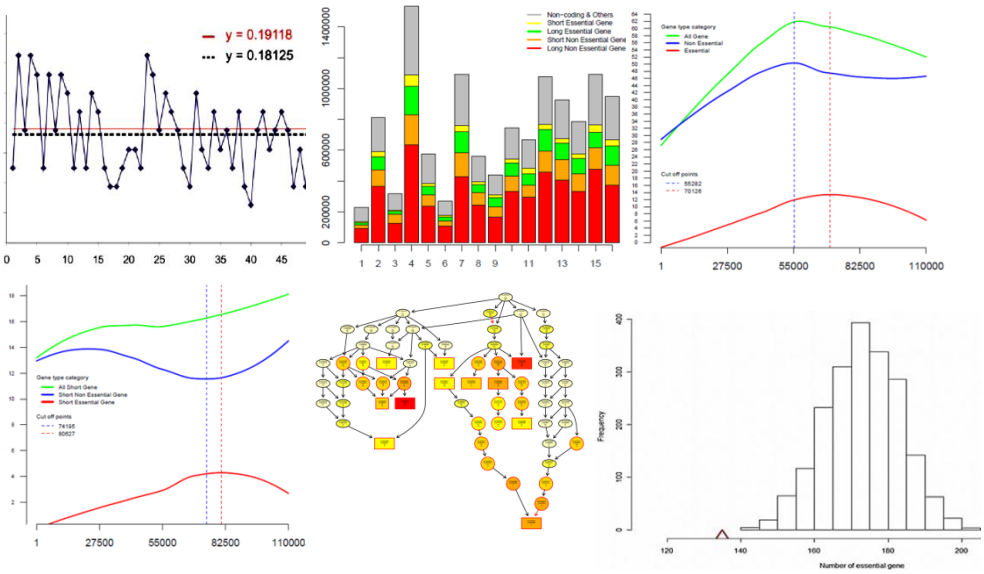
CHALMERS | UNIVERSITY OF GOTHENBURG



Bioinformatics and Statistical Methods for Identifying Enrichment of Functional Gene Classes in Telomeric Regions of Chromosomes

Mohammad Tanvir Ahamed

Masters Program in Bioinformatics and Systems Biology
Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Göteborg, Sweden 2013

Bioinformatics and Statistical Methods for Identifying Enrichment of Functional Gene Classes in Telomeric Regions of Chromosomes

Masters of Science thesis in the Master's Program
Bioinformatics and Systems Biology

Mohammad Tanvir Ahamed

Department of Mathematical Sciences
Division of Bioinformatics and Systems Biology
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Göteborg, Sweden 2013

Bioinformatics and Statistical Methods for Identifying Enrichment of Functional Gene Classes in Telomeric Regions of Chromosomes

MOHAMMAD TANVIR AHAMED

Supervised by : Professor Olle Nerman

Department of Mathematical Sciences,
Division of Bioinformatics and Systems Biology
Chalmers University of Technology and University of Gothenburg,
SE-412 96 Göteborg, Sweden
Phone: +46 (0) 31-772-1000

Cover:
GO term specific analysis and telomeric cutoff point for essential and non-essential genes.

Abstract

It has been noted that the telomeric regions of *Saccharomyces cerevisiae* has fewer essential genes than expected from random shuffling. Further the general effect of single gene silencing of non-essential genes in the telomeric regions with an average has less effect on viability than for non-essential genes in other chromosomal regions. It has also been suggested that the genes in the telomeric regions are less stable with higher mutation and recombination rates. And this could be an evolutionary positive property for adaption of genes with changing environment, provided that there are back up systems for the genes.

In this work, we took a look at some different statistical properties of the telomeres and the genes in the telomeric regions. Some of the studied properties are: How dense the code is in the telomeric region compared to the rest of the genome? What length distribution do the genes have in the telomeric region in comparison to the general length distribution? What GO-annotated classes are over-represented in telomeres? Can we find protein sequence clusters that are over-represented in the telomeres?

