN range.

Functional similarity

The functions of human protein coding genes represented by their Gene Ontology terms were obtained from the biomart interface of Ensembl version 77 [46]. We considered the GO domains ‘Biological process’ as well as ‘Molecular function’ separately for functional similarity measurement. The functional similarity within each duplicate pair were calculated by their GO annotations using the following formula adapted from the Bayesian data integration method [44, 70]-

$$\text{Functional Similarity}(i, j) = \frac{2 \times S(i, j)}{|\text{GO terms}(i) + \text{GO terms}(j)|}$$

Where ‘i’ and ‘j’ are duplicated pairs and ‘S(i,j)’ represents the Gene Ontology terms shared between the duplicated pairs ‘i’ and ‘j’.

Subcellular localization

The protein subcellular localization represented by the respective genes’ Gene Ontology terms for the GO domain ‘Cellular component’ were obtained from the biomart interface of Ensembl version 77 [46]. Considering the Gene Ontology terms of a gene and its paralog, we obtained the subcellular compartment sharing for each SSD and WGD duplicate pairs. With the same formula used for functional similarity calculation mentioned previously, we calculated the subcellular compartment sharing for each duplicate pairs and compared the SSD and WGD genes of different dN ranges (as mentioned above).

Gene expression

The RNA-seq gene expression data of human were taken from two databases- The gene expression values of 9113 duplicated genes in 27 different tissues (namely adipose tissue, adrenal gland, appendix, bone marrow, cerebral cortex, colon, duodenum, oesophagus, gallbladder, heart muscle, kidney, liver, lung, lymph node, ovary, pancreas, placenta, prostate, salivary gland, skin, small

intestine, spleen, stomach, testis, thyroid gland, urinary bladder and uterus) were extracted from the human protein atlas Release 9 (<http://www.proteinatlas.org/>) [49, 71] and 9393 duplicate genes in 32 different tissues (namely adipose tissue, adrenal gland, ovary, appendix, bladder, bone marrow, cerebral cortex, colon, duodenum, endometrium, oesophagus, fallopian tube, gall bladder, heart, kidney, liver, lung, lymph node, pancreas, placenta, prostate, rectum, salivary gland, skeletal muscle, skin, small intestine, smooth muscle, spleen, stomach, testis, thyroid and tonsil) were obtained from Expression Atlas (<http://www.ebi.ac.uk/gxa>) [50, 72], which present stable repositories of experimental RNA-seq gene expression data in human tissues. The Pearson correlation coefficient (see formula below) was used to determine the expression profile similarity within the paralogous copies.

$$\text{Pearson correlation coefficient}(r) = \frac{N \sum ij - (\sum i)(\sum j)}{\sqrt{[N \sum i^2 - (\sum i)^2][N \sum j^2 - (\sum j)^2]}}$$

Where ‘i’ and ‘j’ are paralogous pairs, ‘N’ is the total number of tissues, ‘ $\sum ij$ ’ is the sum of the products of paired expression signal intensities, ‘ $\sum i$ ’ sum of expression signal intensities for gene ‘i’, ‘ $\sum j$ ’ is the sum of expression signal intensities for gene ‘j’, ‘ $\sum i^2$ ’ is sum of squared expression signal intensities of gene ‘i’, ‘ $\sum j^2$ ’ is sum of squared expression signal intensities of gene ‘j’.

Evolutionary rate

The oldest and widely used measurement of evolutionary rate calculates the evolutionary rate by using either dN values [73], or the $\frac{dN}{dS}$ ratio [74, 75], where ‘dN’ denotes Nonsynonymous nucleotide substitution per nonsynonymous sites and ‘dS’ stands for Synonymous nucleotide substitution per synonymous sites. For our study, we obtained one-to-one Mouse (*Mus musculus*) and Chimpanzee (*Pan troglodytes*) orthologs for each human genes to obtain the dN and dS values from the biomart interface of Ensembl version 77 [46]. Mutation saturation was controlled by discarding all dS values ≥ 3 [76]. We discarded the genes having paralogous copies from both small-scale and whole-genome duplications and used the nonredundant set of 9386 genes with only SSD or only WGD paralogs, but not both. Considering these SSD-only and WGD-only pairs, we obtained two distinct sets of genes: 1. Genes (and its paralogous copies) involved in Small-scale duplication only (SSD only) (containing 3478 genes), and 2. Genes involved in Whole-genome duplication only (WGD only) (containing 5908 genes). The dN values and

$\frac{dN}{dS}$ ratios between these groups were compared and used as the measurement of evolutionary rate.

Multifunctionality

The Multifunctionality of a gene and its encoded protein was measured by two approaches: A. Using their Gene Ontology annotation [77] for the GO domain 'biological process' from Ensembl Genome Browser [46], we calculated the unique biological processes of which a gene and its encoded protein(s) take part and used as the measurement of multifunctionality [51, 52], B. Additionally, we also considered the number of functional protein domains as proxy of Multifunctionality using Pfam protein families database. Finally, we compared the multifunctionality of SSD-only and WGD-only genes.

Gene essentiality

The human gene essentiality data were obtained from the Online Gene Essentiality (OGEE) database (<http://ogeedb.embl.de/#overview>) [55]. After matching this essentiality data with our dataset, we finally obtained gene essentiality information of 2692 SSD-only and 5730 WGD-only genes. We compared the proportion of essential genes between these duplicate sets.

Disease association

Human disease genes were obtained from 'The Human Gene Mutation Database' (<http://www.hgmd.cf.ac.uk/ac/index.php>) [63]. After discarding redundancy, we were able to identify 9668 disease genes of which, 9299 genes were matched to our dataset. This contains both the monogenic and the polygenic disease genes and is considered as human disease-associated genes. All other genes were termed 'non-disease genes' ($N = 13148$). We compared the proportion of disease genes among the SSD-only ($N = 3478$) and WGD-only ($N = 5908$) sets.

Software

We used the SPSS package (version 13) [78] and our in-house PERL-script for all statistical analyses. The R package [79] was used for data representation.

Availability of supporting data

The dataset of human small-scale and whole-genome duplicate pairs used in the study is available in Additional Table S1.

Ethics statement

The human data used in the study were collected from publicly available databases. Therefore ethics was not required for our study.

Additional files

Additional file 1: Figure S1. The differences between human small-scale and whole-genome duplicate pairs using the closest paralogs. **Figure S2.** Functional similarity between human small-scale duplicates with different sequence identity thresholds and whole-genome duplicate pairs.

Figure S3. Subcellular co-localization between human small-scale and whole-genome duplicate pairs. **Figure S4.** Differences in gene expression correlation between human small-scale and whole-genome duplicate pairs. (PDF 662 kb)

Additional file 2: Contains **Figure S5.** The differences between human young small-scale duplicates (Young-SSD) and old small-scale duplicates (Old-SSD) with whole-genome duplicates. (PDF 209 kb)

Additional file 3: Contains **Figure S6.** The differences between *Xenopus tropicalis* small-scale and whole-genome duplicates. (PDF 379 kb)

Additional file 4: Contains **Table S1.** The human Small-Scale and Whole-Genome duplicate pairs used in the study. (XLSX 705 kb)

Abbreviations

PPI: protein-protein interaction network; SSD: small-scale duplicates; WGD: whole-genome duplicates; GO: Gene ontology; dN: Nonsynonymous nucleotide substitution per nonsynonymous sites; dS: Synonymous nucleotide substitution per synonymous sites.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DA conceived, designed and performed the work. DA wrote the manuscript, TCG helped in planning of the work and drafting the manuscript. Both authors reviewed the manuscript.

Acknowledgements

We thank Dr. Sandip Chakraborty and Dr. Soumita Podder for technical advices. We also thank the Editor and two anonymous Reviewers for their constructive suggestions to improve our study. This work was supported by University Grants Commission (UGC) (Grant Sanction Letter No.F.2-8/2002 (SA-I) dated 04.10.2012 to D.A.).

Received: 28 September 2015 Accepted: 13 January 2016

References

- Ohno S, Wolf U, Atkin NB. Evolution from fish to mammals by gene duplication. *Hereditas*. 1968;59(1):169–87.
- Ohno S. *Evolution by Gene Duplication*. New York: Springer-Verlag; 1970.
- Stephens SG. Possible significances of duplication in evolution. *Adv Genet*. 1951;4:247–65.
- Clark AG. Invasion and maintenance of a gene duplication. *Proc Natl Acad Sci U S A*. 1994;91(8):2950–4.
- Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet*. 2010;11(2):97–108.
- Teshima KM, Innan H. Neofunctionalization of duplicated genes under the pressure of gene conversion. *Genetics*. 2008;178(3):1385–98.
- Taylor JS, Raes J. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet*. 2004;38:615–43.
- Kondrashov FA, Kondrashov AS. Role of selection in fixation of gene duplications. *J Theor Biol*. 2006;239(2):141–51.
- Liang H, Li W-H. Functional compensation by duplicated genes in mouse. *Trends Genet*. 2009;25(10):441–2.
- Gu ZL, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. Role of duplicate genes in genetic robustness against null mutations. *Nature*. 2003;421(6918):63–6.
- Li J, Musso G, Zhang Z. Preferential regulation of duplicated genes by microRNAs in mammals. *Genome Biol*. 2008;9(8):R132.
- Chang AY-F, Liao B-Y. DNA methylation rebalances gene dosage after mammalian gene duplications. *Mol Biol Evol*. 2012;29(1):133–44.

13. Qian W, Liao B-Y, Chang AY-F, Zhang J. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.* 2010;26(10):425–30.
14. Ganko EW, Meyers BC, Vision TJ. Divergence in expression between duplicated genes in *Arabidopsis*. *Mol Biol Evol.* 2007;24(10):2298–309.
15. Li WH, Yang J, Gu X. Expression divergence between duplicate genes. *Trends Genet.* 2005;21(11):602–7.
16. Li Z, Zhang H, Ge S, Gu X, Gao G, Luo J. Expression pattern divergence of duplicated genes in rice. *BMC Bioinformatics.* 2009;10(Suppl 6):S8.
17. Marques AC, Vinckenbosh N, Brawand D, Kaessmann H. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. *Genome Biol.* 2008;9(3):R54.
18. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. *Genetics.* 2000;154(1):459–73.
19. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics.* 1999;151(4):1531–45.
20. Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol.* 2007;8(10):R209.
21. Adams KL, Wendel JF. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol.* 2005;8(2):135–41.
22. Wendel JF. Genome evolution in polyploids. *Plant Mol Biol.* 2000;42(1):225–49.
23. Stebbins GL. *Chromosomal Evolution in Higher Plants*. New York: Addison-Wesley; 1971.
24. Blanc G, Barakat A, Guyot R, Cooke R, Delseny I. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell.* 2000;12(7):1093–101.
25. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature.* 2004;428(6983):617–24.
26. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature.* 1997;387(6634):708–13.
27. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, et al. Genome evolution in yeasts. *Nature.* 2004;430(6995):35–44.
28. Brunet FG, Croliis HR, Paris M, Aury J-M, Gilbert P, Jaillon O, et al. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol.* 2006;23(9):1808–16.
29. Zhou RJ, Cheng HH, Tiersch TR. Differential genome duplication and fish diversity. *Rev Fish Biol Fisher.* 2001;11(4):331–7.
30. Allendorf FW, Thorgaard GH. Tetraploidy and the evolution of salmonid fishes. In: Turner BJ, editor. *Evolutionary Genetics of Fishes*. New York: Plenum Press; 1984. p. 1–53.
31. Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* 2005;3(10):1700–8.
32. McLysaght A, Hokamp K, Wolfe KH. Extensive genomic duplication during early chordate evolution. *Nat Genet.* 2002;31(2):200–4.
33. Nakatani Y, Takeda H, Kohara Y, Morishita S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* 2007;17(9):1254–65.
34. Chakraborty S, Ghosh TC. Evolutionary rate heterogeneity of core and attachment proteins in yeast protein complexes. *Genome Biol Evol.* 2013;5(7):1366–75.
35. Papp B, Pal C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. *Nature.* 2003;424(6945):194–7.
36. He XL, Zhang JZ. Higher duplicability of less important genes in yeast genomes. *Mol Biol Evol.* 2006;23(1):144–51.
37. Birchler JA, Veitia RA. The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell.* 2007;19(2):395–402.
38. Freeling M, Thomas BC. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 2006;16(7):805–14.
39. Makino T, McLysaght A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A.* 2010;107(20):9270–4.
40. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science.* 2000;290(5494):1151–5.
41. Fares MA, Keane OM, Toft C, Carretero-Paulet L, Jones GW. The Roles of Whole-Genome and Small-Scale Duplications in the Functional Specialization of *Saccharomyces cerevisiae* Genes. *PLoS Genet.* 2013;9:e1003176.
42. Robinson-Rechavi M, Laudet V. Evolutionary rates of duplicate genes in fish and mammals. *Mol Biol Evol.* 2001;18(4):681–3.
43. Jordan IK, Wolf YI, Koonin EV. Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol Biol.* 2004;4:22.
44. Guan Y, Dunham MJ, Troyanskaya OG. Functional analysis of gene duplications in *Saccharomyces cerevisiae*. *Genetics.* 2007;175(2):933–43.
45. Singh PP, Arora J, Isambert H. Identification of ohnolog genes originating from whole genome duplication in early vertebrates, based on synteny comparison across multiple genomes. *PLoS Comput Biol.* 2015;11(7):e1004394.
46. Flicek P, Armode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. *Nucleic Acids Res.* 2014;42(D1):D749–55.
47. Byun-McKay SA, Geeta R. Protein subcellular relocalization: a new perspective on the origin of novel genes. *Trends Ecol Evol.* 2007;22(7):338–44.
48. Qian W, Zhang J. Protein subcellular relocalization in the evolution of yeast singleton and duplicate genes. *Genome Biol Evol.* 2009;1:198–204.
49. Uhlen M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. *Science.* 2015;347(6220):1260419.
50. Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, et al. Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 2014;42(D1):D926–32.
51. Satake M, Kawata M, McLysaght A, Makino T. Evolution of vertebrate tissues driven by differential modes of gene duplication. *DNA Res.* 2012;19:305–16.
52. Podder S, Mukhopadhyay P, Ghosh TC. Multifunctionality dominantly determines the rate of human housekeeping and tissue specific interacting protein evolution. *Gene.* 2009;439(1–2):11–6.
53. Salathe M, Ackermann M, Bonhoeffer S. The effect of multifunctionality on the rate of evolution in yeast. *Mol Biol Evol.* 2006;23(4):721–2.
54. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. *Nucleic Acids Res.* 2014;42(Database):D222–30.
55. Chen W-H, Minguez P, Lercher MJ, Bork P. OGEE: an online gene essentiality database. *Nucleic Acids Res.* 2012;40(D1):D901–6.
56. Acharya D, Mukherjee D, Podder S, Ghosh TC. Investigating different duplication pattern of essential genes in mouse and human. *PLoS One.* 2015;10(3):e0120784–4.
57. Hsiao T-L, Vitkup D. Role of duplicate genes in robustness against deleterious human mutations. *PLoS Genet.* 2008;4(3):e1000014–4.
58. Dean EJ, Davis JC, Davis RW, Petrov DA. Pervasive and Persistent Redundancy among Duplicated Genes in Yeast. *Plos Genet.* 2008;4(7):e1000113.
59. Wagner A. Gene duplications, robustness and evolutionary innovations. *Bioessays.* 2008;30(4):367–73.
60. Forslund K, Schreiber F, Thanintorn N, Sonnhammer ELL. OrthoDisease: tracking disease gene orthologs across 100 species. *Brief Bioinform.* 2011;12(5):463–73.
61. Dickerson JE, Robertson DL. On the origins of mendelian disease genes in man: the impact of gene duplication. *Mol Biol Evol.* 2012;29(1):61–9.
62. Chen WH, Zhao XM, van Noort V, Bork P. Human monogenic disease genes have frequently functionally redundant paralogs. *PLoS Comput Biol.* 2013;9(5):e1003073.
63. Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Curr Protoc Bioinform.* 2012;Chapter 1:Unit1.13–Unit1.13.
64. Zhang JZ. Evolution by gene duplication: an update. *Trends Ecol Evol.* 2003;18(6):292–8.
65. Emanuelsson O, von Heijne G. Prediction of organellar targeting signals. *BBA-Mol Cell Res.* 2001;1541(1–2):114–9.
66. Chen K, Zhang Y, Tang T, Shi S. Cis-regulatory change and expression divergence between duplicate genes formed by genome duplication of *Arabidopsis thaliana*. *Chinese Sci Bull.* 2010;55(22):2359–65.
67. Ha M, Kim E-D, Chen ZJ. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proc Natl Acad Sci U S A.* 2009;106(7):2295–300.
68. Leach LJ, Zhang Z, Lu C, Kearsey MJ, Luo Z. The role of Cis-Regulatory motifs and genetical control of expression in the divergence of yeast duplicate genes. *Mol Biol Evol.* 2007;24(11):2556–65.
69. Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics.* 2013;14(1):117.

70. Podder S, Ghosh TC. Insights into the molecular correlates modulating functional compensation between monogenic and polygenic disease gene duplicates in human. *Genomics*. 2011;97(4):200–4.
71. Uhlen M, Bjorling E, Agaton C, Szigartyo CA, Amini B, Andersen E, et al. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics*. 2005;4(12):1920–32.
72. Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, Filippov A, et al. Gene Expression Atlas update—a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res*. 2012;40(D1):D1077–81.
73. Begum T, Ghosh TC. Understanding the effect of secondary structures and aggregation on human protein folding class evolution. *J Mol Evol*. 2010;71(1):60–9.
74. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, et al. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci U S A*. 2005;102(15):5483–8.
75. Chen FC, Liao BY, Pan CL, Lin HY, Chang AY. Assessing determinants of exonic evolutionary rates in mammals. *Mol Biol Evol*. 2012;29(10):3121–9.
76. Begum T, Ghosh TC. Elucidating the genotype-phenotype relationships and network perturbations of human shared & specific disease genes from an evolutionary perspective. *Genome Biol Evol*. 2014;6(10):2741–53.
77. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. 2004;32 suppl 1:D258–61.
78. Nie N, Bent D, Hull C. SPSS: statistical package for the social sciences. New York: McGraw-Hill; 1970.
79. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comput Graph Stat*. 1996;5:299–314.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit





Cite this: *Mol. BioSyst.*, 2017, **13**, 2521

Insights into human intrinsically disordered proteins from their gene expression profile†

Arup Panda, Debarun Acharya * and Tapash Chandra Ghosh*

Expression level provides important clues about gene function. Previously, various efforts have been undertaken to profile human genes according to their expression level. Intrinsically disordered proteins (IDPs) do not adopt any rigid conformation under physiological conditions, however, are considered as an important functional class in all domains of life. Based on a human tissue-averaged gene expression level, previous studies showed that IDPs are expressed at a lower level than ordered globular proteins. Here, we examined the gene expression pattern of human ordered and disordered proteins in 32 normal tissues. We noticed that in most of the tissues, ordered and disordered proteins are expressed at a similar level. Moreover, in a number of tissues IDPs were found to be expressed at a higher level than ordered proteins. Rigorous statistical analyses suggested that the lower tissue-averaged gene expression level of IDPs (reported earlier) may be the consequence of their biased gene expression in some specific tissues and higher protein length. When we considered the gene repertoire of each tissue we noticed that a number of human tissues (brain, testes, etc.) selectively express a higher fraction of disordered proteins, which help them to maintain higher protein connectivity by forming disordered binding motifs and to sustain their functional specificities. Our results demonstrated that the disordered proteins are indispensable in these tissues for their functional advantages.

Received 26th May 2017,
Accepted 9th October 2017

DOI: 10.1039/c7mb00311k

rsc.li/molecular-biosystems

Introduction

Extensive research on intrinsically disordered proteins (IDPs) over the past few decades has led to a paradigm shift in our understanding of protein structural biology. These studies marked disordered proteins as a unique structural class, distinct from globular proteins in a number of structural and functional characteristics.^{1–3} Differences between ordered and disordered proteins are manifested in multiple layers, starting from their sequence composition to functional consequences and evolutionary aspects. At the primary structure level, IDPs are devoid of hydrophobic and aromatic residues and highly enriched with polar and charged amino acids.^{2,3} At the functional level, disordered proteins are enriched with processes complementary to the functions of globular proteins and are implicated in various regulatory and signaling cascades, such as control of cell division, apoptosis, post-translational modification, and transcription, etc.^{4–7} Since IDPs are composed of low complexity regions and are enriched with highly mutable hydrophilic residues these proteins tend to evolve at a faster rate as compared to globular proteins.^{8,9} Although IDPs lack three-dimensional structures under physiological conditions,

these proteins can adopt well-defined conformations upon interaction with partner proteins (coupled folding and binding).¹⁰ This unique feature of disordered proteins enables them to bind with a large number of partner molecules. Thus, disordered proteins often act as hubs in protein–protein interaction networks.¹¹ IDPs were initially regarded as a rare class of proteins. However, considering their abundance in different domains of life recent studies have suggested that IDPs constitute a very large class of proteins. Although there are controversies regarding the extent of the disorder, these studies suggested a general trend that IDPs are more common among complex genomes such as multi-cellular eukaryotes, however, are less abundant in unicellular bacterial and archaeal genomes.^{12–17} Because of their functional advantages, recently it was proposed that IDPs play important roles in the evolution of complex organisms and their strategies to cope with environmental stresses.^{18–21}

Although considerable progress has been achieved in our understanding of the characteristics of disordered proteins, many intriguing questions still remain elusive. One of the major goals of molecular biology is to profile transcripts in terms of their tissue distribution. Expression level provides a crucial indication of whether a gene is functional in a tissue or not. Moreover, gene expression profiles have major implications for understanding human disease etiology for the development of novel therapeutics.^{22,23} Therefore, previously a number of initiatives have been undertaken to estimate the expression

Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata 700 054, West Bengal, India. E-mail: stararup@gmail.com, debarun@jcbose.ac.in, tapash@jcbose.ac.in; Fax: +91-33-2355-3886; Tel: +91-33-2355 6626

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c7mb00311k

levels of human genes at genomic scales.^{22,24} However, until now little attention has been paid to investigating the gene expression signatures of disordered proteins at the tissue level. A few previous studies that have estimated their gene expression level considered the average gene expression values across all the tested tissues.^{25,26} Thus, to date, we have no clear understanding of whether these proteins are expressed in all human tissues and to what extent. Therefore, in this study, we took an initiative to profile disordered proteins in terms of their gene expression level across various human tissues. Based on a mean gene expression level of several human tissues, previously it was ascertained that as compared to globular proteins, most of the disordered proteins tend to be expressed at a lower level.^{25,26} However, tissue wise gene expression values, as we analyzed in this study, revealed a contrasting trend. Our study suggested that depending upon the nature of the tissue, disordered proteins may be expressed at a higher, lower or similar level as compared to ordered globular proteins. Moreover, here we found evidence that several human tissues selectively express a higher fraction of disordered proteins which help to sustain their functional specificities.