We have found fairly a lot of interesting properties and at least partly our results also support the earlier suggestions. Finally, for the future, we suggest that comparison of our different finding corresponding telomeric statistical properties in *Saccharomyces cerevisiae* should be performed with other yeast species, like *Schizosaccharomyces pombe*, which is evolutionary distant enough to be genomically fairly reshuffled. As usual, in multivariate statistics, the statistical properties are correlated (Length correlates to viability, function, etc.) and causality is hard to deduce, but may be easier to understand using more organisms.

The main findings of the thesis were that, there is less code in the extreme telomeric region. In percentage, long essential genes in the telomeric region are very few. The numbers of genes in the long non-essential gene category are larger but also quite few compared to elsewhere. And of those that reside in the telomeric region, there are many genes related to metal ion transport, disaccharide and oligosaccharide metabolic and catabolic process. The pipeline of methods used in the present research also identifies some gene function related to helicase activity that has been pointed out in earlier research.

Keywords: *Saccharomyces Cerevisiae* Genome; Telomeric Gene; Essential Gene; Non-essential Gene; Gene Ontology; Hyper-geometric Distribution; Term Redundancy;

Acknowledgement

To acknowledge my primary debt of gratitude of course goes to my thesis supervisor Professor Olle Nerman, Department of Mathematical Science, Chalmers University of Technology, Sweden for his clear, careful guidance and direction. He carefully read the entire report, gently point out the errors and enthusiastically offering many useful suggestions for improvement.

I wish to acknowledge my debt of gratitude to Professor Bing Zhang, Department of Biomedical Informatics, Vanderbilt University School of Medicine, USA. He patiently replied all of my e-mails and supported my research work by providing necessary information and data related to GO terms, which were used in "WebGestalt" "WEB-based GEne SeT AnaLysis Toolkit".

I am very grateful to Professor Anders Blomberg, Functional Genomics, Lundberg Laboratory, Gothenburg University, Sweden for his careful guidance, direction and suggestion to improve my thesis work in a constructive and informative way. I am also very appreciative to Dr. Jonas Warringer, Assistant Professor, Dept. of Chemistry and Molecular Biology, Gothenburg University for his time and useful suggestion during a meeting and post thesis presentation with Professor Olle Nerman and Professor Anders Blomberg.

I am very much thankful to Rhoda Kinsella (Ph.D.), Ensembl Production Project Leader, European Bioinformatics Institute (EMBL-EBI) for his rapid support to solve the problem regarding BioMart web server. I am grateful to Guangchuang Yu, Bioinformatics technician at Jinan University, China for his wonderful support regarding "GOSemSim" package in Bioconductor. In addition, I express my thanks to all users from "r-help@r-project.org" and "users@biomart.org", without whose help the long R code behind the thesis work would not have been possible.

I am thankful to Mariana Buongermino Pereira, PhD student in Bioinformatics at the Department of Mathematical Sciences, Chalmers University of Technology. As a thesis presentation opponent, she reads my thesis elaborately and supported with many corrections and improvement

I want to express my thanks to all of my fellows, especially to Adnan Hashim (Phd Student in Bioinformatics at University of Salerno, Italy) and Jing Guo (Informatician at AstraZeneca R&D, Sweden) without whose help this research project would not have been possible. I am also thankful to Mohammad Mazharul Hoque (MS student in Nano Scale Science and Technology, Chalmers), Imtiaz Hoque, (Project and System Engineer at Solvina AB Gothenburg, Sweden) for their support and love.

I am very grateful to my family, especially to my Father, Mother and Suria Jahan for their enormous patience and love during my thesis work and study period at Chalmers, Sweden.

I want to give special thanks to Md Asaduzzaman, Associate Professor ISRT, University of Dhaka who inspire me to study Bioinformatics with statistics background and Chalmers University of Technology, Sweden to give me an opportunity to fulfill my wish.

Finally, I want to thank All Mighty to give me the strength to complete this thesis work and give me a chance to thank all the people mentioned above, who helped and support me during the project and study.