Methods

Data collection

Tissue wise gene expression values of human protein-coding sequences were obtained from Uhlén *et al.*,²⁷ In this dataset, average FPKM (fragments per kilobase of exon model per million mapped reads) values were provided for a total of 20 344 genes across 32 human tissues (adipose tissue, adrenal gland, appendix, bone marrow, brain, colon, duodenum, endometrium, esophagus, fallopian tube, gallbladder, heart muscle, kidney, liver, lung, lymph node, ovary, pancreas, placenta, prostate, rectum, salivary gland, skeletal muscle, skin, small intestine, smooth muscle, spleen, stomach, testis, thyroid gland, tonsil, urinary bladder). All these genes were tested for evidence at a protein level through various biochemical approaches; for details see Uhlén *et al.*²⁷ Here, we discarded 3174 genes either with no evidence or with evidence only at the transcript (RNA) level and further removed 535 genes with undetectable gene expression values (FPKM < 1 in all tissues). Following the gene annotation of Uhlén *et al.*, protein coding sequences of these genes were retrieved from Ensembl release 75.²⁸ For genes with more than one transcript, we considered the longest transcript. Sequences containing internal stop codons and partial codons were detected by CodonW (J Peden, <http://codonw.sourceforge.net>) and removed.

Prediction of protein intrinsic disorder content

Disorder predictions for human proteins were retrieved from the Database of Disordered Protein Predictions (D2P2) database.²⁹ Currently, D2P2 houses disorder predictions for more than 10 429 760 unique proteins from 1765 individual genomes. Each protein in this database was checked with nine disorder prediction algorithms, namely VL-XT, VSL2b, PrDOs, PV2, IUPred-S, IUPred-L,

Espritz-X, Espritz-N and Espritz-D, and searched for several other biologically relevant information such as the number of phosphorylation sites, domain annotations, *etc.* D2P2 allows users to retrieve disorder predictions in several useful formats. To calculate the disorder content of proteins in our dataset we retrieved a prediction for all human proteins currently available in this database. However, we considered disorder predictions only when we found an exact match between the sequences in the D2P2 database and the sequences in our dataset. We found disorder predictions for 15 472 proteins in our dataset all of which were considered for this analysis. To estimate the fraction of disordered residues in each protein, we considered the residues predicted as disordered residues by at least five of the nine algorithms. The disorder content was calculated as the fraction of the total number of such disordered residues in a protein to the length of that protein. We also checked the consistency of the results by calculating protein disordered content considering residues predicted as disordered by at least 6 and 7 algorithms.

Calculations of tissue selectivity

To determine the genes that are selectively expressed in different tissues we considered two approaches. At first, we followed the tissue annotation of Uhlén *et al.*, from where we retrieved gene expression values.²⁷ Based on the expression profile they classified human genes into six general categories (i) tissue enriched genes, (ii) group enriched genes, (iii) tissue enhanced genes, (iv) mixed genes, (v) genes which are expressed in all tissues and (vi) genes which are not expressed in any tissue. Among these categories, tissue enriched genes were defined with most stringent criteria, 5-fold higher FPKM in one tissue as compared to all the remaining tissues. To compile the list of genes that are selectively expressed in each tissue, we considered the genes that were annotated as 'tissue enriched genes' and associated with only one tissue. However, the genes that were identified as tissue-selective genes by this approach lack any statistical validation. Therefore, we considered another approach that defines tissue-selective genes through rigorous statistical analysis.^{30,31} Following Chang *et al.* and Greco *et al.* for each tissue–gene pair we calculated a tissue-selectivity score S_{ij} from the gene expression matrix as:

$$S_{ij} = W_i \times X_{ij}$$

Here, X_{ij} is the normalized gene expression (FPKM) value of gene 'i' in tissue 'j' and W_i is a gene-specific weight. The gene-specific weight W_i was calculated as follows:

$$W_i = \frac{1}{(N-1)} \sum_{k=1}^N (1 - X_{ik})$$

Here, X_{ik} is the FPKM value of gene 'i' in tissue 'k' and N is the total number of tested tissues.

The normalized gene expression value X_{ij} was calculated by dividing the FPKM value of gene 'i' in tissue 'j' with its highest FPKM across all the tested tissues.

$$X_{ij} = \frac{Y_{ij}}{\max\{Y_{ik}\}_{k=1}^N}$$

Tissue selectivity score S_{ij} ranges between zero and one, where one denotes a higher propensity of tissue-selective expression. The significance threshold for the tissue-selectivity score was computed through a permutation test. Briefly, we generated 1000 arbitrary gene expression datasets by sampling tissue-gene pairs randomly and calculated tissue-selectivity scores for each such dataset. For each tissue gene pair, we calculated the threshold value as the number of times the random tissue selective scores are greater than the real tissue selective score divided by the number of randomized datasets (1000). For a gene, if we found a tissue with FPKM > 100 with threshold value $< 10^{-2}$ then the gene was considered to be selectivity expressed in that tissue.^{30,31}

Prediction of molecular recognition features (MoRFs)

Protein binding sites embedded within disordered regions were predicted by the ANCHOR algorithm^{32,33} and fMoRFpred algorithm.³⁴ ANCHOR predicts protein-protein interaction sites that undergo disorder to order transition upon binding on the basis of pairwise inter-residue interaction energies irrespective of its amino acid composition and its secondary structure.³² This method was proposed to give an unbiased estimate of protein binding capacity.¹² fMoRFpred predicts MoRF regions with the help of support vector machine based on 20 features related to the structural and biochemical characteristics of the input protein sequence. This algorithm was tested with several benchmarking datasets and validated against experimentally supported results in small scales.³⁴ For each residue in the input sequence, fMoRFpred provides a binary classification where '1' denotes an MoRF residue and '0' a non-MoRF residue. Currently, fMoRFpred supports prediction for proteins less than 1000 residues in length. Here, we could predict MoRF regions for 13 281 of 15 472 proteins in our dataset and then we calculated the percentage of MoRF residues in those proteins.

Protein-protein interaction data

Human protein-protein interaction data were retrieved from BioGRID protein interaction repository (v-3.4.144).³⁵ Currently, BioGRID houses the largest number of interaction pools as compared to the other human interaction databases like HPRD,³⁶ MIPS,³⁷ FlyBase,³⁸ *etc.* Therefore, for systematic analysis of the interaction network, we chose the BioGRID database. Currently, there are interaction data for 21 270 unique human proteins collectively annotated with 279 852 non-redundant interactions. To compute protein connectivity, we considered human binary protein interactions with experimental evidence of physical connections. We removed self-interactions and counted the number of unique interaction partners that a protein connects with (protein connectivity).

Functional enrichment analysis

To determine the functional categories that are significantly over-represented among the genes that are selectively expressed in different human tissues we used the GOrilla Gene Ontology (GO) enrichment analysis tool.^{39,40} GOrilla automatically retrieves GO

terms (biological process, molecular functions, and cellular components) from gene names or identifiers and compares their distribution either in a ranked gene list or between a target and a background list of genes through rigorous statistical analysis. Along with the details (identifier, description, *P* values, *etc.*) of the terms that are significantly overrepresented in the target list, GOrilla provides a graphical overview of their hierarchical relationships. Here, we compared the distribution of GO terms in tissue-selective genes with respect to their distribution in the total of 15 472 human genes considered for this study.

Statistical analyses

All statistical tests were performed using the SPSS package. Following their non-parametric distribution, we compared the measures of different variables (protein disorder content, gene expression level, and protein length) by the Kruskal-Wallis *H* test, an extended version of the Mann-Whitney *U* test, applicable for comparing distributions between multiple independent groups. To determine significant differences we considered adjusted *P* values corrected for multiple comparisons. For correlation analysis, we calculated non-parametric Spearman's Rank correlation coefficient ρ , where significant correlations were denoted by $P < 0.05$.

Results

Gene expression level of disordered proteins across 32 human tissues

To analyze the gene expression pattern of human disordered proteins at the tissue level we considered the dataset provided by Uhlén *et al.*,²⁷ with two restrictions that (i) only genes with detectable expression (FPKM value ≥ 1) at least in one tissue and (ii) only genes with evidence at the protein level were selected. Genes with no evidence at the protein level were regarded as missing genes or non-coding genes and were suggested to be removed from the list of human protein-coding sequences.²⁷ Disorder predictions were retrieved from the D2P2²⁹ database and disorder content was estimated based on the consensus of 5 of 9 prediction algorithms (see materials and methods). Following Edwards *et al.*,²⁵ we categorized our dataset into five bins in ranges of 0–20% (ordered), 20–40% (moderately disordered), 40–60% (disordered), 60–80% (highly disordered) and 80–100% (extremely disordered) predicted disorder content. As has been suggested earlier,²⁵ here we noticed that both highly disordered (predicted disorder content $> 60\%$) and extremely disordered proteins (predicted disorder content $> 80\%$) are relatively rare in the human proteome (Fig. S1 in Supplementary file 1, ESI†). Following previous studies,^{22,27} an FPKM value of 1 was taken as a threshold to estimate the genes expressed in different tissues. Interestingly, among the genes expressed in different tissues (with FPKM > 1), $\sim 3\%$ of genes were found to be extremely disordered (predicted disorder content $> 80\%$) and $\sim 11\text{--}12\%$ of genes were predicted as highly disordered (predicted disorder content $> 60\%$). Next, we calculated the mean gene expression intensities of ordered and disordered proteins in each individual tissue (Fig. S2 and Table S1 in Supplementary file 1, ESI†).

In 21 of 32 tested tissues (adipose tissue, adrenal gland, appendix, colon, duodenum, esophagus, fallopian tube, gallbladder, heart muscle, lung, pancreas, placenta, prostate, rectum, salivary gland, small intestine, smooth muscle, stomach, thyroid gland, tonsil, and urinary bladder), we did not find a significant difference in gene expression level between any order and disorder categories ($P > 0.05$ for all the pairwise comparison among five disorder categories by Kruskal–Wallis H Test). However, in tissues like the brain, endometrium, lymph node, ovary, skeletal muscle, skin, spleen, and testes, disordered (bin3) and/or highly disordered (bin4) proteins were found to be expressed at a significantly higher level as compared to ordered proteins (bin1) ($P < 0.05$). In contrast, an opposite trend was noticed in three tissues – liver, kidney and bone marrow, where ordered proteins were found to be expressed at a relatively higher level than disordered proteins (Fig. S2 and Table S1 in Supplementary file 1, ESI†). We also tested whether the observed variations in mean gene expression levels between proteins in different disorder bins are due to random chance. For this analysis we generated 100 arbitrary gene expression matrices from our real gene expression dataset by random permutation of tissue gene pairs. Next, in each random dataset we found out the tissues where ordered and disordered proteins differ significantly in their mean gene expression level. Considering all those random datasets (32×100 tissue wise comparisons) we found significant differences in 142 tissues (~ 1.5 tissues per random dataset) (Supplementary results S2 in Supplementary file 3, ESI†). In $\sim 50\%$ of tissues where we found significant differences, disordered proteins were found to be expressed at a higher level, while in the remaining $\sim 50\%$ ordered proteins were found to be expressed at a higher level. Moreover, here we did not find any general trend in these tissues. Altogether this suggested that the observed variations are not due to random chance. In addition, the mean gene expression intensity values may have been biased by the very high expression of a few genes in some tissues. To check this possibility we calculated average gene expression intensities after removing the genes with expression intensity > 1000 (Fig. S3 in Supplementary file 1, ESI†) and > 5000 (Fig. S4 in Supplementary file 1, ESI†) in any of the tested tissues and considered median values instead of mean values (Supplementary file 2, ESI†). Further, we re-annotated proteins into five disordered bins based on the disorder content

predicted by the consensus of 6 (Fig. S5 in Supplementary file 1, ESI†) and 7 algorithms (Fig. S6 in Supplementary file 1, ESI†) and compared their mean gene expression levels. When we compared among these datasets (Fig. S2–S6 in Supplementary file 1, ESI†) we noticed a similar trend that in most of the human tissues there is no significant difference in gene expression between any disorder bins (Table S1 in Supplementary file 1, ESI†). For most of the tissues in which we found a significant difference we didn't find any consistent trend, however disordered proteins were found to be expressed at a lower level in the liver and kidneys across all these datasets, while at a higher level in the testes, ovaries and to some extent the brain. Previously, it was ascertained that disordered proteins tend to be expressed at a lower level than ordered globular proteins.^{25,26} However, these results imply that in most of the human tissues proteins are expressed at a similar level irrespective of their order and disorder tendencies. Moreover, here we found evidence that disordered proteins may be expressed at a higher level than ordered proteins depending upon the tissue physiology.

Tissue averaged gene expression level of disordered proteins

Based on the tissue averaged gene expression values, previously it was shown that human disordered proteins (predicted disorder content 40–80%), tend to be expressed at a comparatively lower level than ordered globular proteins.^{25,26} Thus our results are in apparent conflict with the results shown based on tissue averaged values. To check whether tissue averaged values would reflect a different scenario than what we found in individual tissues, we considered the average gene expression level of all the 32 tissues. As has been reported earlier, here we noticed that disordered and highly disordered proteins (predicted disorder content 40–80%) indeed have significantly lower tissue averaged gene expression levels than ordered proteins (Fig. 1A). However, all significant differences disappeared when we calculated the mean values without considering the tissues (liver, pancreas, and salivary gland) where we found large variation in gene expression between ordered and disordered proteins (Fig. 1B). Thus, these results suggest that the lower tissue averaged gene expression level of disordered proteins, as reported earlier, may have been caused by biased gene expression of these proteins in some specific tissues. To further evaluate the effect of other

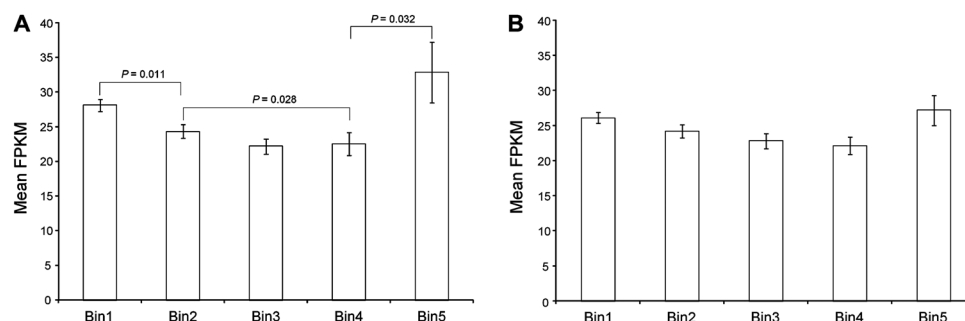


Fig. 1 Tissue-averaged mean gene expression levels of proteins in different disorder bins. (A) Average values calculated considering all 32 tissues, (B) considering 29 tissues (without considering liver, pancreas, and salivary gland where we found large variation in gene expression between ordered and disordered proteins).

factors we considered the impact of protein length. Gene length was regarded as a major determinant of gene expression level^{41,42} and was also shown to be correlated with protein disorder content.^{8,18} In accordance, here we noticed a significant negative correlation between protein length and tissue averaged gene expression level (Spearman $\rho = -0.132$, $P = 1 \times 10^{-6}$). Consequently, proteins in moderately and highly disorder bins were found to have a significantly higher length as compared to ordered proteins (Fig. 2), suggesting that disordered proteins may have lower gene expression due to their higher protein length. To analyze how protein length influences the correlation between gene expression level and protein disorder content we controlled this effect using partial correlation analysis. The weak correlation between gene expression and protein disorder content was found to disappear (Spearman $\rho = -0.018$, $P = 2.3 \times 10^{-2}$ for correlation between protein disorder content and average gene expression) controlling protein length. To evaluate whether the observed distribution of tissue averaged gene expression intensities has really been influenced by protein length we compared gene expression levels between ordered and disordered proteins of comparable length (in protein length bins). When we controlled the effect of protein length in this way we found no significant difference in mean gene expression intensity between ordered and disordered proteins in most of these bins (Table S2 in Supplementary file 1, ESI†). However, one probable reason for not finding a significant difference may be the lower sample size *i.e.* the number of ordered and disordered proteins to compare in each length bin. To consider this possibility, we randomly sampled 500 proteins from each of the ordered, moderately disordered, disordered and highly disordered protein groups such that the average gene lengths of these groups do not differ significantly. We then checked whether the tissue averaged gene expression level varies significantly between these groups. The extremely disordered group of proteins was not considered for this analysis due to the insufficiency of the dataset required for the randomization procedure. We repeated the procedure 1000 times, however, in more than 95% of cases we

did not find any significant difference (at 95% confidence level) in the tissue averaged gene expression level between any disorder bins. Thus, these results suggest that a lower tissue averaged gene expression level among the moderately and highly disordered proteins as has been reported earlier may be the consequence of their higher gene length.

Disordered proteins and tissue functionality

Our results suggested that although in most of the human tissues ordered and disordered proteins are expressed at a similar level, in some specific tissues disordered proteins tend to be expressed at a higher level than ordered proteins. To delve into this issue further, we analyzed gene expression specificities of ordered and disordered proteins in each individual tissue. Genes that are expressed predominantly in a particular tissue were considered to be important for functional specificities of that tissue.^{30,31,43} Therefore, here we considered the genes that are selectively expressed in each of the 32 tissues identified by two approaches (see Materials and methods). From Uhlén's *et al.*,²⁷ we retrieved 1707 tissue enriched genes, as compared to 1086 tissue-selective genes identified by the second method.^{30,31} For most of the tissues, we noticed a high degree of overlap between the lists of tissue-selective genes identified by these two methods (Fig. S7 in Supplementary file 1, ESI†). Moreover, genes which were identified as tissue-selective genes by both these methods (602 genes) were found to have the same tissue specificity. When we compared their protein disorder content, we found that genes that are selectively expressed in tissues like the testes, brain, *etc.* have a higher protein disorder content as compared to the genes that are expressed selectively in the liver, pancreas, kidney *etc.* tissues (Fig. 3). The higher protein disorder content of tissue-selective genes may suggest that disordered residues are indispensable for the proper functioning of the former group of tissues. In this context, we found it interesting to analyze why the genes that are selectively expressed in the former group of tissues encode more disordered residues as compared to the other groups of tissue-selective genes. Previously, it was ascertained that proteins that connect with a large number of partners in their interaction network (hub proteins) are more disordered as compared to the proteins that interact with few partners.^{11,44} Consequently, we compared different groups of tissue-selective genes in terms of their protein connectivity. In favor of their higher disorder content, here we noticed that genes that are selectively expressed in tissues like the testes and brain, *etc.* share higher protein connectivity than the genes that are selectively expressed in the liver or kidneys (Fig. 4). Proteins with higher connectivity were shown to encode a large number of disordered binding regions (protein binding sites within disordered regions) for their binding promiscuity.³² Therefore, we predicted disordered binding sites using two algorithms – (1) ANCHOR³³ and (2) fMoRFpred,³⁴ both of which suggested that the genes that are selectively expressed in the former group of tissues (testes, brain, *etc.*) encode a greater fraction of such motifs than the liver and kidney *etc.* tissue-selective genes (Fig. 5). This may suggest that a higher fraction of disordered residues among the

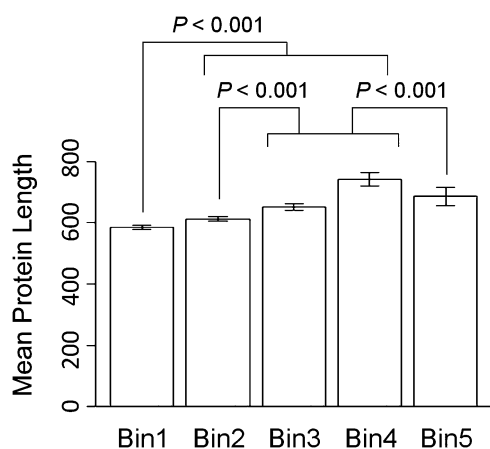


Fig. 2 Average length of proteins in different disorder bins. Significant differences for pair-wise comparison between different groups were evaluated through Kruskal Wallis H test and shown with P -values.

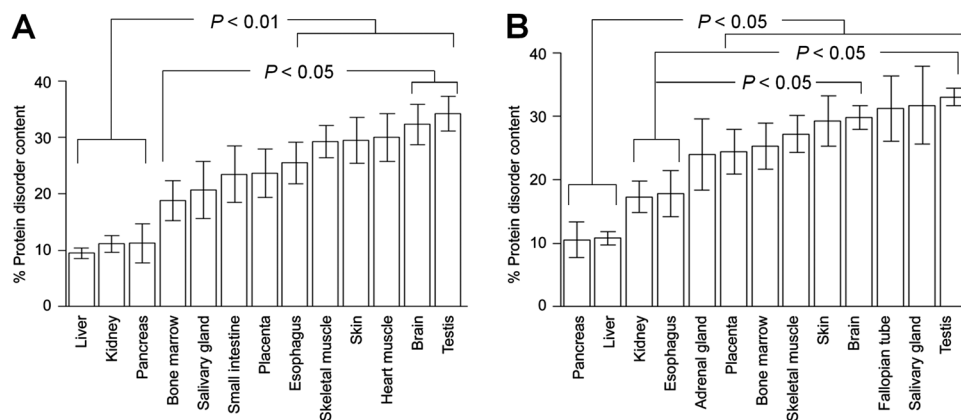


Fig. 3 Average protein disorder content of different groups of tissue-enriched genes. A protein disorder content was retrieved from the D2P2 database and tissue-selective genes were identified using two methods: (A) Uhlen *et al.* and (B) Chang *et al.*, and Greco *et al.* Here, tissues with 30 or more selective genes were shown. Significant differences in protein disorder content between the different groups of tissue selective genes were evaluated through Kruskal Wallis *H* test shown with *P*-values.

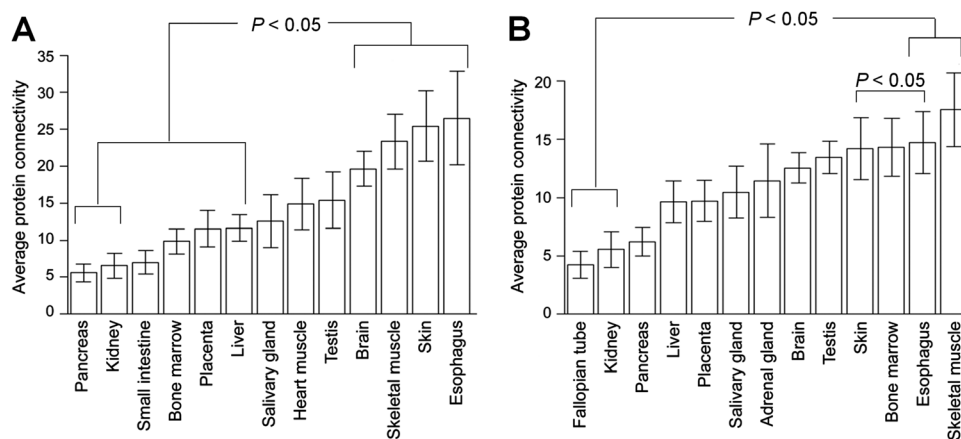


Fig. 4 Average protein connectivity of different groups of tissue-enriched genes. Only tissues with 30 or more tissue enriched genes were shown. (A) Tissue-selective genes retrieved from Uhlen *et al.* and (B) those identified following Chang *et al.*, and Greco *et al.* Significant differences between the different groups of tissue-selective genes were evaluated through Kruskal Wallis *H* test and shown with *P*-values.

former groups of tissue-selective genes is a prerequisite for forming protein–protein interaction sites. Next, we tested the influence of gene functionalities. Previous studies have grouped different functional keywords according to their ordered and disordered tendencies.^{45–47} In particular, proteins involved in signal transduction, regulation, protein transport and development and differentiation-related processes were shown to be more disordered as compared to the proteins which mainly function in ion transport, metabolic and enzymatic activities.^{4,45–47} When we analyzed the functional association of different groups of tissue-selective genes we noticed that genes that are selectively expressed in the testes, brain, and ovaries are enriched with disorder-related functions (cell cycle, reproductive processes, signaling, regulation, and cell differentiation, *etc.*) while the genes that are expressed mainly in the liver and kidneys are enriched with terms that rely on globular proteins (ion transport, transmembrane transport, metabolic processes, and regulation of metabolic processes, *etc.*) (Supplementary file 4, ESI†). These inherent biases towards disorder related functions

may also account for the higher disorder content among the former groups of tissue-selective genes.

Discussion

Analysis of the gene expression pattern across tissues and organs was considered to be crucial for the understanding of human disease and biology. Expression levels can provide important clues about the phenotypes and functionalities of genes across different tissues and their regulatory mechanisms.^{23,30,31} Although disordered proteins are considered as a predominant class, specifically among higher eukaryotes,^{2,11,16,17,19} to date little attention has been paid to investigating their gene expression profile. In this study, we retrieved high-throughput gene expression data for more than 15 000 human proteins from published literature and analyzed their gene expression signature across 32 normal human tissues. Since disordered proteins are vulnerable towards protein aggregation, previously it was suggested that cells need

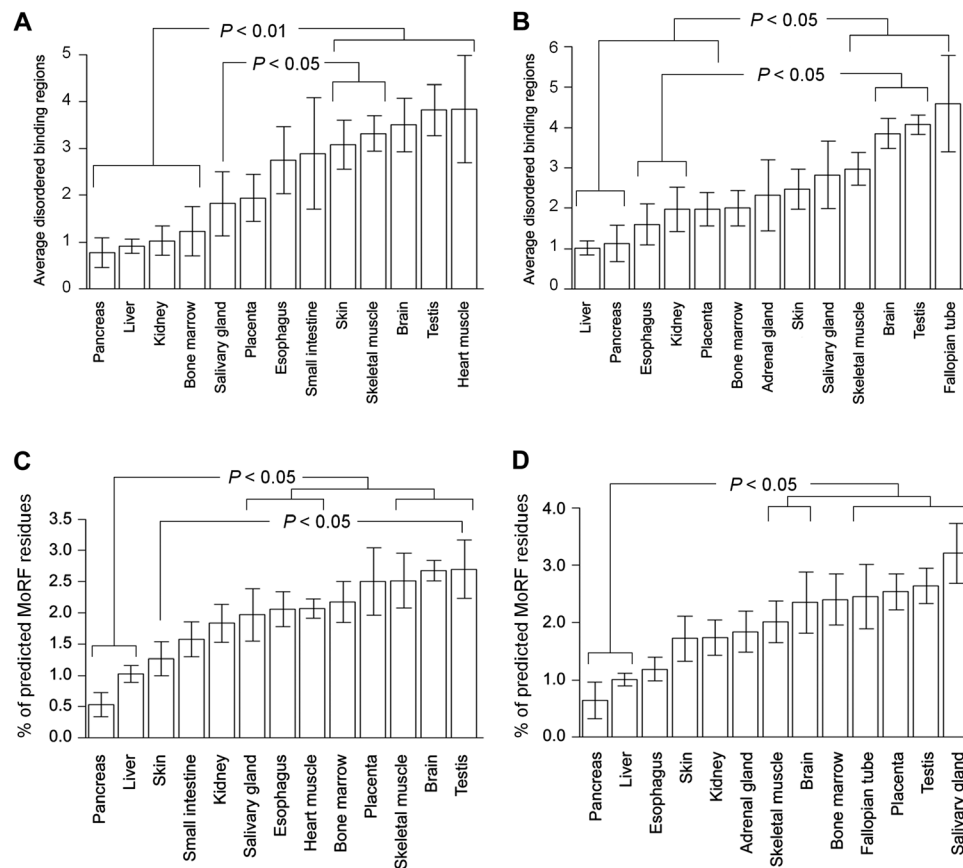


Fig. 5 Average protein disordered binding regions of different groups of tissue enriched genes. Only tissues with 30 or more tissue enriched genes were shown. (A and B) Using ANCHOR for (A) Uhlen *et al.* and (B) Chang *et al.*, and Greco *et al.* datasets. (C and D) Using fMoRFpred for (C) Uhlen *et al.* and (D) Chang *et al.*, and Greco *et al.* datasets. Significant differences between the different groups of tissue selective genes were evaluated through the Kruskal Wallis *H* test shown with *P*-values.

intricate regulatory mechanisms to maintain their concentration below a certain limit.^{25,26,48} However, here we did not find any general trend of low expression of disordered proteins except in a few specific tissues (Fig. S2–S6 and Table S1, Supplementary file 1, ESI†). Moreover, our results suggested that in a number of human tissues disordered proteins tend to be expressed at a higher level than ordered globular proteins. Based on the tissue averaged gene expression intensity, previously Gsponer *et al.*,²⁶ have shown that human disordered proteins tend to be expressed at a lower level than globular proteins. Consequently, disordered proteins were shown to contain a higher proportion of ubiquitination and micro-RNA target sites and high mRNA decay rates suggesting a complex association between gene expression level and protein intrinsic disorder content.²⁵ Considering the mean gene expression level of 32 human tissues, here we observed a similar trend. However, we did not find any significant difference when we calculated mean values without considering three tissues (liver, pancreas, salivary gland) which may imply that the lower tissue-averaged values are caused by the low gene expression level of disordered proteins in some specific tissues. Previous studies suggested that longer genes tend to be expressed at a lower level than shorter genes.^{41,42} Accordingly, here we noticed a similar trend in each and every individual tissue considered in this study.

Considering this, together with the fact that moderately and highly disordered proteins are longer than ordered globular proteins (Fig. 2) here we assumed that protein length may have some influence on the correlation between protein disorder and gene expression level. This became clear from partial correlation analysis where the correlation between gene expression and disorder content disappeared after controlling protein length. In addition, comparing the expression levels of genes having a similar protein length, no significant difference was observed between the disorder groups, suggesting that the protein length, rather than protein disorder content is the major determinant of gene expression level here. Therefore, overall this study suggests that the previously accepted impression that disordered proteins are expressed at a lower level than ordered protein holds true for only a few tissues, and is mostly influenced by their higher protein length.

In the next part, we tried to explore the functional significance of disordered proteins in human tissues by considering the disorder content of tissue-enriched genes. Previously, great interest has been paid to characterizing different human tissues in terms of their transcriptome profile.^{22,24,49,50} These studies suggested that most, if not all, of the human tissues express a few genes predominantly which are crucial for maintaining

their functional differences with other tissues as well as for their development and differentiation.^{22,24,30,31,43} Several methods have been proposed earlier to evaluate whether a gene has an affinity to be expressed in a particular tissue selectively.⁵¹ To underscore the proteins that are important for the proper functioning of different human tissues here we considered two such approaches and identified the genes which show predominant expression in each tissue individually. Functional analysis of selectively expressed genes for the tissues where we found an adequate number of such genes suggested an overall concurrence with the function of the respective tissues. Comparing their protein disorder content, here we noticed a higher fraction of disordered residues among the genes expressed mainly in the testes, brain *etc.* tissues as compared to those expressed predominantly in the liver, pancreas, kidney *etc.* tissues suggesting that disordered proteins may have important functional consequences for the former group of tissues. Consequently, our analysis suggested that the proteins encoded by the former group of tissue-selective genes interact with a higher number of partners in their protein interaction network than the latter group of tissue-selective genes (liver, pancreas, kidney, esophagus, *etc.*). Disordered proteins provide internal flexibility during protein–protein interaction and facilitate promiscuous binding.^{1,2,11} Therefore, highly connected proteins (hub-proteins) were shown to be enriched with intrinsically disordered regions.¹¹ Higher protein disorder among the former group of tissue-selective genes may suggest that disordered regions are crucial to maintain their higher protein connectivity. In order to further explore the role of disordered proteins in tissue functionalities, here we carefully examined the presence of disordered binding sites among the different groups of tissue-selective proteins. Disordered proteins interact through fly-casting mechanisms where they undergo folding upon binding. Disordered binding regions act as elementary units in molecular recognition that facilitate high-specificity and low-affinity interaction, a specific signature of disordered proteins.³² Thus, the higher proportion of disordered binding regions among the former group of tissue-selective genes (Fig. 5) may be considered as an indication that disordered residues help these tissues to sustain their functional specificity by providing structural flexibility for binding promiscuity. We also observed that in tissues where tissue-selective genes are enriched in protein disorder, the disorder associated functions like cell cycle, reproductive processes, signaling, regulation, and cell differentiation, *etc.* are overrepresented. In contrast, in tissues having low disorder content in tissue-selective genes, the globular protein-associated terms like ion transport, transmembrane transport, metabolic processes, and regulation of metabolic processes, *etc.* are overrepresented^{1–5} (Supplementary file 4, ESI†). Our results suggested a strong deterioration in mean gene expression level of disordered proteins only in the liver and kidneys. The liver is the most metabolically active tissue in the human body⁵³ and the kidneys are also associated with the elimination of metabolic wastes. The pancreas is composed of both endocrine and exocrine glands whose main function is to produce enzymes and hormones.⁵⁴ Functional analysis of the genes specific to these

two tissues suggested that these genes are mostly involved in functions which need a relatively lower fraction of disordered residues. Indeed, when we look closely into the genes selectively expressed in these tissues >80% of the genes were found to have predicted disorder content <20% (by consensus of five algorithms). Among the liver-expressed genes, there were 11 genes (AADAC, ADH6, CFHR3, CFI, CPB2, CYP8B1, F11, FGL1, FMO3, PON1, SERPINA6) with <1% of predicted disordered residues most of which are enzymes involved in different catalytic mechanisms. Among the pancreas-enriched genes, we found six (AMY2A, AMY2B, CPA1, CPB1, CTRB2, FBXW12, GRPR, PNLIP) genes with predicted disordered residues <1% four of which encode proteins with enzymatic functions. On the other hand, >50% of genes selectively expressed in the testes and brain fall into different disorder categories with a predicted disorder content of more than 20%. The testes are male reproductive organ whose main function is to develop male-specific characteristics.⁵⁰ Most of the proteins expressed selectively in the testes are involved in spermatogenesis, a process that needs intricate regulation.⁵⁵ Genes showing elevated expression in the testes are tightly regulated starting from synthesis to degradation and are mostly involved in different types of molecular binding.⁵² Proteins involved in a binding mechanism will certainly need a high amount of disordered residues to interact with a large number of partners as we observed in our study. Among the testis-expressed genes, we noticed 15 completely disordered proteins (PAGE1, TNP1, PRM2, VCY1B, VCY, PAGE5, VCX3A, PCP2, PRM1, VCX2, TNP2, VCX, GAGE2A, VCX3B, SRRM5) which play key roles in different phases of spermatogenesis and are involved in nuclear signaling and regulatory processes. The brain is the most complex organ of the human body which expresses genes mostly associated with developmental processes and synaptic signaling.⁵⁶ Here we found 11 brain-specific genes (AMER2, FAM107A, MAPT, BAALC, ERMN, VGF, CPLX1, SRRM4, CPLX2, NRG1, MBP) with predicted disorder content >90% which are involved in various neurological processes. Altogether, our study relates to the specialized functionalities of the tissue enriched genes of both groups, from the reproductive process or the cellular differentiation in the testes⁵² to the cellular signaling indispensable for the functionality of brain⁵⁶ in the disorder-rich class and from the metabolic processes and their regulation in tissues like the liver⁵³ in the disorder poor class.

Conclusions

Disordered proteins provide flexibility in protein functionalities. Due to their binding promiscuity, IDPs are considered as hubs in protein interaction networks where they interact with several other proteins. Considering the risk associated with increased use of disordered proteins, previously it was suggested that the gene expression level of disordered proteins is tightly regulated at multiple layers of transcriptional control machinery.²⁶ However, the probability of interaction largely depends upon the availability of interacting proteins.⁵⁷ Therefore, reduction of the gene expression level of disordered proteins may prove detrimental

to the interaction network. So, the negative correlation between protein intrinsic disorder and gene expression level in humans as was obtained by previous studies seems debatable. In this study, we explored the gene expression profile of human disordered proteins across 32 normal human tissues. Our results indicated that disordered proteins do not have any definite association with gene expression levels, instead lower gene expression of these proteins resulted from their biased gene expression in some specific tissues and their higher protein length. Moreover, here we found evidence that tissues like the testes, ovaries, brain, *etc.* predominantly express genes encoding disordered residues to sustain their high protein connectivity through a higher number of disordered protein binding sites and are associated with functions that are signatures of disordered proteins.

Abbreviation

IDP	Intrinsically disordered proteins
FPKM	Fragments per kilo base of exon model per million mapped reads.
D2P2	Database of Disordered Protein Predictions
MoRF	Molecular Recognition Features
GO	Gene Ontology

Conflicts of interest

The authors declare that they have no conflict of interest.

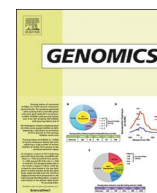
Acknowledgements

We would like to thank the anonymous reviewers for their constructive suggestions and useful comments. The authors thank Bose Institute and the Department of Science and Technology, Government of India for their financial support. The authors are also thankful to Mr Sanjib Gupta for technical help and Dr Tina Begum and Dr Sandip Chakraborty for useful discussions.

References

- 1 J. Gsponer and M. M. Babu, *Prog. Biophys. Mol. Biol.*, 2009, **99**, 94–103.
- 2 J. Habchi, P. Tompa, S. Longhi and V. N. Uversky, *Chem. Rev.*, 2014, **114**, 6561–6588.
- 3 V. N. Uversky, *Int. J. Biochem. Cell Biol.*, 2011, **43**, 1090–1103.
- 4 A. K. Dunker, I. Silman, V. N. Uversky and J. L. Sussman, *Curr. Opin. Struct. Biol.*, 2008, **18**, 756–764.
- 5 A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva and Z. Obradovic, *Biochemistry*, 2002, **41**, 6573–6582.
- 6 P. E. Wright and H. J. Dyson, *Nat. Rev. Mol. Cell Biol.*, 2015, **16**, 18–29.
- 7 G. J. P. Rautureau, C. L. Day and M. G. Hinds, *Int. J. Mol. Sci.*, 2010, **11**, 1808–1824.
- 8 S. C.-C. Chen, T.-J. Chuang and W.-H. Li, *Mol. Biol. Evol.*, 2011, **28**, 2513–2520.
- 9 A. Panda, T. Begum and T. C. Ghosh, *PLoS One*, 2012, **7**, e48336.
- 10 M. Arai, K. Sugase, H. J. Dyson and P. E. Wright, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 9614–9619.
- 11 C. Haynes, C. J. Oldfield, F. Ji, N. Klitgord, M. E. Cusick, P. Radivojac, V. N. Uversky, M. Vidal and L. M. Iakoucheva, *PLoS Comput. Biol.*, 2006, **2**, e100.
- 12 E. Schadt, P. Tompa and H. Hegyi, *Genome Biol.*, 2011, **12**, R120.
- 13 B. Xue, A. K. Dunker and V. N. Uversky, *J. Biomol. Struct. Dyn.*, 2012, **30**, 137–149.
- 14 J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones, *J. Mol. Biol.*, 2004, **337**, 635–645.
- 15 A. K. Dunker, P. Romero, Z. Obradovic, E. C. Garner and C. J. Brown, *Genome Inf.*, 2000, **11**, 161–171.
- 16 M. Y. Lobanov and O. V. Galzitskaya, *Int. J. Mol. Sci.*, 2015, **16**, 19490–19507.
- 17 Z. Peng, J. Yan, X. Fan, M. J. Mizianty, B. Xue, K. Wang, G. Hu, V. N. Uversky and L. Kurgan, *Cell. Mol. Life Sci.*, 2015, **72**, 137–151.
- 18 A. Panda and T. C. Ghosh, *Gene*, 2014, **548**, 134–141.
- 19 N. Pietrosevoli, J. A. García-Martín, R. Solano and F. Pazos, *PLoS One*, 2013, **8**, e55524.
- 20 A. Panda, S. Podder, S. Chakraborty and T. C. Ghosh, *Genomics*, 2014, **104**, 530–537.
- 21 S. Chakraborty, J. S. Byers, S. Jones, D. M. Garcia, B. Bhullar, A. Chang, R. She, L. Lee, B. Fremin and S. Lindquist, *Cell*, 2016, **167**, 369–381.
- 22 L. Fagerberg, B. M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpour, A. Danielsson and K. Edlund, *Mol. Cell. Proteomics*, 2014, **13**, 397–406.
- 23 D. Hebenstreit, M. Fang, M. Gu, V. Charoensawan, A. van Oudenaarden and S. A. Teichmann, *Mol. Syst. Biol.*, 2011, **7**, 497.
- 24 M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester and S. Hober, *Nat. Biotechnol.*, 2010, **28**, 1248–1250.
- 25 Y. J. K. Edwards, A. E. Lobley, M. M. Pentony and D. T. Jones, *Genome Biol.*, 2009, **10**, R50.
- 26 J. Gsponer, M. E. Futschik, S. A. Teichmann and M. M. Babu, *Science*, 2008, **322**, 1365–1368.
- 27 M. Uhlen, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjödelt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. Szigarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P. H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen and F. Pontén, *Science*, 2015, **347**, 1260419.
- 28 P. Flicek, M. R. Amodé, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. García Girón, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M. McLaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. J. Trevanion, A. Vullo,