Mohammad Tanvir Ahamed
Bioinformatics and Systems Biology
Chalmers University of Technology
Gothenburg, Sweden

Dedication

To My Parents

Table of Contents

Chapter Four: Results

List of figures

List of tables

Abbreviations

GO = Gene Onlology

SNP = Single Nucleotide Polymorphism

DEG = Database of Essential Gene

SGD = Saccharomyces Genome Database

BP = Biological Process

MF = Molecular Function

CC = Cellular Function

Chapter One: Introduction

1.1. Background of the study

The biology of telomere has a very wide application in the field of human health and aging. The discovery of telomerase and research on telomere capping by Elizabeth Blackburn, Carol Greider, and Jack Szostak, were honored with the 2009 Nobel Prize in Medicine. All the prize winners conducted research in single-cell organisms, including budding yeast. Therefore, *Saccharomyces cerevisiae* is a premier organism for telomere research.

Eukaryote chromosomes have linear DNA molecules has a physical end which is known as telomere. It is estimated that about 10,000 DNA damaging event (Mutation, SNP, etc.) occurs every day in a cell of human body (1). Among the entire DNA damaging event, probably Double-Stranded DNA Breaks (DSBs) is a more complex event which creates chromosome ends at internal sites on chromosome. So it is a complex task to understand the mechanism how cell makes the difference between the natural end and telomere from DSBs. Telomeres keep the stable maintenance of the chromosome. So cell must keep telomeres safe from degradation, unexpected mutation or fused with other ends. However, in DSBs, DNA must be repaired by either homologous or non-homologous recombination, and this repair often involves in regulating degradation of DSB. In fact, unrepaired DSBs results in the cell cycle stop to provide more repairing time. Capping is used to describe how telomere prevents their degradation and recombination fusion (2) (3). It is suspected that, as a consequence of capping, there are fewer genes in the telomeric region. In many organisms, telomeric genes are subjected to a specific type of function. Another key role of telomere is to provide substrate for a specific mechanism of replication. Telomere replication is carried out by ribonucleic protein named telomerase, which is mechanistically connected to reverse transcriptase (4).

Figure 1: DNA arrangement at telomeres indicating the sub-telomeric X and Y' elements as well as the terminal repeated

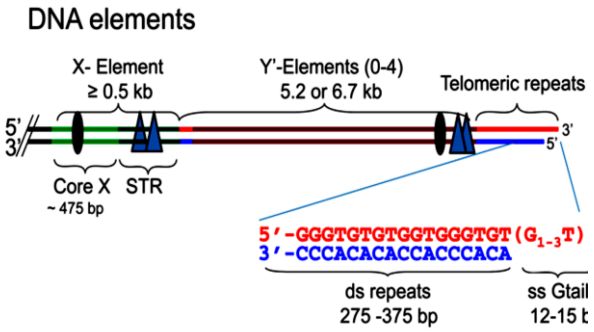


Figure 1 describe DNA arrangement at telomeres indicating the sub-telomeric X and Y' elements as well as the terminal repeated sequence (33)

In most organisms, telomeres are maintained by telomerase. Telomeres in *Saccharomyces cerevisiae* consists of non-protein coding repeated DNA. There are 300 ± 75 bp of simple repeats, typically abbreviated C1-3A/TG1-3.

Saccharomyces cerevisiae telomeric DNA is unusual, although not unique, in being heterogeneous. This sequence heterogeneity is due to a combination of effects: in a given extension cycle, only a portion of the RNA template is used, and the RNA template and telomeric DNA align in distinct registers in different extension cycles (5).