- S. P. Wilder, M. Wilson, A. Zadissa, B. L. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. J. P. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. R. Zerbino and S. M. J. Searle, *Nucleic Acids Res.*, 2014, **42**, D749–D755.
- 29 M. E. Oates, P. Romero, T. Ishida, M. Ghalwash, M. J. Mizianty, B. Xue, Z. Dosztányi, V. N. Uversky, Z. Obradovic, L. Kurgan, A. K. Dunker and J. Gough, *Nucleic Acids Res.*, 2013, **41**, D508–D516.
- 30 C.-W. Chang, W.-C. Cheng, C.-R. Chen, W.-Y. Shu, M.-L. Tsai, C.-L. Huang and I. C. Hsu, *PLoS One*, 2011, **6**.
- 31 D. Greco, P. Somervuo, A. Di Lieto, T. Raitila, L. Nitsch, E. Castrén and P. Auvinen, *PLoS One*, 2008, **3**, e1880.
- 32 B. Mészáros, I. Simon and Z. Dosztányi, *PLoS Comput. Biol.*, 2009, **5**, e1000376.
- 33 Z. Dosztányi, B. Mészáros and I. Simon, *Bioinformatics*, 2009, **25**, 2745–2746.
- 34 J. Yan, A. K. Dunker, V. N. Uversky and L. Kurgan, *Mol. Biosyst.*, 2016, **12**, 697–710.
- 35 C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers, *Nucleic Acids Res.*, 2006, **34**, D535–D539.
- 36 S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. K. B. Gandhi, K. N. Chandrika, N. Deshpande and S. Suresh, *Nucleic Acids Res.*, 2004, **32**, D497–D501.
- 37 H.-W. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd and B. Weil, *Nucleic Acids Res.*, 2002, **30**, 31–34.
- 38 P. McQuilton, S. E. S. Pierre, J. Thurmond and C. FlyBase, *Nucleic Acids Res.*, 2011, gkr1030.
- 39 E. Eden, R. Navon, I. Steinfeld, D. Lipson and Z. Yakhini, *BMC Bioinf.*, 2009, **10**, 48.
- 40 E. Eden, D. Lipson, S. Yagev and Z. Yakhini, *PLoS Comput. Biol.*, 2007, **3**, e39.
- 41 A. O. Urrutia and L. D. Hurst, *Genome Res.*, 2003, **13**, 2260–2264.
- 42 C. I. Castillo-Davis, S. L. Mekhedov, D. L. Hartl, E. V. Koonin and F. A. Kondrashov, *Nat. Genet.*, 2002, **31**, 415–418.
- 43 S. Liang, Y. Li, X. Be, S. Howes and W. Liu, *Physiol. Genomics*, 2006, **26**, 158–162.
- 44 A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva and V. N. Uversky, *FEBS J.*, 2005, **272**, 5129–5148.
- 45 S. Vucetic, H. Xie, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky, *J. Proteome Res.*, 2007, **6**, 1899.
- 46 H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky and Z. Obradovic, *J. Proteome Res.*, 2007, **6**, 1882.
- 47 H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky, *J. Proteome Res.*, 2007, **6**, 1917.
- 48 M. M. Babu, R. van der Lee, N. S. de Groot and J. Gsponer, *Curr. Opin. Struct. Biol.*, 2011, **21**, 432–440.
- 49 M. Krupp, J. U. Marquardt, U. Sahin, P. R. Galle, J. Castle and A. Teufel, *Bioinformatics*, 2012, **28**, 1184–1185.
- 50 D. Djureinovic, L. Fagerberg, B. Hallström, A. Danielsson, C. Lindskog, M. Uhlén and F. Pontén, *Mol. Hum. Reprod.*, 2014, gau018.
- 51 N. Kryuchkova-Mostacci and M. Robinson-Rechavi, *Briefings Bioinf.*, 2016, bbw008.
- 52 M. T. Anand and B. V. L. S. Prasad, *J. Hum. Reprod. Sci.*, 2012, **5**, 266.
- 53 C. Kampf, A. Mardinoglu, L. Fagerberg, B. M. Hallström, K. Edlund, E. Lundberg, F. Pontén, J. Nielsen and M. Uhlen, *FASEB J.*, 2014, **28**, 2901–2914.
- 54 M. J. Muraro, G. Dharmadhikari, D. Grün, N. Groen, T. Dielen, E. Jansen, L. van Gurp, M. A. Engelse, F. Carlotti and E. J. P. de Koning, *Cell Syst.*, 2016, **3**, 385–394.
- 55 S. R. Grimes, *Gene*, 2004, **343**, 11–22.
- 56 E. Sjöstedt, L. Fagerberg, B. M. Hallström, A. Häggmark, N. Mitsios, P. Nilsson, F. Pontén, T. Hökfelt, M. Uhlén and J. Mulder, *PLoS One*, 2015, **10**, e0130028.
- 57 A. Bossi and B. Lehner, *Mol. Syst. Biol.*, 2009, **5**, 260.



Evolutionary rate heterogeneity between multi- and single-interface hubs across human housekeeping and tissue-specific protein interaction network: Insights from proteins' and its partners' properties

Kakali Biswas^a, Debarun Acharya^a, Soumita Podder^{a,b}, Tapash Chandra Ghosh^{a,*}

^a Bioinformatics Centre, Bose Institute, P-1/12, C.I.T. Scheme VII M, Kolkata 700 054, India

^b Department of Microbiology, Raiganj University, Raiganj, Uttar Dinajpur 733134, India

ARTICLE INFO

Keywords:

Tissue-specific hubs
Housekeeping hubs
Multi-interface hubs
Single-interface hubs
dN/dS ratio
Functional divergence
Conformational diversity

ABSTRACT

Integrating gene expression into protein-protein interaction network (PPIN) leads to the construction of tissue-specific (TS) and housekeeping (HK) sub-networks, with distinctive TS- and HK-hubs. All such hub proteins are divided into multi-interface (MI) hubs and single-interface (SI) hubs, where MI hubs evolve slower than SI hubs. Here we explored the evolutionary rate difference between MI and SI proteins within TS- and HK-PPIN and observed that this difference is present only in TS, but not in HK-class. Next, we explored whether proteins' own properties or its partners' properties are more influential in such evolutionary discrepancy. Statistical analyses revealed that this evolutionary rate correlates negatively with protein's own properties like expression level, miRNA count, conformational diversity and functional properties and with its partners' properties like protein disorder and tissue expression similarity. Moreover, partial correlation and regression analysis revealed that both proteins' and its partners' properties have independent effects on protein evolutionary rate.

1. Introduction

Cells are the fundamental unit of life. Except for the unicellular ones, every living organism possesses diverse types of cells adapted to perform specialized functions. The functions of each cell are mediated by the molecular machinery, of which proteins play an essential part. Proteins interact with each other and perform almost all the fundamental life processes. Such interactions involve interfaces or domains, which execute the functions of the protein. Protein domains play a crucial part in molecular evolution since these are used as structural building blocks and may create proteins with discrete functions due to exon shuffling [1–3]. The advancement in high-throughput protein interaction data helps to analyze protein functions from the network perspective. Moreover, within the whole protein interaction network, there are some small, densely linked components formed by the interactions between proteins, nucleic acids, and other small molecules, and are weakly connected to the rest of the protein-interaction network. These components are termed as modules [4]. Recent advances in discovery and revision of the proteins in modules using computational biology have enabled us to model these protein-protein interactions as a network where proteins represent the nodes with interactions as links between the nodes.

Inside the protein-protein interactions network (PPIN), proteins with a high degree of connectivity are found to be essential and are likely to perturb the PPIN upon deletion, malfunction or misregulation [5]. These proteins, named as hub, are distinct from lesser connected proteins or non-hubs and are evolutionary more conserved. Although most of the earlier studies featuring hub proteins from evolutionary perspective compared hub and non-hub proteins in PPIN, more recent studies aim at detailed analysis of hub proteins. One such study by Han et al. classified hub proteins into two groups - multi-interface hub proteins (MI or party hubs) and single interface hub proteins (SI or date hubs) based on protein domain architecture and correlated expression of the interacting partners [6]. Comparing the evolutionary rate between these MI and SI hubs revealed discrete differences—MI proteins were found to be evolutionarily more conserved than SI proteins [7], which may be mainly due to selective constraint acting on a larger region in MI proteins, as it usually possesses more interacting surfaces. Additionally, the party hubs mediate within-module interactions (intra-module), whereas date hubs integrate between modules (inter-module) [7]. However, the SI proteins acting on various modules face stronger consequences when deleted than the less pervasive densely connected MI proteins, due to their association with diverse functions [8]. Besides, a few studies have been carried out to understand the structural

* Corresponding author.

E-mail address: tapash@jcbiose.ac.in (T.C. Ghosh).

<https://doi.org/10.1016/j.ygeno.2017.11.006>

Received 23 June 2017; Received in revised form 10 November 2017; Accepted 29 November 2017
0888-7543/ © 2017 Elsevier Inc. All rights reserved.

(conformational) and functional role of these hub proteins [9,10]. The functions of a protein are mediated mainly by its structure. Although each protein is thought to possess definite three-dimensional conformation determined by its amino acid sequence, that may not be the only conformation adopted by the protein within a cellular system [11]. The magnitude of conformational diversity encompasses structural changes like fluctuation of protein's side chains and the movement of loops and secondary structures, even to the global rearrangement in protein tertiary structure [12].

Further insights into human PPIN classified topological variation based on gene expression data. Based on gene expression breadth, all genes are grouped as either tissue-specific (TS), or housekeeping (HK). Previous studies revealed many differences between the HK and TS genes in humans. Human HK genes are more compact in structure, containing shorter intron length, 5' UTR length and coding sequence length [13]. Consistent with this, HK genes are enriched in shorter repetitive sequences such as Alu-elements, but depleted in longer repetitive sequences like Long Interspersed Nuclear Element 1 (LINE-1) elements [14]. Additionally, elucidation of evolutionary rate differences among these two groups resulted in similar findings across organisms as diverse as unicellular fungi to humans, the housekeeping genes (HK) evolve slower than tissue-specific genes (TS) [15]. Accordingly, the whole PPI network was also grouped into tissue-specific or local network and housekeeping or global network, where TS hubs (TSH) evolve faster than HK hubs (HKH). These TSH also feature longer genes, less protein expression abundance, tight regulation and greater protein intrinsic disorder content than HKH [54]. Additionally, within the PPI network, HK genes are more central and are associated with core cellular processes whereas TS genes are more peripheral with modified core cellular processes as well as regulatory and developmental functions [16–18]. However, these findings remain confounding as some TS genes are reported to evolve slower than even this HK class of genes [19–21]. To address this issue, Podder et al. classified human proteins into MI and SI counterparts and analyzed the evolutionary rate of TS and HK genes between these two groups. They found that within MI proteins, both TS- and HK-genes show similar evolutionary rates, whereas, within SI proteins, HK genes evolve slower than TS genes [10]. Furthermore, recent studies based on PPI-network properties highlights the impact of the partner proteins on proteins' evolutionary rate [16,22], as the interacting partners also contribute to the central node evolution via the domain-domain interaction [23]. Such analysis on HK- and TS-hubs revealed that interacting partners of the TSH are more conserved than HKH with diverse subcellular localization [22]. However, these studies lack detailed insights into the protein interaction network-based properties and the influence of interacting partners on the evolutionary rate. Therefore, a detailed spatially resolved analysis is required to explain the evolutionary rate variation between these two hub classes.

In this study, we delved deep into the understanding of protein evolutionary rate based on their expression breadth (whether housekeeping or tissue-specific) and the contribution of domain number (whether single or multiple) to it. We tried to identify at which level the evolutionary conservation endures. Furthermore, we sought to explore which among the two: protein's own property or partner properties influence the evolutionary rate of proteins the most.

2. Materials and methods

2.1. Retrieval of dataset

We obtained tissue specific gene expression data from EMBL-EBI expression atlas (<https://www.ebi.ac.uk>) for “baseline” expression where the expression level of each gene in normal and untreated conditions. Then we calculated tissue specificity index τ [24] of each gene for tissue specificity using the following formula [10]—

$$\tau = \frac{\sum_{j=1}^{\eta_H} \left(1 - \left[\frac{\log_2 S_H(i,j)}{\log_2 S_H(i, \max)} \right] \right)}{\eta_H - 1}$$

(where, η_H = number of human tissues examined and $S_H(i, \max)$ = highest expression signal of gene i across the η_H tissues). The value ranges from 0 to 1, where genes with τ -values close to ‘0’ are considered to be more towards housekeeping and those with τ -values close to ‘1’ are considered as tissue-specific (TS). $\tau = 0$ represents equal expression of the gene across all tissue, i.e. housekeeping (HK) genes. We sorted our dataset according to an increasing τ values and obtained genes from extreme 20% of the population from both ends. Thereby, we obtained 1198 HK and 7767 TS genes.

2.2. Protein connectivity data retrieval and interacting domain identification

Protein-protein interaction data was obtained from BioGRID (release 3.4.130) (<https://thebiogrid.org/>) [25]. Genes with at least five interacting partners were considered to be highly connected or hub proteins. We obtained human protein sequences from the UCSC genome browser (<http://genome.ucsc.edu>). Interacting domains were retrieved from Pfam repository (<http://pfam.sanger.ac.uk/>) [55]. The hypothesis behind the Pfam data retrieving was that the interacting domains confer binding capability to protein regions. The cut-off values used for domain assignment are (1) e-value of alignment $e < 1.0 \times 10^{-4}$; (2) domain length > 12 ; (3) matched sequence length $> 80\%$ of domain length [26]. In particular, single interface proteins were designated as having few interaction interfaces (two at most) and multi-interface proteins having more than two interacting interfaces [27]. The numbers of HKH_MI and HKH_SI proteins are respectively 303 and 895. The numbers of MI and SI proteins belonging to TSH PPIN are 1705 and 6062, respectively.

2.3. Estimation of evolutionary rate

The evolutionary rates of human genes were calculated by dividing non-synonymous substitution rate (dN) with synonymous substitution rate (dS). The dN and dS values were retrieved from BioMart interface of Ensembl Version 87 (<http://www.ensembl.org/biomart/martview>) [28] for *Homo sapiens* (GRCh37) using one to one Human-Mouse as well as Human-Chimpanzee orthologous pairs.

2.4. Prediction of miRNA targets sites and gene expression level assessment

The number of miRNA targets per gene were obtained from TargetScan (release 6.2) (<http://www.targetscan.org>) [29] for its more reliable data over other databases. Tissue-wise RNA-seq gene expression data was obtained from the human protein atlas [30]. Average gene expression level of HK genes was calculated by considering only those tissues where it shows higher than mean expression level calculated for all tissues. Expression level for TS genes represents only the tissue where the desired gene is expressed at its highest level.

2.5. Collection of conformational and functional annotation

Protein conformational diversity data was acquired from CoDNAs database [31]. The database utilizes a total of 70,467 PDB structures (Protein Data Bank, a repository of biological macromolecular structure) [32], representing a set of 9398 monomeric proteins of the protein data bank. Conformational diversity was measured as the maximum RMSD (root-mean-square deviation measuring the average distance between the superimposed atoms) between available conformers of a protein. RMSD values were normalized to RMSD100 for all proteins with > 40 residues [33]. This provided us with 1094 human proteins with corresponding conformational diversity values.

Next, we acquired the protein-coding human genes with functional annotation from Ensembl Genome Browser (<http://www.ensembl.org/>) [28] for “biological process” GO classification for individual gene and its paralog. Functional divergence was determined using Czekanowski-Dice distance formula [34].

Functional distance (i, j)

$$= \frac{\text{Number of (Terms (i) } \Delta \text{ Terms (j))}}{[\text{Number of (Terms (i) } \cup \text{ Terms (j))} + \text{Number of (Terms (i) } \cap \text{ Terms (j))}]}$$

Here, i represents GO terms of individual human genes, j represents GO terms of the paralogous genes, Δ corresponds to the symmetrical difference between the GO term sets of two genes, \cup and \cap represents the non-redundant and common GO terms, respectively.

2.6. Protein disorder content estimation

Protein disordered residues were predicted from one of the top disorder predictors: IUPred algorithm [35,36]. It provides a fair estimation of disorder residue by assigning disorder tendency score for each residue by their ability to form favorable pair-wise contacts with neighboring amino acids [37]. Protein disorder content was defined as the fraction of the total number of such disordered residues within a protein. Moreover, flexible loops were trimmed down from the calculation by taking only 30 or more consecutive predicted disordered residues at a stretch. Other stretches were denominated as ordered regions [38].

2.7. Protein tissue expression similarity calculation

As described earlier, proteins were designated as tissue-specific or housekeeping depending on their τ (tau) value. Furthermore, the name of each tissue where the protein is expressed with the highest level of expression along with the higher bin of tau value was denoted for that tissue-specific (Top 20%). Now, these tissue names for each gene data was integrated with protein-protein interaction data among protein and its partner. Tissue expression similarity between a protein (y) and its interacting partner (z) was calculated as

Tissue expression similarity (y, z)

$$= 1 - \frac{\text{Number of (Tissues (y) } \Delta \text{ Tissues (z))}}{[\text{Number of (Tissues (y) } \cup \text{ (Tissues (z))} + \text{Number of (Tissues (y) } \cap \text{ (Tissues (z))}]}$$

Here, ‘Tissues(y)’ and ‘Tissues(z)’ represent the name of tissues where the protein ‘y’ and its interacting partner ‘z’ is expressed, respectively., Δ corresponds to the symmetrical difference between the tissues where the two proteins are expressed, \cup and \cap represents the nonredundant and common tissues, where proteins ‘y’ and ‘z’ are expressed.

2.8. Statistical analyses

All the statistical tests were performed using the SPSS (20.0) package [39]. Non-parametric Spearman's correlation test was used to evaluate the correlation coefficient between two parameters. Difference between parameters was calculated with Mann-Whitney U test. Linear and categorical regression analysis was performed using ANOVA model for understanding the relationship of the parameters with dN/dS ratio.

3. Results

3.1. Analysis on evolutionary rate difference among different hub protein classes

In this study, we explored the effect of tissue-specificity in modulating the evolutionary rate differences of human multi- and single-interface hubs. Previous studies suggest that highly connected or hub-proteins evolve slower than lowly connected or non-hub proteins [40].

Additionally, housekeeping genes are well-known for their slower evolutionary rate (than the tissue-specific genes) and so are the multi-interface hubs (than the single-interface hubs) [15,27]. In our analysis, we used the high-throughput RNA-seq data from the Human Protein Atlas [30] to obtain housekeeping and tissue-specific genes and physical protein interaction data from Biogrid [25] and co-expression data from [30] to obtain multi- and single-interface proteins and achieved similar trends for both the cases, that is, multi-interface (MI or party-hubs) evolve slower than single-interface (SI or date hub) proteins (Table 1). However, the evolutionary rate differences between MI and SI proteins within the housekeeping and tissue-specific groups are not yet clear. Therefore, we subdivided human housekeeping-(HKH) and tissue-specific hub (TSH) genes into MI and SI proteins and obtained four classes: HKH_MI, HKH_SI, TSH_MI and TSH_SI (Supplementary Tables S1A and S1B in Supplementary File 1). Comparing the evolutionary rate differences between MI and SI proteins in HKH and TSH groups using dN/dS ratio revealed that significant difference exists in the case of TSH_MI and TSH_SI (TSH_MI < TSH_SI) but not in the case of HKH_MI and SI (Table 1). However, as the MI proteins contain larger regions under selective constraint, we investigated the influence of protein size on our findings. We classified the proteins in our dataset into ‘Small’ and ‘Large’ classes depending on the median protein length. We found that the protein length has no influence in our dataset as the trend remains the same in both the length bins (Fig. 1). To explain this further, we studied the most probable parameters leading to such

Table 1

Average dN/dS ratio of different hub-proteins calculated using Human-Mouse and Human-Chimpanzee orthologs. P-value indicates significance level derived from Mann-Whitney U test [“*” denotes significant differences].

Orthologous gene pair	Category	Average dN/dS	Significance level	
A. Difference between the evolutionary rate (dN/dS ratio) of tissue-specific hubs (TSH) and housekeeping hubs (HKH) using mouse and chimpanzee as outgroups				
Human-Mouse	TSH (n = 3691)	0.158	$P = 1.00 \times 10^{-6}$, $\alpha < 0.001$	
	HKH (n = 457)	0.094		
Human-Chimpanzee	TSH (n = 5248)	0.332	$P = 9.50 \times 10^{-4}$, $\alpha < 0.001$	
	HKH (n = 449)	0.287		
B. Difference between the evolutionary rate (dN/dS ratio) of multi-interface (MI) and single-interface (SI) hubs using mouse and chimpanzee as outgroups				
Human-Mouse	MI (n = 890)	0.133	$P = 1.00 \times 10^{-6}$, $\alpha < 0.001$	
	SI (n = 3258)	0.156		
Human-Chimpanzee	MI (n = 1292)	0.293	$P = 1.00 \times 10^{-6}$, $\alpha < 0.001$	
	SI (n = 4405)	0.339		
C. Difference between the evolutionary rate (dN/dS ratio) of MI- and SI-hubs within TSH and HKH genes using mouse and chimpanzee as outgroups				
Human-Mouse	TSH_MI (n = 770)	0.141	$P = 1.00 \times 10^{-6}$, $\alpha < 0.001$	
	TSH_SI (n = 2921)	0.163		
	HKH_MI (n = 120)	0.832		$P = 6.33 \times 10^{-2}$, $\alpha > 0.05$
	HKH_SI (n = 337)	0.991		
Human-Chimpanzee	TSH_MI (n = 1167)	0.214	$P = 1.00 \times 10^{-6}$, $\alpha < 0.001$	
	TSH_SI (n = 4081)	0.284		
	HKH_MI (n = 125)	0.132		$P = 4.38 \times 10^{-1}$, $\alpha > 0.05$
	HKH_SI (n = 329)	0.285		

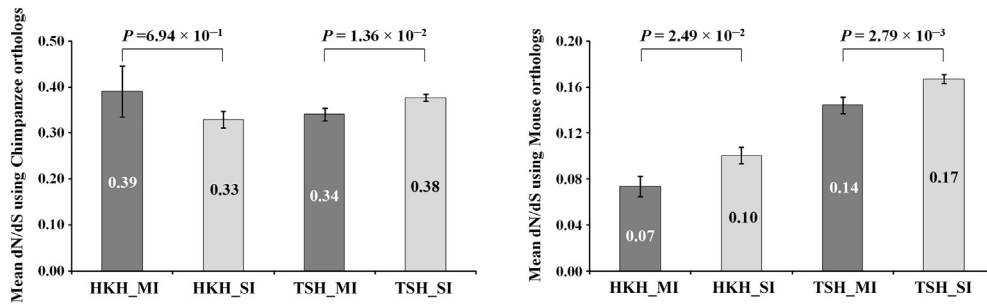
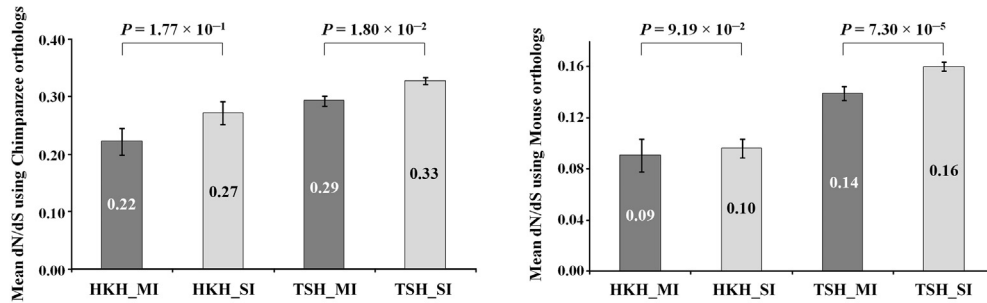


Fig. 1. Evolutionary rate (dN/dS ratio) differences between MI and SI proteins of HKH and TSH class of genes in 'small' (protein length = 24 amino acids - 473 amino acids; $N = 3383$) and 'large' (protein length = 474 amino acids - 8924 amino acids; $N = 3391$) proteins based on the median protein length (= 474 amino acids).

A. Small Proteins



B. Large Proteins

consequence. We focused on not only the protein but also their interacting partners' structural as well as functional properties guiding such evolutionary rate differences.

3.2. Role of gene expression and regulation on evolutionary rate of MI and SI hubs

One of the major determinants of protein evolution is expression level. Highly expressed genes evolve slowly, a phenomenon known as E-R anti-correlation [41]. We calculated the average expression level of MI and SI hubs across HKH and TSH classes and observed that MI proteins are significantly more highly expressed than SI proteins in TSH class whereas, no significant difference in average expression level was observed in these two groups of HKH class (Table 1, Fig. 2). Additionally, genes with a higher number of miRNA targets are also reported to be conserved [15]. Consistent with this, we found that

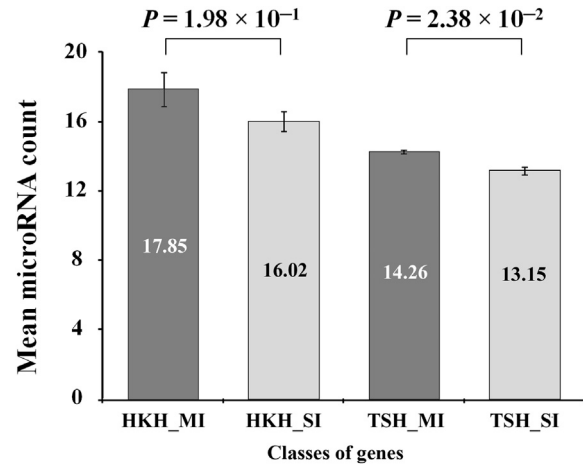


Fig. 3. Average number of miRNA target per gene difference among MI and SI proteins of HKH and TSH class of genes. P value indicates significance level derived from Mann-Whitney U test.

TSH_MI hubs are targeted by significantly more miRNAs than TSH_SI hubs (Fig. 3). However, the HKH_MI and HKH_SI did not show any difference in average miRNA number among them. Together, these results may serve as a probable reason for evolutionary discrepancy among TSH_MI/SI and HKH_MI/SI.

3.3. Role of structural and functional properties on the evolutionary rate of MI and SI hubs

Recent studies on protein functions primarily focused on this conformational diversity of proteins [11], which is found to be negatively correlated with evolutionary rate [42], mainly because it increases the functional diversity of proteins. Hence, we look for the conformational diversity of MI and SI hubs present in HK and TS PPIN. Accordingly, we found that MI proteins possess a significantly higher conformational diversity than SI proteins only for TSH class and not the HKH counterpart (Table 2).

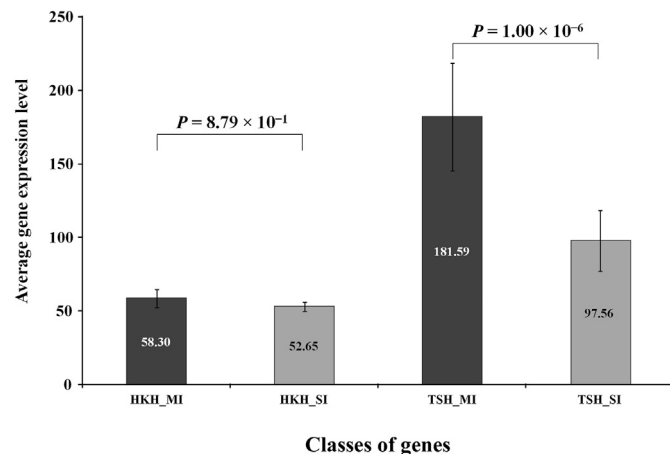


Fig. 2. Average gene expression level difference among MI and SI proteins of HKH and TSH class of genes. P value indicates significance level derived from Mann-Whitney U test.

Table 2

Average values for structural and functional properties of TSH and HKH. *P*-value indicates significance level derived from Mann-Whitney *U* test [‘*’ denotes significant differences].

Parameters	Classes of genes	Average value	Significance level
Average conformational diversity	TSH_MI (n = 198)	1.35	$P = 3.38 \times 10^{-2}$, $\alpha < 0.05$
	TSH_SI (n = 229)	1.21	
	HKH_MI (n = 84)	1.33	$P = 4.54 \times 10^{-1}$, $\alpha > 0.05$
	HKH_SI (n = 137)	1.21	
Average functional diversity per gene	TSH_MI (n = 427)	0.65	$P = 8.89 \times 10^{-3}$, $\alpha < 0.01$
	TSH_SI (n = 1267)	0.61	
	HKH_MI (n = 43)	0.65	$P = 3.72 \times 10^{-1}$, $\alpha > 0.05$
	HKH_SI (n = 82)	0.67	
Average core function per gene	TSH_MI (n = 820)	2.14	$P = 3.19 \times 10^{-2}$, $\alpha < 0.05$
	TSH_SI (n = 2602)	2.05	
	HKH_MI (n = 209)	2.44	$P = 4.29 \times 10^{-1}$, $\alpha > 0.05$
	HKH_SI (n = 548)	2.32	

Additionally, protein functional diversity between the paralogous pairs has long been treated as one of the key guiding factors of protein evolution [43–47]. Although gene duplication initially leads to the relaxation of purifying selection, the subsequent functional divergence between paralogs imposes selective constraints and slows down the evolutionary rate [48,49]. In this study, we noticed that MI proteins have a significantly higher functional divergence than SI proteins within the TSH class but not in HKH class (Table 2), indicating the selective constraints are higher for MI-TSH groups, which may be the underlying cause of their slower evolutionary rates. Furthermore, it was also reported that proteins performing core biological processes like metabolism, protein synthesis and its transport are largely conserved across species compared to the proteins involved in more regulatory processes like transcription factor binding or signal transduction [47]. Using gene ontology (GO) terms for the GO domain ‘biological process’ (GO-BP) [47] we noticed that number of core functions differ in MI and SI only within TSH but not in HKH (Table 2). However, the number of regulatory functions does not differ between the MI and SI proteins within both TSH and HKH classes. Thus, differences in conformational diversity along with functional diversity and association with core functions may impose higher selection pressure on TSH_MI compared to

TSH_SI, whereas such differences are not attributable to MI and SI classes of HKH proteins.

3.4. Role of tissue-specificity similarity and protein intrinsic disorder content of protein partners on its evolutionary rate

Tissue-specific proteins making fewer interactions evolve faster than highly interacting housekeeping proteins [16]. An earlier study also deciphered the influence of interacting partners’ properties on a protein’s evolutionary rate [50]. Additionally, analysis of TSH and HKH genes’ partners revealed that partners of TSH genes evolve slower than partners of HKH genes [22]. Thereby, we sought to investigate whether the tissue distribution of TSH genes and their interacting partners has any role in evolutionary rate. To do this, we constructed a tissue-specific similarity index according to the protein and its partner’s tissue expression profile (explained in the Materials and methods section). Interestingly, we obtained a negative correlation ($\rho = -0.189$, $n = 4128$, $P = 1 \times 10^{-6}$) between tissue expression similarity with evolutionary rate, which also demonstrated that when a gene and its’ interacting partner have a higher tissue-expression similarity, they are evolutionary more conserved than gene having interacting partners with lower tissue expression similarity (Fig. 4). Almost all of the housekeeping genes share similar tissue similarity with their partners as they are ubiquitously expressed in all tissue types.

Moreover, in an interaction network, SI proteins are more disordered than MI proteins [26,51] and perform transient interactions with their partners. However, when the interacting partners’ intrinsic disorder content was analyzed in both TSH and HKH, we found significantly higher protein disorder content in interacting partners of TSH_SI than that of TSH_MI. Such a significant difference was not observed between the two HKH groups (Fig. 5). Thus, both partner proteins’ tissue expression similarity, as well as intrinsic disorder content, may impact on a dissimilar evolutionary rate between TSH_MI and TSH_SI.

3.5. Influence of the studied factors on evolutionary rate

To examine whether each of the above mentioned parameters has a significant influence on evolutionary rate, we performed Spearman’s rank correlation analysis by considering dN/dS as scalar dependent variable and all other parameters as explanatory variables. We found that dN/dS ratio upholds significant negative correlations with mean miRNA count, expression level, conformational diversity, functional diversity, core functional processes, domain similarity, partners’ average disorder content and tissue similarity with partners (Table 3). Next, we intend to find out whether protein’s own properties or its partners’ properties are more influential in guiding protein evolutionary rate or if they act in a mutually exclusive way. For this we have performed partial correlation analysis in two ways— we have controlled all

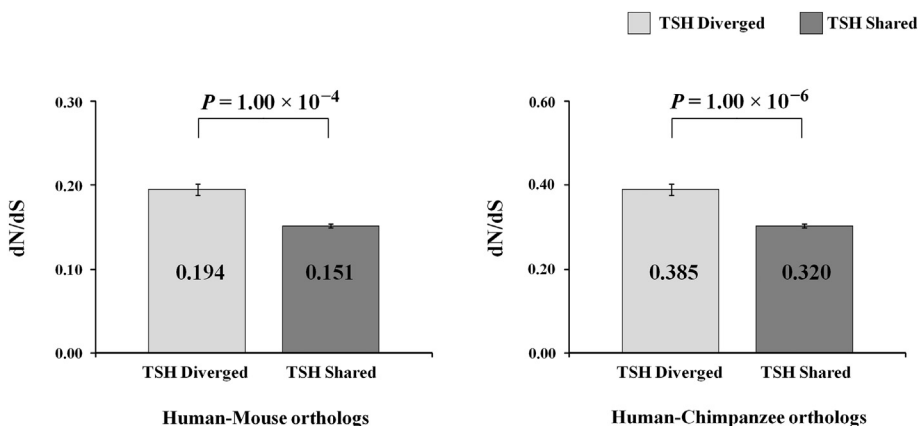


Fig. 4. Evolutionary rate (dN/dS ratio) differences between tissue-specific genes with similar (TSH_{shared}) and different (TSH_{diverged}) tissue-specificity similarity with their interacting partners. Human-Mouse and Human-Chimpanzee 1:1 orthologs were used to calculate the dN/dS ratio. *P*-values are provided in the figure.

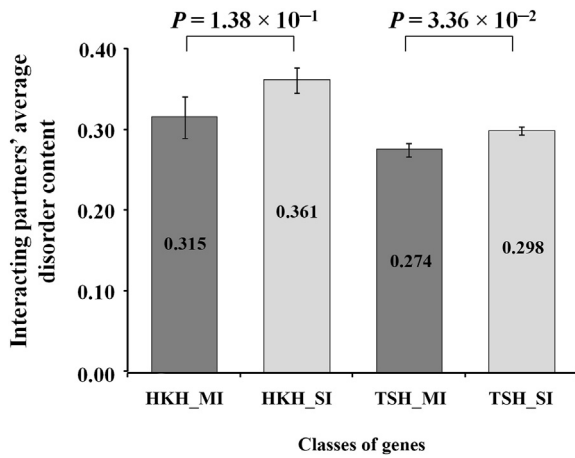


Fig. 5. Interacting partners' average disorder content for a given hub protein with difference among MI and SI proteins of HKH and TSH class of genes. P value indicates significance level derived from Mann-Whitney U test.

Table 3

Values of nonparametric correlation analysis using dN/dS ratio as a scalar dependent variable [** denotes significant differences].

Explanatory variables	Spearman's rho (ρ) correlation coefficient	Significance level (two-tailed)
Number of miRNA per gene ($n = 3519$)	-0.271	$P = 1 \times 10^{-6}$, $\alpha < 0.001$
Gene expression level ($n = 4148$)	-0.073	$P = 2 \times 10^{-6}$, $\alpha < 0.001$
Conformational diversity ($n = 4148$)	-0.105	$P = 1 \times 10^{-6}$, $\alpha < 0.001$
Average of disorder residue in interacting partner ($n = 1710$)	0.171	$P = 1 \times 10^{-6}$, $\alpha < 0.001$
Average functional divergence ($n = 742$)	-0.075	$P = 4.19 \times 10^{-2}$, $\alpha < 0.05$
Tissue expression similarity of protein and its partner (4128)	-0.113	$P = 1 \times 10^{-6}$, $\alpha < 0.001$

the partners' properties (such as partners' average disorder content and tissue similarity with partners) and noticed the correlation between protein's own properties and evolutionary rate and also vice versa. The result is delineated in (Table 4) which indicates that both the protein's and its interacting partners' properties guide the evolutionary rate in a mutually exclusive way. By using linear regression analysis we have confirmed that evolutionary rate (dN/dS ratio) of proteins are independently influenced by its own properties like number of miRNAs per gene ($\beta = -0.142$, $P = 1.45 \times 10^{-2}$) as well as protein's interacting partner properties such as tissue expression similarity ($\beta = 0.080$, $P = 4.01 \times 10^{-2}$) are also important for determining the evolutionary rate of hub proteins across housekeeping and tissue specific genes.

4. Discussion

Integrating the protein-protein interaction (PPI) network with high-throughput gene expression data, researchers divided all human PPI network into sub-network of housekeeping or global and tissue-specific or local interacting parts. Highly connected (hub) proteins within PPI network are further divided into multi-interface (MI or party-hub) and single-interface (SI or date-hubs) hubs, based on the number of their interacting interface. MI-hubs, in general, evolve slower than SI hubs

Table 4

Values of partial correlation analysis using dN/dS ratio as a scalar dependent variable [** denotes significant differences].

Explanatory variables	Correlation value (r)	Significance level
<i>Control for partners' properties</i>		
miRNA ($n = 307$)	-0.152	$P = 7.37 \times 10^{-3}$, $\alpha < 0.01$
Core functions ($n = 307$)	-0.116	$P = 4.17 \times 10^{-2}$, $\alpha < 0.05$
<i>Control for proteins' intrinsic properties</i>		
Average disorder content in partners ($n = 304$)	0.115	$P = 4.38 \times 10^{-2}$, $\alpha < 0.05$

[7] due to evolutionary constraints acting on larger surfaces. When the hub proteins from both the housekeeping (HK) and tissue-specific network (TS) were classified into MI and SI hubs, we found that MI proteins evolve slower than SI proteins in the TS PPIN, but not in the HK PPIN (Table 1), a trend slightly different from previous study [10]. Similar results were obtained after splitting all proteins in 'Small' (below-median) and 'Large' (above-median) bins, depending on their length, indicating the protein size has no significant impact on the observations (Fig. 1). As evolutionary rate exhibits a strong negative correlation with gene expression level ($\rho = -0.168$, $P = 9.75 \times 10^{-4}$), we presumed that comparison of gene expression level between MI and SI genes within HK- and TS-hubs might provide insight into their evolutionary rate difference. We found a significantly higher gene expression level in MI proteins in TSH class, whereas in HKH class, both MI and SI express at a similar level (Fig. 2). Thus, gene expression level seems to be a major determinant influencing the evolutionary rate of MI and SI proteins within HK and TS network. However, gene expression is regulated by numerous factors, of which miRNAs are a predominant regulator. Accordingly, the hub proteins are likely to have a high level of miRNA regulation with diverse local and global coordinated regulation [52]. Since regulatory stringency is supposed to be similar in all housekeeping genes, we did not get any significant difference in number of miRNA targets between MI and SI hubs. Whereas, tissue-restricted genes with diverse local sub-networks hold different regulatory constraint between MI and SI hubs, reflected by a greater number of miRNA per gene in MI/TSH proteins (Fig. 3), despite their higher gene expression, which is quite contradictory. However, our result is in agreement with the fact that genes with more miRNA target sites evolve slowly [53]. Now, the interaction between proteins in PPIN may be aided by multiple conformations of the same protein. This diverse conformation of a protein facilitates greater selection pressure on the protein-coding gene to maintain the structural domain/s via which the proteins interact. A strong negative correlation ($\rho = -0.186$, $P = 1.75 \times 10^{-4}$) between dN/dS and protein conformational diversity, as observed in our study also strengthen this hypothesis. Additionally, for duplicated genes, functional diversity between paralogs is a significant contributor to protein evolutionary rate, as it builds up selective constraints that were reduced immediately after gene duplication. It is fascinating to note that the global interacting proteins (HKH-MI and HKH-SI) with cellular maintenance purposes do not show a significant difference in conformational diversity or functional diversity. Conversely, local network of TSH proteins significantly differs in both conformational and functional diversity. This may be due to the fact that sub-networks within TS PPIN might encounter diverse selective pressure for maintaining these various expressional, conformational and functional similarities with their interacting partners.

Next, we intended to identify the contribution of interacting partners on proteins' evolutionary rate. As proteins collaborate to function as a unit, the impact of its partner on its evolutionary rate must be sought out. A significant negative correlation between tissue expression

similarities with evolutionary rate suggests that when a protein and its interacting partners possess a higher tissue expression similarity, it exhibit more evolutionary conservation, which is also supported by the differences within TSH proteins when they show similar (TSH_{shared}) and different (TSH_{diverged}) tissue-specificity similarity with their interacting partners (Fig. 4). The interacting partners of TSH_SI proteins was found to content higher protein intrinsic disorder content than TSH_MI class (Fig. 5), indicating their higher propensity to form transient tissue-specific interactions that are signatures of this group of proteins. Moreover, interactions involving proteins with lower tissue-expression similarities are also essential to maintain the connections required for the combined performance of the proteins in PPI network. Therefore, linking housekeeping and tissue-specific genes are much vital for maintaining the overall performance of a human body.

Furthermore, we performed a statistical analysis combining the impact of both proteins' own property (such as expression level, number of miRNA count, conformational diversity and other functional properties) and its partners' properties (like intrinsic protein disorder and tissue expression similarity of the interacting protein partners) on proteins' evolutionary rate. Our findings suggest that genomic novelties are more introduced by intermodular hubs or SI-hubs in the tissue-specific network only. Whereas, MI proteins remain highly conserved within this network for performing core cellular processes and are under more stringent regulation. Conversely, the housekeeping genes with greater cellular maintenance functions might not permit the HKH_SI to undergo mutation, as it could be lethal to the system. Our findings illustrate that evolutionary rate of proteins is equally governed by both its partner properties along with protein's own properties.

5. Conclusion

Our study demonstrates that lower evolutionary rate of MI hubs than SI hubs is only present in the TSH but not in HKH of human PPIN, an analysis done for the first time. We here, provide statistical evidence to establish that both structural and molecular properties of protein as well as interacting partners implicated for determining protein evolutionary rate. Thus, our study makes new findings in exploring interacting partner's properties in the conservation of global and local protein interaction networks.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ygeno.2017.11.006>.

Abbreviations

TS	tissue-specific
HK	housekeeping
PPIN	protein-protein interaction network
TSH	tissue-specific hub
HKH	housekeeping hub
MI	multi-interface protein
SI	single interface protein
GO	gene ontology
BP	biological processes
dN	nonsynonymous nucleotide substitution per nonsynonymous site
dS	synonymous nucleotide substitution per synonymous site
miRNA	micro-RNA

Contributors

Conceived and designed the experiments: KB, SP, TCG. Performed the experiments: KB. Analyzed the data: KB, DA, SP, TCG. Wrote the paper: KB, DA. All authors read and approved the final article for submission.

Conflict of interests

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the UGC: Rajiv Gandhi National Fellowship (Sanction No. RGNF-2012-13-SC-WES-32829).

References

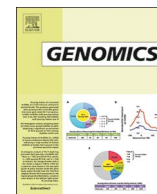
- [1] P. Bork, Building protein-structure and function from modular units, *FEBS Lett.* 286 (1991) 47–54.
- [2] H. Hegyi, M. Gerstein, The relationship between protein structure and function: a comprehensive survey with application to the yeast genome, *J. Mol. Biol.* 288 (1999) 147–164.
- [3] I.D. Campbell, A.K. Downing, Building protein-structure and function from modular units, *Trends Biotechnol.* 12 (1994) 168–172.
- [4] V. Spirin, L.A. Mirny, Protein complexes and functional modules in molecular networks, *Proc. Natl. Acad. Sci.* 100 (2003) 12123–12128.
- [5] X. He, J. Zhang, Why do hubs tend to be essential in protein networks? *PLoS Genet.* 2 (2006) 826–834.
- [6] J.D.J. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L.V. Zhang, D. Dupuy, A.J.M. Walhout, M.E. Cusick, F.P. Roth, M. Vidal, Evidence for dynamically organized modularity in the yeast protein–protein interaction network, *Nature* 430 (2004) 88–93.
- [7] H.B. Fraser, Modularity and evolutionary constraint on proteins, *Nat. Genet.* 37 (2005) 351–352.
- [8] H.B. Fraser, Coevolution, modularity and human disease, *Curr. Opin. Genet. Dev.* 16 (2006) 637–644.
- [9] B. Kahali, S. Ahmad, T.C. Ghosh, Exploring the evolutionary rate differences of party hub and date hub proteins in *Saccharomyces cerevisiae* protein–protein interaction network, *Gene* 429 (2009) 18–22.
- [10] S. Podder, P. Mukhopadhyay, T.C. Ghosh, Multifunctionality dominantly determines the rate of human housekeeping and tissue specific interacting protein evolution, *Gene* 439 (2009) 11–16.
- [11] A.M. Monzon, D.J. Zea, M.S. Fornasari, T.E. Saldano, S. Fernandez-Alberti, S.C.E. Tosatto, G. Parisi, Conformational diversity analysis reveals three functional mechanisms in proteins, *PLoS Comput. Biol.* 13 (2017).
- [12] L.C. James, D.S. Tawfik, Conformational diversity and protein evolution—a 60-year-old hypothesis revisited, *Trends Biochem. Sci.* 28 (2003) 361–368.
- [13] E. Eisenberg, E.Y. Levanon, Human housekeeping genes are compact, *Trends Genet.* 19 (2003) 362–365.
- [14] C.D. Eller, M. Regelson, B. Merriman, S. Nelson, S. Horvath, Y. Marahrens, Repetitive sequence environment distinguishes housekeeping genes, *Gene* 390 (2007) 153–165.
- [15] L.Q. Zhang, W.H. Li, Mammalian housekeeping genes evolve more slowly than tissue-specific genes, *Mol. Biol. Evol.* 21 (2004) 236–239.
- [16] A. Bossi, B. Lehner, Tissue specificity and the human protein interaction network, *Mol. Syst. Biol.* 5 (2009).
- [17] O. Souiai, E. Becker, C. Prieto, A. Benkahla, J. De Las Rivas, C. Brun, Functional integrative levels in the human interactome recapitulate organ organization, *PLoS One* 6 (2011).
- [18] X.F. Zhang, L. Ou-Yang, D.Q. Dai, W. MY, Y. Zhu, H. Yan, Comparative analysis of housekeeping and tissue-specific driver nodes in human protein interaction networks, *BMC Bioinforma.* 17 (2016).
- [19] K. Biswas, S. Chakraborty, S. Podder, T.C. Ghosh, Insights into the dN/dS ratio heterogeneity between brain specific genes and widely expressed genes in species of different complexity, *Genomics* 108 (2016) 11–17.
- [20] L.D. Hurst, N.G.C. Smith, Do essential genes evolve slowly? *Curr. Biol.* 9 (1999) 747–750.
- [21] R. Nielsen, C. Bustamante, A.G. Clark, S. Glanowski, T.B. Sackton, M.J. Hubisz, A. Flédal-Alon, D.M. Tanenbaum, D. Civello, T.J. White, et al., A scan for positively selected genes in the genomes of humans and chimpanzees, *PLoS Biol.* 3 (2005) 976–985.
- [22] M. Kiran, H.A. Nagarajaram, Interaction and localization diversities of global and local hubs in human protein-protein interaction networks, *Mol. BioSyst.* 12 (2016) 2875–2882.
- [23] Z. Itzhaki, E. Akiva, Y. Altuvia, H. Margalit, Evolutionary conservation of domain–domain interactions, *Genome Biol.* 7 (2006).
- [24] I. Yanai, H. Benjamin, M. Shmush, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification, *Bioinformatics* 21 (2005) 650–659.
- [25] A. Chatr-aryamontri, B.J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D.C. Chen, C. Stark, A. Breitkreutz, N. Kolas, L. O'Donnell, et al., The BioGRID interaction database: 2015 update, *Nucleic Acids Res.* 43 (2015) D470–D478.
- [26] P.M. Kim, A. Sboner, Y. Xia, M. Gerstein, The role of disorder in interaction networks: a structural analysis, *Mol. Syst. Biol.* 4 (2008).
- [27] P.M. Kim, L.J. Lu, Y. Xia, M.B. Gerstein, Relating three-dimensional structures to protein networks provides evolutionary insights, *Science* 314 (2006) 1938–1941.
- [28] A. Yates, W. Akanni, M.R. Amodé, D. Barrell, K. Billis, D. Carvalho-Silva,

- C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, et al., Ensembl 2016, *Nucleic Acids Res.* 44 (2016) D710–D716.
- [29] B.P. Lewis, C.B. Burge, D.P. Bartel, Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets, *Cell* 120 (2005) 15–20.
- [30] M. Uhlén, L. Fagerberg, B.M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, et al., Tissue-based map of the human proteome, *Science* 347 (2015).
- [31] A.M. Monzon, C.O. Rohr, M.S. Fornasari, G. Parisi, CoDNAS 2.0: a comprehensive database of protein conformational diversity in the native state, *Database-J. Biol. Databases Curation* Jan. 1 2016 (2016), <http://dx.doi.org/10.1093/database/baw038> (baw038).
- [32] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, *Nucleic Acids Res.* 28 (2000) 235–242.
- [33] O. Carugo, S. Pongor, A normalized root-mean-square distance for comparing protein three-dimensional structures, *Protein Sci.* 10 (2001) 1470–1473.
- [34] D. Acharya, D. Mukherjee, S. Podder, T.C. Ghosh, Investigating different duplication pattern of essential genes in mouse and human, *PLoS One* 10 (2015) e0120784.
- [35] Z. Dosztanyi, V. Csizmek, P. Tompa, I. Simon, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, *Bioinformatics* 21 (2005) 3433–3434.
- [36] Z. Dosztanyi, M. Sandor, P. Tompa, I. Simon, Prediction of protein disorder at the domain level, *Curr. Protein Pept. Sci.* 8 (2007) 161–171.
- [37] Z. Dosztanyi, V. Csizmek, P. Tompa, I. Simon, The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins, *J. Mol. Biol.* 347 (2005) 827–839.
- [38] S. Light, R. Sagit, D. Ekman, A. Elofsson, Long indels are disordered: a study of disorder and indels in homologous eukaryotic proteins, *Biochim. Biophys. Acta Proteins Proteomics* 2013 (1834) 890–897.
- [39] N.H. Nie, D.H. Bent, C.H. Hull, SPSS: Statistical Package for the Social Sciences, McGraw-Hill, New York, 1970.
- [40] I.K. Jordan, L. Mariño-Ramírez, Y.I. Wolf, E.V. Koonin, Conservation and coevolution in the scale-free human gene coexpression network, *Mol. Biol. Evol.* 21 (2004) 2058–2070.
- [41] D.A. Drummond, J.D. Bloom, C. Adami, C.O. Wilke, F.H. Arnold, Why highly expressed proteins evolve slowly, *Proc. Natl. Acad. Sci. U. S. A.* 102 (2005) 14338–14343.
- [42] D.J. Zee, A.M. Monzon, M.S. Fornasari, C. Marino-Buslje, G. Parisi, Protein conformational diversity correlates with evolutionary rate, *Mol. Biol. Evol.* 30 (2013) 1500–1503.
- [43] Gu ZL, D. Nicolae, Lu HHS, W.H. Li, Rapid divergence in expression between duplicate genes inferred from microarray data, *Trends Genet.* 18 (2002) 609–613.
- [44] X. Gu, Functional divergence in protein (family) sequence evolution, *Genetica* 118 (2003) 133–141.
- [45] R.A. Studer, M. Robinson-Rechavi, Large-scale analysis of orthologs and paralogs under covarion-like and constant-but-different models of amino acid evolution, *Mol. Biol. Evol.* 27 (2010) 2618–2627.
- [46] P. Zhang, Gu ZL, W.H. Li, Different evolutionary patterns between young duplicate genes in the human genome, *Genome Biol.* 4 (2003).
- [47] N. Lopez-Bigas, S. De, S.A. Teichmann, Functional protein divergence in the evolution of *Homo sapiens*, *Genome Biol.* 9 (2008).
- [48] I.K. Jordan, Y.I. Wolf, E.V. Koonin, Duplicated genes evolve slower than singletons despite the initial rate increase, *BMC Evol. Biol.* 4 (2004).
- [49] D. Acharya, T.C. Ghosh, Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution, *BMC Genomics* 17 (2016) 1–14.
- [50] T. Makino, T. Gojobori, The evolutionary rate of a protein is influenced by features of the interacting partners, *Mol. Biol. Evol.* 23 (2006) 784–789.
- [51] S. Podder, T.C. Ghosh, Exploring the differences in evolutionary rates between monogenic and polygenic disease genes in human, *Mol. Biol. Evol.* 27 (2010) 934–941.
- [52] H. Liang, W.-H. Li, MicroRNA regulation of human protein–protein interaction network, *RNA-Publ. RNA Soc.* 13 (2007) 1402–1408.
- [53] C. Cheng, N. Bhardwaj, M. Gerstein, The relationship between the evolution of microRNA targets and the length of their UTRs, *BMC Genomics* 10 (2009).
- [54] M. Kiran, H.A. Nagarajaram, Global versus local hubs in human protein–protein interaction network, *J. Proteome Res.* 12 (12) (2013) 5436–5446.
- [55] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall, E.L. Sonnhammer, The Pfam protein families database, *Nucleic Acids Res.* 30 (1) (2002) 276–280.



Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

The role of introns in the conservation of the metabolic genes of *Arabidopsis thaliana*

Dola Mukherjee^a, Deeya Saha^a, Debarun Acharya^a, Ashutosh Mukherjee^b, Sandip Chakraborty^a, Tapash Chandra Ghosh^{a,*}

^a Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata, 700 054, West Bengal, India

^b Department of Botany, Vivekananda College, 269, Diamond Harbour Road, Thakurpukur, Kolkata, 700063, West Bengal, India

ARTICLE INFO

Keywords:

Conservation patterns
Metabolic genes
Intron enrichment
Protein versatility
Mutational load

ABSTRACT

In *Arabidopsis thaliana*, primary metabolic genes (PMGs) are more evolutionarily conserved and intron-rich than secondary metabolic genes. We observed that PMGs are more primitive and pan-taxonomically persistent as compared to secondary (SMGs) and non-metabolic genes (NMGs). This difference in primitiveness and persistence is primarily correlated with intron number and is independent of gene expression level. We propose a twofold explanation behind higher intron enrichment in PMGs. Firstly, introns might increase protein versatility amongst PMGs through alternative splicing, providing selective advantage of PMGs and making them more persistent across diverse plant taxa. Also, multifunctional PMGs may acquire functional domains by increasing the intronic burden. Additionally, single nucleotide polymorphisms (SNPs) accumulate at a higher rate in introns as compared to exons. Moreover, a strong negative correlation between cumulative exonic SNPs density and intron number indicates that introns may protect the exonic regions against the deleterious effect of these mutations, making them more conserved.

1. Introduction

Introns are non-coding sequences that interrupt the coding regions of eukaryotic genes. They act as a hallmark of eukaryotic protein coding genes [1–3] and are important components of genome adaptation [4]. Although, spliceosomal introns are common amongst eukaryotic genomes, their density varies greatly across genomes as well as genes within the same genome [5] and deciphering the uneven phylogenetic distribution of introns is a major challenge for evolutionary genomics [4]. Understanding the function and evolution of introns have gained much attention since its discovery in the late 1970s [6]. The rapidly accumulating fully sequenced eukaryotic genomes are also allowing high-resolution reconstruction of evolutionary history of introns [7].

However, introns are thought to impose a considerable burden to the host [7], and there could be at least three possible deleterious effects on gene expression [4]: First, spliceosomal introns requires a spliceosome [7] and thus, splicing multiple introns is biologically expensive [8,9]. Second, intron transcription is costly in terms of time and energy [10–12]. Third, malfunction of any of the snRNPs will have a general detrimental effect on the cell [7]. Moreover, some studies showed that highly expressed genes are compact, especially, concerning intron size [13,14]. Finally, the mutational hazard hypothesis says that

non-coding sequences have slightly deleterious effects on fitness because of the hazard of accumulating deleterious mutations [15–17]. Thus, to minimize the mutational hazard, selection would preferentially remove the excess DNA from genomes [5].

On the other hand, some recent studies highlight various advantages of having introns [7,18]. It has been reported that introns increase the protein diversity by exon shuffling and alternative splicing [5,19,20]. Some introns also regulate gene expression [5]. Moreover, introns play a pivotal role in mRNA export, transcription coupling, splicing, etc. [21] and also give rise to non-coding RNAs that participate in regulatory processes [22]. Introns can also boost the gene expression, and this positive effect is called intron-mediated enhancement (IME) [23].

Indeed, the relationships between gene expression and intron numbers have been a matter of debate. For example, Vinogradov showed that in humans, housekeeping genes and tissue specific genes differed in their genomic complexities and regulation [24]. While the former category harbored compact, broadly and highly expressed genes, the later was tissue specific. Such observations on the properties of housekeeping genes were assessed using an older dataset. However, a different trend was observed in the model plant *Arabidopsis thaliana*, where primary metabolic genes, being mostly housekeeping in nature exhibited not only elevated expression but also higher intron number

* Corresponding author.

E-mail address: tapash@jcbiose.ac.in (T.C. Ghosh).

<https://doi.org/10.1016/j.ygeno.2017.12.003>

Received 26 September 2017; Received in revised form 6 December 2017; Accepted 8 December 2017
0888-7543/ © 2017 Elsevier Inc. All rights reserved.

[25].

Some recent studies have indicated that the relation between gene expression and introns are much more complex than previously thought. While in animals like *Caenorhabditis elegans* and *Homo sapiens*, highly expressed genes contain less and compact introns [14], in plants like *Oryza sativa* and *Arabidopsis thaliana*, it was found that highly expressed genes contained more and longer introns than genes expressed at a low level [26]. However, when the intron length between model plant and animal were compared, the introns were found to be relatively shorter in the model plant *Arabidopsis thaliana* than the mammalian mouse model, indicating the cost of transcription is negligible [4], which may favor intron retention. Previous studies indicated that variation of intron size is influenced by various factors [27]. The metabolic requirements and spatiotemporal economy might also act as selective forces to resume surplus DNA [27]. For example, house-keeping genes that are required to express at a certain level in every cell comprise shorter introns than other genes in humans [28]. On the contrary, Gorlova et al. [20] showed that evolutionary conserved and primitive genes are more functionally important and have a more intron enrichment in human, which opens up the opportunity for novel functions. Genes expressed in pollens of *A. thaliana* have smaller introns than genes expressed in sporophytes [29]. However, it is unclear to what extent the genomic configuration of plant has been shaped by functional requirement and natural selection [13].

It was earlier reported that in *Arabidopsis thaliana*, primary metabolic pathway genes contain significantly more introns than secondary metabolic pathway genes [25]. Additionally, the primary metabolic pathway genes are evolutionary more conserved than secondary metabolic pathway genes on the basis of the ratio of synonymous and non-synonymous substitution rates (d_N/d_S). However, no correlation has been found between d_N/d_S and intron number. This may not be surprising as d_N/d_S , by definition, addresses the evolutionary rate of the coding regions. Thus, the difference of intron number of these two categories of genes in *A. thaliana* is still enigmatic. So, to address this issue, we have taken a different approach here. Encouraged by the work of Gorlova et al. [20], we have introduced, in this study, two new indices named *Persistence Index (PI)* and *Age Index (AI)* to see whether this intron number variation is correlated with the evolutionary history as well as the taxonomic distribution of the concerned gene within the plant kingdom. We have taken this approach as in plants, primary metabolic pathways are almost omnipresent while secondary metabolic pathways are restricted to specific plant groups [30]. Moreover, a gene's level of evolutionary conservation reflects its functional significance [31,32].

Thus, the objective of our study is to find whether higher intron enrichment of primary metabolic pathway genes (PMGs) over secondary metabolic pathway genes (SMGs) confer any selective advantages to them which can answer the primitiveness and pan-taxonomic distribution of PMGs. For analysis of PI and AI, we have selected six other plant species along with *A. thaliana* whole genome sequences are available. These include one dicot and two monocot species, one species each from pteridophyta, bryophyta and algae. Our analysis showed that in *A. thaliana*, these two indices differ in PMGs, SMGs and NMGs (Non Metabolic pathway Genes) and both PI and AI are significantly correlated with intron number. Moreover, introns accumulate more single nucleotide polymorphisms in PMGs than SMGs as well as NMGs and may act as buffer to protect the coding region of the genes to accumulate mutations. Our study shows that introns confer some advantages for evolutionary conservation of primary metabolic pathway genes in *A. thaliana*.

2. Materials and methods

2.1. Dataset preparation

We collected the whole genome data of *Arabidopsis thaliana* from

Biomart interface [33] of Ensembl Plants [34] (<http://plants.ensembl.org/>). The metabolic gene dataset was prepared from KEGG Database (<http://www.genome.jp/kegg>) [35]. Initially, we obtained a dataset of 2512 metabolic genes out of which 2030 were PMGs and 482 were SMGs. We filtered out 209 metabolic genes from our dataset which participated in both primary and secondary metabolism. Finally, we had 1821 PMGs and 273 SMGs. The rest of the protein coding genes that did not participated in metabolism were categorised as non-metabolic genes or NMGs. We compiled a dataset of 24,903 NMGs. The complete gene list of PMG, NMG and SMG are provided in the Supplementary file 1.

2.2. Estimation of conservation of genes

We used the pan-taxonomic distribution of metabolic genes as a measure of conservation of the metabolic genes of *A. thaliana* rather than the protein level conservation. Previously, Gorlova et al. have formulated the conservation index as a measure of genes' degree of preservation [20]. The concept of Conservation index as perceived by Gorlova et al. [20] was further redefined by us as persistence index (PI) and age index (AI) to study the pan-taxonomic distribution of *A. thaliana* genes amongst the various plant taxa. PI reflects the distribution of orthologous genes of *A. thaliana* between the other plant taxa while AI denotes the primitiveness of the orthologous genes. We have detected orthologous set of genes in six of the below mentioned plant species: *A. lyrata* (dicot), *Sorghum bicolor* (monocot), *Oryza sativa* var. *japonica* (monocot), *Selaginella moellendorffii* (lycophyte), *Physcomitrella patens* (moss) and *Chlamydomonas reinhardtii* (alga) [36]. These species were ranked on the basis of their evolutionary distance from *A. thaliana*. We assigned rank 0 to those genes which are unique to *A. thaliana* while rank 6 was assigned to those genes which have orthologs on *C. reinhardtii*. Persistence index (PI) = $\sum x_i / (N - 1)$, where x_i represents the count of orthologous gene across the selected plant taxa and N is the total of plant species selected apart from *A. thaliana*. Age index (AI) = $x_i / (N - 1)$, here x_i represents the rank where the primitive most ortholog of *A. thaliana* genes could be traced. The indices value ranges from 0 to 1. '0' depicting the genes confined only to *A. thaliana* and recent origin while '1' representing the most persistent and orthologs that could be traced to all other groups and hence more primitive. To explain the indices better, we put a hypothetical example where a gene of *A. thaliana* is present in 3 other groups so its PI is 0.5, now if the most primitive ortholog could be traced to *Chlamydomonas*, and then the AI is 1. We also checked the primitiveness of the *A. thaliana* genes by using Phylostratigraphy (<https://lighthouse.ucsf.edu/proteinhistorian/>). Here, we have categorised the *A. thaliana* genes according to their phylogenetic origin into three groups-Arabidopsis (recent), Magnoliophyta (medium) and Viridiplantae (ancient).

2.3. Gene expression

Microarray expression data for *A. thaliana* was obtained from PLEXdb (www.plexdb.org/) [37]. The accession number of expression dataset is AT40 and the microarray platform used was ATH1-121501.

2.4. Intron enrichment

Both intron counts within each gene along with the intron length considered separately for studying the intron enrichment of the respective genes. The intronic coordinates were obtained from Biomart of Ensembl Plants (<http://plants.ensembl.org/biomart>).

2.5. Other genomic parameters

Intron count, intron length, transcript length, GO terms accessions and Pfam accessions were downloaded from Biomart of Ensembl Plant (<http://plants.ensembl.org/biomart>).

Multifunctionality was calculated by summing up the number of GO biological process terms assigned to each gene identifier [38]. Domain number was obtained by summing up the Pfam [39] accession against each gene identifier.

2.6. Single nucleotide polymorphism (SNPs)

Data for Single Nucleotide polymorphism (SNPs) of the genome of *A. thaliana* was also obtained from Biomart of Ensembl Plants. The coordinates of the SNPs were mapped to both the exonic and intronic positions of genes of *A. thaliana*. The mapping of coordinates was done by using in-house PERL script.

2.7. Statistical tests

Statistical analyses were performed using SPSS v.13. Mann-Whitney *U* test [40] was used to compare the average values of different variables between two classes of genes since the values were not normally distributed in our dataset. For correlation analysis, we performed the Spearman's rank correlation coefficient ρ [41], where the significant correlations were denoted by $P < 0.05$. Z-test was also carried out to study the proportion difference between groups.

3. Results and discussions

3.1. PMGs are more intron rich than SMGs and NMGs in *A. thaliana* as well as in other plant groups

A previous study showed that PMGs are more intron-rich than SMGs in *A. thaliana* [25]. Here, we have also studied the non-metabolic genes of *A. thaliana* to get a complete picture of the intronic distribution in *A. thaliana* regarding metabolic and non-metabolic genes. We have considered a total of 2094 genes as metabolic genes and 24903 genes as non-metabolic genes (NMGs). Of these metabolic genes, 1821 genes are associated with primary metabolism while 273 genes are related with secondary metabolism. It was observed that PMGs on an average have higher intron number as compared to NMGs and SMGs (Fig. 1) (Mann-Whitney *U* test, $P < 10^{-6}$).

We then, analysed whether this trend (PMGs have higher intron number than SMGs and NMGs) is present in other groups of plants too. We have studied the differences between the average intron number in NMGs, PMGs and SMGs in all the seven species (Fig. 2). For this, we have considered the PMGs, SMGs and NMGs of *A. thaliana* and their orthologous genes from the other six species. It was found that in all the

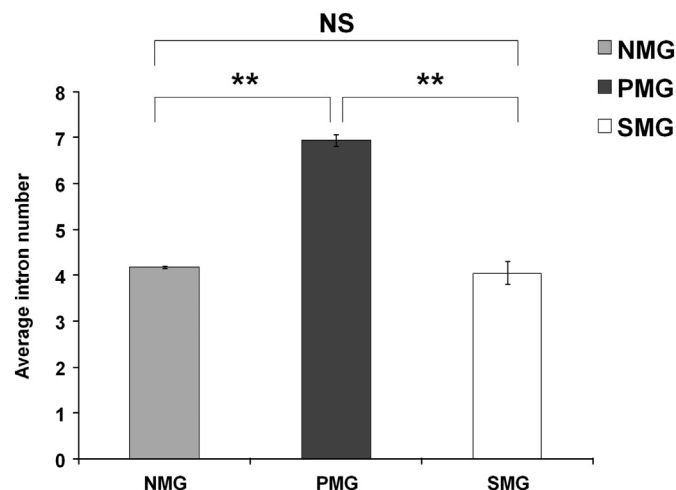


Fig. 1. Bar diagram showing the difference of intron number amongst different groups, PMGs, NMGs and SMGs. ** denotes $P < 0.01$, * denotes $P < 0.05$ and NS denotes Not significantly different values.

investigated species, PMGs always showed more introns than SMGs and NMGs. However, in *C. reinhardtii* (an aquatic alga), there is no significant differences between the three groups while in *P. patens*, *S. moellendorffii* and in the two monocot species, significant difference was found between NMGs and PMGs. However, in the two species of *Arabidopsis*, significant differences between PMGs and NMGs as well as between PMGs and SMGs have been found with respect to intron number. From these results, it can be concluded that PMGs gathered significantly more introns than NMGs or SMGs over time. It was also clear that early land plants showed a similar pattern before the monocot-dicot divergence. After that, these two groups showed significant differences with respect to intron number in PMGs, SMGs and NMGs.

3.2. PMGs are more primitive and conserved than SMGs and NMGs in *A. thaliana*

We have estimated taxonomic distribution of PMGs, SMGs and NMGs using two unique indexes, i) Persistence index and ii) Age index. It was observed that in *Arabidopsis*, protein coding genes showed a marked variation in their degree of persistence across different plant species. The persistence index as well as age index of a given gene ranges from 0 (present only in *Arabidopsis* and of most recent origin) to 1 (present in all the investigated species and genes with most ancient origin). It was observed that primary metabolic genes (PMGs) show higher level of primitiveness and persistence as compared to NMGs and secondary metabolic genes (SMGs) (Fig. 3 A and B). Although, PMGs possessed significantly higher PI as well as AI values compared to NMGs and SMGs ($P < 0.01$), SMGs did not show any significant difference of PIs and AIs as compared to NMGs ($P > 0.05$) (Fig. 3). It was observed that there was a significantly strong positive correlation (Spearman's $\rho = 0.993$, $P = 10^{-6}$, $N = 26997$) between PI and AI indicating that genes with most ancient origin are the ones that are more persistent across wide range of plant genomes. We also checked the phyletic age of the metabolic genes using Phylostratigraphy (<https://lighthouse.ucsf.edu/proteinhistorian/>). It was observed that majority of the genes of PMGs have ancestral origin than SMGs and NMGs. However, the proportion of PMGs decrease gradually with the gene age. We also observed that the NMGs are mostly quite recent in their origin (Fig. 4). As PMGs in *A. thaliana* are more ancient in terms of their origin, as showed more PI, AI and ancient phyletic origin than the other two groups and they also retained more introns over time, there must be some selective advantage of retaining more and more introns in PMGs. Therefore, from here onwards, we would investigate the role of intron number in guiding the persistence of *Arabidopsis* genes.

3.3. Intron number is correlated with persistence index in *A. thaliana*

We observed a strong positive association between persistence index and intron enrichment (Spearman's $\rho_{PI-intron\ number} = 0.297$, $P < 10^{-6}$, $N = 26997$, Spearman's $\rho_{PI-intron\ length} = 0.225$, $P < 10^{-6}$, $N = 26998$). In animal genomes, previous studies suggested that persistence of genes or its conservation is highly correlated to its intron enrichment [20,42]. In agreement with these studies, our study also found a strong association between intron enrichment and gene persistence index. In addition to it, our study also showed that genes that are older and has wider pan-taxonomic distribution, have higher intron enrichment as compared with genes of more recent origin. Next, we intended to find out whether total intron lengths or intron number was the more prominent predictor of persistence. Intron number per gene was found to correlate highly with total intron length (Spearman's $\rho = 0.809$, $P < 10^{-6}$, $N = 26997$). So, we performed a partial correlation between PI and intron number after controlling for total intron length (Spearman's $\rho = 0.113$, $P < 10^{-6}$) and observed that there was a significant impact of intron number over PI. On the contrary, when intron number was controlled and correlation between

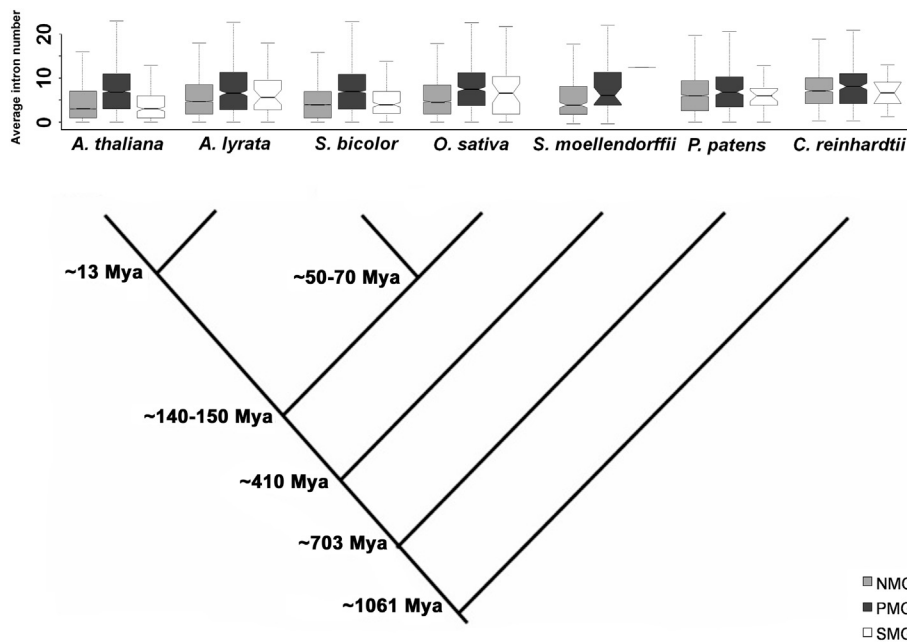


Fig. 2. Comparison of intron number between PMGs, SMGs and NMGs across diverse taxonomic plant groups. ** denotes $P < 0.01$, * denotes $P < 0.05$ and NS denotes Not significantly different values.

total intron length and conservation was noticed, it was observed that there was a very weak significant correlation between them (Spearman's $\rho = 0.044$, $P < 10^{-6}$). Thus, the effect of intron length was negligible over conservation.

3.4. Difference in conservation of PMGs, NMGs and SMGs in *A. thaliana* is independent of gene expression levels

Gene expression level has been shown to be a major determinant of protein level conservation in plants and animals [43]. Henceforth, we were interested to study the effect of intron number over gene expression level of *A. thaliana*. It has been previously proposed that intron number negatively influences gene expression level in animals [20,42]. However, we obtained a strong positive correlation between intron number and gene expression level (Spearman's $\rho = 0.253$, $P < 10^{-6}$, $N = 21049$). Our results are in agreement with previous work [26]

which also showed that highly expressed genes in plants contain more introns. Hence we were interested to study the effect of expression over conservation of PMGs. It was observed that PMGs have significantly higher expression level as compared to NMGs and SMGs. It was also observed that gene expression level positively correlates with PI (Spearman's $\rho_{PI} = 0.312$, $P < 10^{-6}$). Next, we intended to explore whether the difference of PI between PMGs, NMGs and SMGs were due to their difference in the expression level. In this context, we binned gene expression values into four bins-Bin1 (containing genes having gene expression value 2.00–5.00), Bin2 (gene expression value 5.00–8.00), Bin3 (gene expression value 8.00–11.00) and Bin4 (gene expression value > 11.00). Bin4 also showed absence of any SMGs. It was observed that in each bin, PMGs has significantly higher persistence as compared to NMGs and SMGs except Bin3 where expression level of PMGs and SMGs were insignificant (Fig. 5A). This indicates that difference of PI between PMGs and NMGs as well as PMGs and SMGs

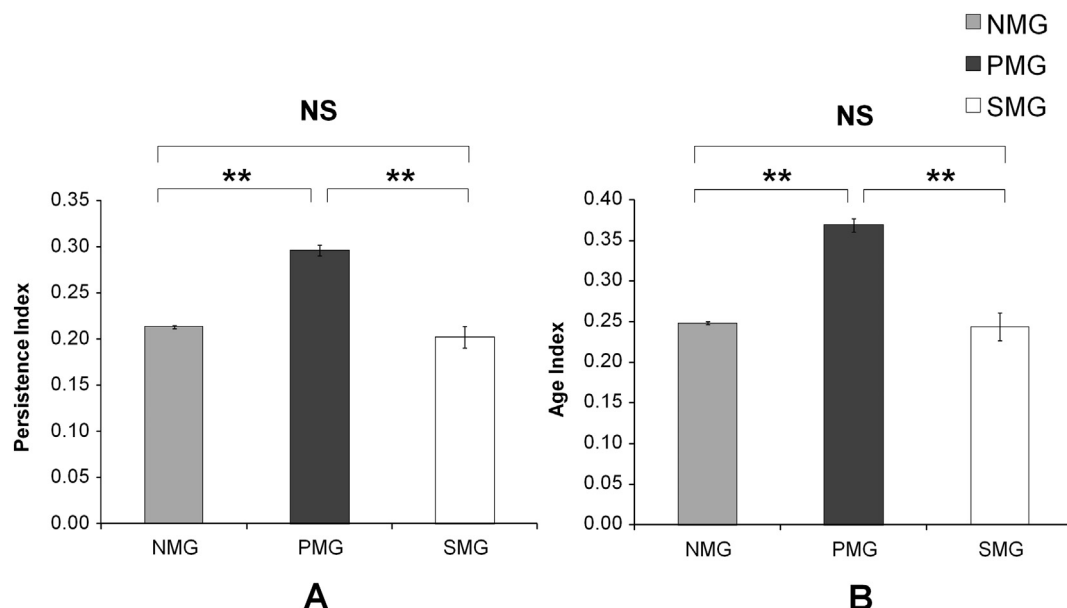


Fig. 3. Bar diagram showing the significant difference of (A) Persistence index and (B) Age Index, amongst different groups, PMGs, NMGs and SMGs. ** denotes $P < 0.01$, * denotes $P < 0.05$ and NS denotes Not significantly different values.

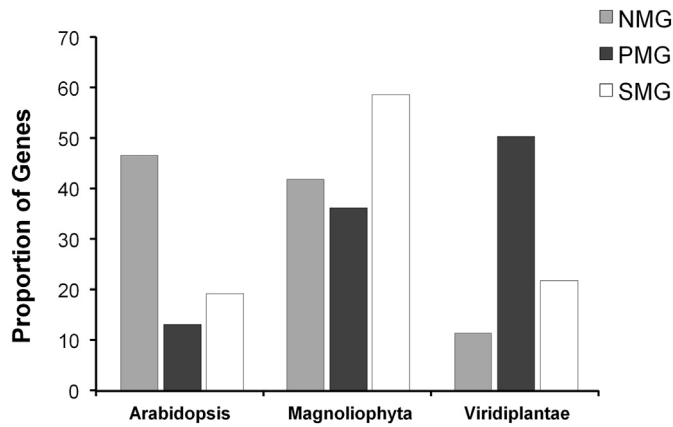


Fig. 4. Proportion of genes in three categories of phyletic age. All the proportions are significantly different (Chi-square test, $P < 0.001$).

are independent of expression level. It is also noticeable that there is a constant rise in PI up to category 3 of expression level and at category 4 (Highest expression values) there were no SMGs, so we could only compare PMGs and NMGs at this category. Interestingly, category 4 showed a lower magnitude of conservation level. This once again indicated that relationship between gene expression and conservation is non-linear. Next we analysed whether intron enrichment has an influence in governing the PI difference between PMGs, NMGs and SMGs. For this we binned intron number into four bins-Bin1 (containing genes having intron number 1–10), Bin2 (intron number 11–20), Bin3 (intron number 21–30) and Bin4 (intron number > 30). Bin 1 represented the group of genes with least number of introns, while Bin4 had genes with highest intron numbers. Bin3 and Bin4 also showed absence of any SMGs. It was observed that PI of PMGs, NMGs and SMGs did not follow any particular trend (Fig. 5B). This once again revealed that intron enrichment has a role in determining the conservation of the genes.

We propose a two-fold explanation behind such an observation. Firstly, it has been proposed earlier that older genes are under more complex regulation [44]. Introns have a definitive effect over gene expression regulation in both animals [45] and plants [46]. Hence, acquisition of large number of introns in plants could be due to the result of more complex regulation of older and highly persistent genes. Given the fact that intron number is correlated with gene conservation and intron number gradually increases with increase in degree of persistence, it is questionable that, what roles introns have in maintaining gene's degree of conservation.

3.5. Introns increase protein versatility

Previously it has been proposed that gradual segmentation of a given gene into smaller exonic regions interrupted by introns may facilitate alternative splicing and thus might increase protein versatility of the concerned gene [47]. It is quite obvious that genes with high protein diversity would tend to be more conserved than genes that yield fewer number of protein isoforms [20]. In other words, as genes grew older with time, it acquired many different number of spliced variants which increases its diversity in both transcript and protein level [48].

Intron enrichment is considerably higher in PMGs and they also acquire higher number of spliced variants as compared to NMGs and SMGs. As PMGs are older and are more persistent across diverse plant taxa, it is more likely for them to gain different molecular functions with time. It has been previously proposed that PMGs are more multifunctional in nature as compared to SMGs [25]. This high multifunctionality may be attributed to higher number of spliced variants and increased protein versatility. Previous studies have suggested intron number might increase protein versatility through alternative splicing [49]. Here we investigated that whether introns in *A. thaliana* increases protein diversity by the mechanism of alternative splicing. In this context, we estimated the number of unique proteins and unique transcripts per gene of PMGs, SMGs and NMGs respectively. It was observed that PMGs possessed significantly higher splice variant and protein diversity as compared to SMGs and NMGs (Table 1). We found that transcript and protein diversity (measured as number of unique transcript ids/protein ids) has a significant strong positive correlation with intron number (Spearman's $\rho_{\text{intron no-transcript count}} = 0.214$, $P < 10^{-6}$; Spearman's $\rho_{\text{intron no-protein count}} = 0.210$, $P < 10^{-6}$). The increased diversity in transcript and protein level could be due to alternative splicing of these genes. It was also revealed that PI is also correlated with multifunctionality (Spearman's $\rho_{\text{PI-Multifunctionality using GO biological process terms}} = 0.219$, $P < 10^{-6}$; Spearman's $\rho_{\text{PI-Multifunctionality using Pfam domain number}} = 0.017$, $P = 1.2 \times 10^{-2}$) and transcript count (Spearman's $\rho_{\text{PI-transcript count}} = 0.117$, $P < 10^{-6}$). Overall our data suggests that acquisition of large number of introns could eventually increase protein versatility through exon shuffling mechanisms which may ultimately cause conservation of genes in *A. thaliana*.

In order to gain multiple functions, primary genes might harbour elevated number of functional domains within them. It was observed that PMGs were indeed enriched in functional domains as compared to the SMGs (Mann-Whitney U Test, $P < 10^{-6}$). Thus, we hypothesize that intron enrichment in primary genes could be correlated with functional domain acquisition. In agreement to our hypothesis, we observed a significant correlation between functional domain count and

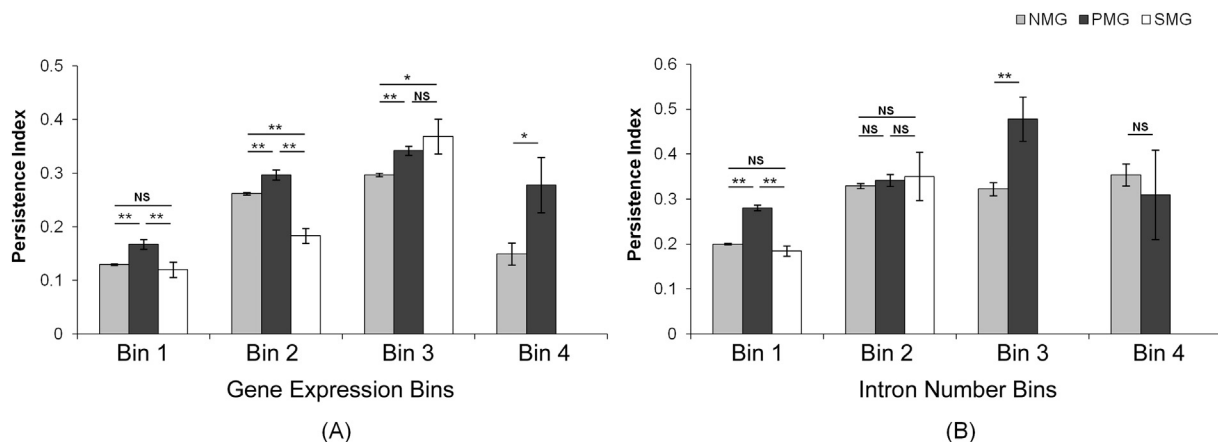


Fig. 5. Bar diagram showing the difference of Persistence index between PMGs and NMGs in each bin of (A) gene expression and (B) intron number. The expression values were divided into 4 bins: $2.0 < \text{Bin1} < 5.0$, $5.0 < \text{Bin2} < 8.0$, $8.0 < \text{Bin3} < 11.0$, $\text{Bin4} > 11$. The intron numbers were divided into 4 bins such that: $1 \leq \text{Bin1} \leq 10$, $11 \leq \text{Bin2} \leq 20$, $21 \leq \text{Bin3} \leq 30$, $\text{Bin4} \geq 30$. ** denotes $P < 0.01$, * denotes $P < 0.05$ and NS denotes Not significantly different values.

Table 1

Details of Mann Whitney *U* test of transcript and protein diversity between PMGs, NMGs and SMGs of *A. thaliana*.

	PMGs	NMGs	SMGs	P-values
Transcript diversity				
Mean	1.44	1.28	1.25	$P_{\text{PMG-NMG}} = 10^{-6}$
Standard deviation	0.78	0.64	0.57	$P_{\text{PMG-SMG}} = 10^{-6}$ $P_{\text{NMG-SMG}} = 0.53$
Protein diversity				
Mean	1.43	1.30	1.25	$P_{\text{PMG-NMG}} = 10^{-6}$
Standard deviation	0.79	0.69	0.57	$P_{\text{PMG-SMG}} = 10^{-6}$ $P_{\text{NMG-SMG}} = 0.46$

intron number in PMGs (Spearman's $\rho_{\text{domain no-intron no}} = 0.176$, $P < 10^{-6}$). Our results thus, indicate that introns increase protein function by acquisition of functional domains, and thus plays important role in protein multifunctionality.

3.6. Introns may serve as buffer for mutations in coding regions

Another probable explanation behind acquisition of high number of introns could be the fact that introns being themselves non-coding might retain mutational disturbances and thus buffers the coding exons from mutations, as explained by Jo and Choi [18]. We, thus, analysed the single nucleotide polymorphisms (SNPs) as the mutational force. A strong negative correlation between intron number and exonic SNP density (Spearman's $\rho = -0.312$, $P < 10^{-6}$) accompanied by a significantly higher enrichment of SNPs in the intronic regions as compared to exonic ones suggest that along with alternative splicing, intron enrichment is helpful for persistence of old genes to protect themselves from mutations. On the other way round, SNPs in intronic region could also guide splicing as observed in many different previous studies [50]. In this study, we have also found that intronic SNP density is significantly correlated with transcript count (Spearman's $\rho = 0.155$, $P < 10^{-6}$) in *A. thaliana*. These results show that SNPs do have a role in alternative splicing mechanisms. Moreover, exonic SNP density was found to have a slight yet significant negative correlation with transcript count (Spearman's $\rho = -0.074$, $P < 10^{-6}$). Thus, proteins with fewer number of splice variants have a slightly more chance of gathering SNPs in the exonic regions. Thus, we have searched the intronic regions for the presence of SNPs and tried to understand their role in conservation.

Previous studies [51] suggested that mutation through single nucleotide polymorphisms (SNPs) are more in the intronic regions of the genes as compared to the exonic counterparts. It has also been suggested that introns could possibly buffer mutations and protect the exons [18]. In this study, we hypothesized that introns may absorb more mutational shocks which allow the genes to retain normal protein function and hence be conserved. To elaborate this, we studied the distribution pattern of SNPs. We have observed a significantly higher count of SNPs in the introns than exons in all the groups (Mann-Whitney *U* test, $P < 0.01$). Here, we observed introns of PMGs have highest SNP density followed by NMGs and least in SMGs and NMGs have significantly more exonic SNPs than PMGs and SMGs (Fig. 6 A and B). However, exonic SNP density of PMGs and SMGs did not vary significantly. In addition to it, as shown above, we obtained very strong negative correlation between total exonic SNPs density and intron number, indicating introns could possibly absorb the mutational load of the genes, which is also supported by the previous notion of Jo and Choi [18]. Finally, intronic SNP density showed a slight yet significant correlation with PI (Spearman's $\rho = 0.079$, $P = 2.35 \times 10^{-9}$). However, there was no correlation of total exonic SNP density with PI (Spearman's $\rho_{\text{PI}} = 0.011$, $P = 0.362$). This shows that intronic SNPs indeed have a role in evolutionary conservation of genes. To further authenticate the study, we generated the SIFT score of the SNPs of the coding

exons of PMGs, NMGs and SMGs using the webserver (SIFT 4.0) [52]. Density of deleterious mutations (no. of deleterious mutations/cds length) was highest in SMG (0.01), followed by NMG (0.008) and PMG (0.006) (Sig at $P < 0.01$, Mann-Whitney *U* test). Moreover, intron number is significantly negatively correlated with this density of deleterious mutations (Spearman's rho of -0.036 , $P < 0.001$). We have also showed that intron number is highest in PMG, followed by NMG and SMG. Surprisingly, this is also true for density of tolerated mutations (no. of tolerated mutations/cds length). It was highest in SMG (0.05), followed by NMG (0.048) and PMG (0.041) (Sig at $P < 0.01$, Mann-Whitney *U* test). Moreover, intron number is significantly negatively correlated with this density of deleterious mutations (Spearman's rho of -0.167 , $P < 0.001$). Thus, it is clear that more number of introns somehow preventing the gene from accumulating more mutations (be it deleterious or tolerated) in the coding regions. However, it may be the fact that as introns rich genes are highly expressed, mutation accumulation is less [26]. We also conducted the cause and effect estimation of intron number and deleterious/tolerated mutations to understand the influence of the factors based on van der Lee et al. [53]. The result in both cases reveals the number of introns to be the cause of mutations be it deleterious or tolerated (Table 2) The presence of introns within the coding regions brings down the overall mutation of the exons, leading to the functional conservation of vital primary metabolic genes.

4. Conclusions

Primitiveness and conservation of metabolic genes is largely correlated with intron number and is expression independent. Unlike that of animal genomes, where housekeeping genes possesses shorter introns and have compact genetic architecture; Primary metabolic genes of *A. thaliana* (which has a basic housekeeping functionality) represent quite a different and unique set of characters. As a matter of fact PMGs share a combination of features that partially resembles both housekeeping and tissue specific genes. PMGs are pan-taxonomically conserved like that of housekeeping genes, but unlike animal housekeeping genes entails higher intron enrichment. Plants being autotrophic can harness their own energy. Hence, energy cost for processing large number of introns might not be a limitation amongst PMGs. At the same time primary genes, in course of evolution could give birth to secondary metabolic genes, which are again tissue specific. Thus, primary genes represent a complex trade-off between housekeeping and tissue specific genetic architectures in *A. thaliana*.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ygeno.2017.12.003>.

Abbreviations

PMG	primary metabolic gene
SMG	secondary metabolic gene
NMG	non metabolic gene
PI	persistence index
AI	age index
SNP	single nucleotide polymorphism

Declarations

Availability of data and materials

All data were obtained from publicly available databases (mentioned in the Materials and methods section) and are freely available online. The dataset used in the study can be found in the Additional file 1 (Microsoft Excel Worksheet). Furthermore, the datasets used and/or analysed during the current study are also available from the corresponding author on reasonable request.

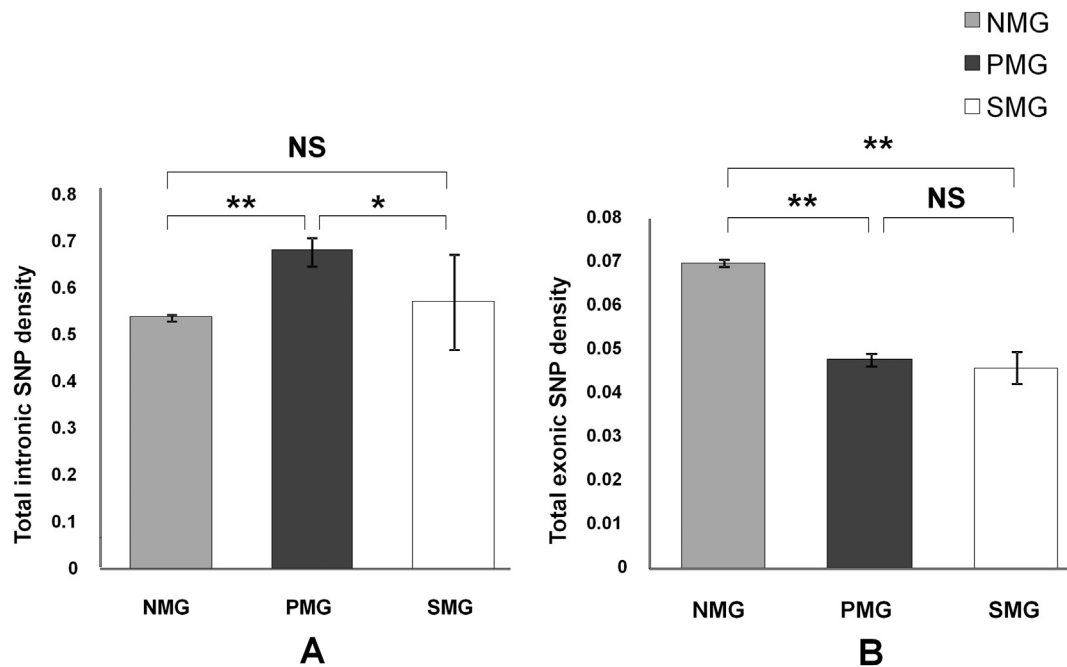


Fig. 6. Bar diagram showing the average values of (A) total intronic SNP density and (B) total exonic SNP density amongst different groups, PMGs, NMGs and SMGs. ** denotes $P < 0.01$, * denotes $P < 0.05$ and NS denotes Not significantly different values.

Table 2

Conditional probability study between intron number and Tolerated/Deleterious mutations.

		Tolerant mutation		
		High (T_H)	Low (T_L)	
Intron number	High (I_H)	4077	7432	11509
	Low (I_L)	9138	5788	14926
		13,215	13220	
		Deleterious mutation		
		High (D_H)	Low (D_L)	
Intron number	High (I_H)	5983	5559	11542
	Low (I_L)	7337	8057	15394
		13320	13616	

Event(E)	Condition (C)	Probability (Event Condition) = $P(E C)/P(C)$
A Deleterious mutation low	High intron number	$P(D_L I_H) = \frac{5559}{11542} = 0.482$
	Low deleterious mutation	$P(I_H D_L) = \frac{5559}{13616} = 0.408$
B Tolerant mutation low	High intron number	$P(T_L I_H) = \frac{7432}{11542} = 0.644$
	Low tolerant mutation	$P(I_H T_L) = \frac{7432}{13220} = 0.562$

Competing interests

The authors declare that no financial and/or non-financial competing interests exist.

Funding

No funding information is applicable for this manuscript.

Authors' contributions

Conceived and designed the experiments: DM, DS, DA, AM, TCG.

Performed the experiments: DM, DS, SC. Analysed the data: DM, DS, AM. Wrote the paper: DM, DS, DA, AM.

Acknowledgements

The authors are thankful to Department of Biotechnology, Government of India, for funding the Bioinformatic Centre at Bose Institute Kolkata, where the work was done.

Conflict of interest: The authors declared that they had no conflict of interests.

References

- [1] A.I. Lamond, RNA splicing — running rings around RNA, *Nature* 397 (6721) (1999) 655–656.
- [2] F. Rodriguez-Trelles, R. Tarrío, F.J. Ayala, Origins and evolution of spliceosomal introns, *Annu. Rev. Genet.* 40 (2006) 47–76.
- [3] S.W. Roy, W. Gilbert, Rates of intron loss and gain: implications for early eukaryotic evolution, *Proc. Natl. Acad. Sci. U. S. A.* 102 (16) (2005) 5773–5778.
- [4] D. Jeffares, Rapidly regulated genes are intron poor (vol 24, pg 375, 2008), *Trends Genet.* 24 (10) (2008) 488–488.
- [5] Y.-F. Yang, T. Zhu, D.-K. Niu, Association of intron loss with high mutation rate in *Arabidopsis*: implications for genome size evolution, *Genome Biol. Evol.* 5 (4) (2013) 723–733.
- [6] G. Parra, K. Bradnam, A.B. Rose, I. Korf, Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants, *Nucleic Acids Res.* 39 (13) (2011) 5328–5337.
- [7] M. Chorev, L. Carmel, The function of introns, *Front. Genet.* 3 (2012) 55.
- [8] J.F. Wendel, R.C. Cronn, I. Alvarez, B. Liu, R.L. Small, D.S. Sanchina, Intron size and genome size in plants, *Mol. Biol. Evol.* 19 (12) (2002) 2346–2352.
- [9] K. Jiang, L.R. Goertzen, Spliceosomal intron size expansion in domesticated grapevine (*Vitis vinifera*), *BMC. Res. Notes* 4 (2011) (52–52).
- [10] D.S. Ucker, K.R. Yamamoto, Early events in the stimulation of mammary tumor virus RNA synthesis by glucocorticoids. Novel assays of transcription rates, *J. Biol. Chem.* 259 (12) (1984) 7416–7420.
- [11] M.G. Izban, D.S. Luse, Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates, *J. Biol. Chem.* 267 (19) (1992) 13647–13655.
- [12] E. Eisenberg, E.Y. Levanon, Human housekeeping genes are compact, *Trends Genet.* 19 (7) (2003) 362–365.
- [13] H. Yang, In plants, expression breadth and expression level distinctly and non-linearly correlate with gene structure, *Biol. Direct* 4 (2009) 45 (discussion 45).
- [14] C.I. Castillo-Davis, S.L. Mekhedov, D.L. Hartl, E.V. Koonin, F.A. Kondrashov, Selection for short introns in highly expressed genes, *Nat. Genet.* 31 (4) (2002) 415–418.
- [15] M. Lynch, The origins of eukaryotic gene structure, *Mol. Biol. Evol.* 23 (2) (2006)

- 450–468.
- [16] M. Lynch, The Origins of Genome Architecture, Vol. 98 Sinauer Associates Sunderland, 2007.
 - [17] M. Lynch, B. Koskella, S. Schaack, Mutation pressure and the evolution of organelle genomic architecture, *Science* 311 (5768) (2006) 1727–1730.
 - [18] B.S. Jo, S.S. Choi, Introns: the functional benefits of introns in genomes, *Genomics Inform.* 13 (4) (2015) 112–118.
 - [19] A. Kalsotra, T.A. Cooper, Functional consequences of developmentally regulated alternative splicing, *Nat. Rev. Genet.* 12 (10) (2011) 715–729.
 - [20] O. Gorlova, A. Fedorov, C. Logotheitis, C. Amos, I. Gorlov, Genes with a large intronic burden show greater evolutionary conservation on the protein level, *BMC Evol. Biol.* 14 (1) (2014) 50.
 - [21] T. Maniatis, R. Reed, An extensive network of coupling among gene expression machines, *Nature* 416 (6880) (2002) 499–506.
 - [22] S.Y. Ying, S.L. Lin, Intronic microRNAs, *Biochem. Biophys. Res. Commun.* 326 (3) (2005) 515–520.
 - [23] D. Mascarenhas, I.J. Mettler, D.A. Pierce, H.W. Lowe, Intron-mediated enhancement of heterologous gene expression in maize, *Plant Mol. Biol.* 15 (6) (1990) 913–920.
 - [24] A.E. Vinogradov, Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* 20 (5) (2004) 248–253.
 - [25] D. Mukherjee, A. Mukherjee, T.C. Ghosh, Evolutionary rate heterogeneity of primary and secondary metabolic pathway genes in *Arabidopsis thaliana*, *Genome Biol. Evol.* 8 (1) (2016) 17–28.
 - [26] X.-Y. Ren, O. Vorst, M.W.E.J. Fiers, W.J. Stiekema, J.-P. Nap, In plants, highly expressed genes are the least compact, *Trends Genet.* 22 (10) (2006) 528–532.
 - [27] Y.S. Rao, Z.F. Wang, X.W. Chai, Wu GZ, M. Zhou, Q.H. Nie, X.Q. Zhang, Selection for the compactness of highly expressed genes in *Gallus gallus*, *Biol. Direct* 5 (2010).
 - [28] E. Eisenberg, E.Y. Levanon, Human housekeeping genes are compact, *Trends Genet.* 19 (7) (2003) 362–365.
 - [29] C. Seoighe, C. Gehring, L.D. Hurst, Gametophytic selection in *Arabidopsis thaliana* supports the selective model of intron length reduction, *PLoS Genet.* 1 (2) (2005) e13.
 - [30] E. Pichersky, D.R. Gang, Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective, *Trends Plant Sci.* 5 (10) (2000) 439–445.
 - [31] M. Schena, The evolutionary conservation of eukaryotic gene-transcription, *Experientia* 45 (10) (1989) 972–983.
 - [32] J.Y. Yuan, Evolutionary conservation of a genetic pathway of programmed cell death, *J. Cell. Biochem.* 60 (1) (1996) 4–11.
 - [33] R.J. Kinsella, A. Kähäri, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, et al., Ensembl BioMarts: a hub for data retrieval across taxonomic space, *Database: The Journal of Biological Databases and Curation* 2011 (2011) bar030.
 - [34] P.J. Kersey, J.E. Allen, M. Christensen, P. Davis, L.J. Falin, C. Grabmueller, D.S.T. Hughes, J. Humphrey, A. Kerhornou, J. Khobova, et al., Ensembl Genomes 2013: scaling up access to genome-wide data, *Nucleic Acids Res.* 42 (2014) D546–D552.
 - [35] M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.* 28 (1) (2000) 27–30.
 - [36] Y.-L. Guo, Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes, *Plant J.* 73 (6) (2013) 941–951.
 - [37] S. Dash, J. Van Hemert, L. Hong, R.P. Wise, J.A. Dickerson, PLEXdb: gene expression resources for plants and plant pathogens, *Nucleic Acids Res.* 40 (D1) (2012) D1194–D1201.
 - [38] The Gene Ontology C, M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, et al., Gene ontology: tool for the unification of biology, *Nat. Genet.* 25 (1) (2000) 25–29.
 - [39] M. Punta, P.C. Coghill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, et al., The Pfam protein families database, *Nucleic Acids Res.* 40 (Database issue) (2012) D290–D301.
 - [40] H.B. Mann, D.R. Whitney, On a test of whether one of 2 random variables is stochastically larger than the other, *Ann. Math. Stat.* 18 (1) (1947) 50–60.
 - [41] C. Spearman, The proof and measurement of association between two things, *Am. J. Psychol.* 15 (1904) 72–101.
 - [42] L. Carmel, I.B. Rogozin, Y.I. Wolf, E.V. Koonin, Evolutionarily conserved genes preferentially accumulate introns, *Genome Res.* 17 (7) (2007) 1045–1050.
 - [43] S. Subramanian, S. Kumar, Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome, *Genetics* 168 (1) (2004) 373–381.
 - [44] M. Warnefors, A. Eyre-Walker, The accumulation of gene regulation through time, *Genome Biol. Evol.* 3 (2011) 667–673.
 - [45] Y. Imamichi, T. Mizutani, Y. Ju, T. Matsumura, S. Kawabe, M. Kanno, T. Yazawa, K. Miyamoto, Transcriptional regulation of human ferredoxin reductase through an intronic enhancer in steroidogenic cells, *Biochim. Biophys. Acta, Gene Regul. Mech.* 1839 (1) (2014) 33–42.
 - [46] A.B. Rose, J.A. Beliakoff, Intron-mediated enhancement of gene expression independent of unique intron sequences and splicing, *Plant Physiol.* 122 (2) (2000) 535–542.
 - [47] J. Morata, S. Bejar, D. Talavera, C. Riera, S. Lois, G. Mas de Xaxars, X. de la Cruz, The relationship between gene isoform multiplicity, number of exons and protein divergence, *PLoS One* 8 (8) (2013).
 - [48] J. Roux, M. Robinson-Rechavi, Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication, *Genome Res.* 21 (3) (2011) 357–363.
 - [49] B.R. Graveley, Alternative splicing: increasing diversity in the proteomic world, *Trends Genet.* 17 (2) (2001) 100–107.
 - [50] R.A. Moyer, D. Wang, A.C. Papp, R.M. Smith, L. Duque, D.C. Mash, W. Sadee, Intronic polymorphisms affecting alternative splicing of human dopamine D2 receptor are associated with cocaine abuse, *Neuropsychopharmacology* 36 (4) (2011) 753–762.
 - [51] J. Evans, J. Kim, K.L. Childs, B. Vaillancourt, E. Crisovan, A. Nandety, D.J. Gerhardt, T.A. Richmond, J.A. Jeddelloh, S.M. Kaeppler, et al., Nucleotide polymorphism and copy number variant detection using exome capture and next-generation sequencing in the polyploid grass *Panicum virgatum*, *Plant J.* 79 (6) (2014) 993–1008.
 - [52] N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, P.C. Ng, SIFT web server: predicting effects of amino acid substitutions on proteins, *Nucleic Acids Res.* 40 (W1) (2012) W452–W457.
 - [53] R. van der Lee, B. Lang, K. Kruse, J. Gspöner, N.S. de Groot, M.A. Huynen, A. Matouschek, M. Fuxreiter, M.M. Babu, Intrinsically disordered segments affect protein half-life in the cell and during evolution, *Cell Rep.* 8 (6) (2014) 1832–1844.



Prof. Ghosh with his lab members, 2014

Credits

Bioinformatics Centre, Bose Institute

Prof. Tapash Chandra Ghosh

Dr. Soumita Podder

Dr. Arup Panda

Dr. Sandip Chakraborty

Ms. Dola Samanta