Similar to most eukaryote organisms, yeast telomeric regions contain sub-telomeric, middle, repetitive elements which are often called TAS elements (Telomere Associated Sequences). *Saccharomyces Cerevisiae* has two classes of TAS element, which are X and Y9. Y9 is found in zero to four tandem copies immediately internal to the telomeric repeats (6). Y9 comes in two sizes, Y9 long (6.7 kb) and Y9 short (5.2 kb) (6), which differ from each other by multiple small insertions / deletions (7). X is present at virtually all telomeres and is much more heterogeneous in sequence and size. Although X is found on all telomeres, it is composed of a series of repeats, many of which are present on only a subset of telomeres. When telomeres contain both X and Y9, X is centromere proximal to Y9 (8).

Sub-telomeric regions are dynamic, that undergo frequent recombination. Moreover, sub-telomeric repeats diverge rapidly even among related yeast strains. X and Y9 both contain potential replication origins or ARS elements (Autonomously Replicating Sequences) whose presence probably contributes to the dynamic nature of sub-telomeric regions (6) (7) (8). X and Y9 have binding sites for multiple transcription factors, whose identity differs from telomere to telomere (6). Because the sequence of sub-telomeric regions and the proteins that bind with them are variable, their presence can confer distinct behaviors on individual telomeres (8).

Gene's physical distribution in the telomeric region is not a true randomly distributed phenomenon. There are fewer genes, especially essential genes in the telomeric region than non-essential genes. In *Saccharomyces cerevisiae* genome in each chromosome, from both telomeric end point up to 25th gene position it was observed an increasing tendency in the number of essential genes (9). These telomeric genes include both essential and non-essential genes. Essential genes, those are indispensable to support cellular life. These genes constitute a minimal gene set required for a living cell. Therefore, the functions encoded by this gene set are essential and could be considered as a foundation of life itself (10) (11). In *Saccharomyces cerevisiae,* there are 1110 essential genes were observed by single gene knockout experiment (12).

A gene location in the chromosome has an important role to determine how the physical characteristics of an organism will change and develop. Physical characteristics like height, eye color, and hair colors are the natural property of an organism which differs among individuals and one population to another. Very few processes of change or development of these physical characteristics can explain by mutation, natural selection and chance. A recent study on 16000 traits of *Caenorhabditis elegans* chromosomes showed that genes located in the centromeric region of a chromosome were less likely to contribute to genetic variation of traits than were genes found at the telomeric region. In other words, a gene's location on a chromosome influenced the range of physical differences among different traits. In *Caenorhabditis elegans,* genes in the centers of chromosomes are tied to more neighbors than are genes near the ends of the chromosomes. As a result, the genes in the center are less able to harbor genetic variation (13).

1.2. Aim of the study

Our aim is to find, any exceptional group of genes that situated in the telomeric region which performs some unique function. To get this, we have set initially two filters, which are "Essential Property" and "Length Property" of a gene. Many more filters like genes physical position, CG content ratio, etc. can be used, but initially only these 2 filters was used for this research

1. **Essential property of a gene:**
   A gene considered as essential if the knockout of that gene cause for growth stop or death of cell under a specific laboratory condition. Knockout of two genes could make the same result. But only one gene that makes that type of results, can be defined as a essential gene.

2. **Length property of a gene:**
   Length could be an important property of a gene. The gene shorter than the median of all genes can be defined as shorter gene and longer gene can be defined as accordingly.

1.3. Overview of the thesis organization

The main concern of this thesis report is to understand the common pattern of genes organization in the telomeric region to explain the evolutionary mechanism. This present research study is based on telomeric genes from *Saccharomyces cerevisiae* genome. Though this study is only based on *Saccharomyces cerevisiae* genome, but this approach is general and suitable for similar analysis in other well annotated species, e.g. *Saccharomyces Pombo* in identifying the degree of effects of the overall process on evolutionary mechanism.

The present study can be outlined in two major sections, which are (1) Base-pair specific analysis and (2) GO term specific analysis. The research report consists of six sections. The first section contains the brief literature review, background and aim of the current study based on *Saccharomyces cerevisiae* Genome. Second section consists of source and nature of the data along with the visual representation of the combine data set, which is created from different data sources for current research analysis. Different statistical and bioinformatics methods and conceptual definition related to the aim of present study are placed in section three. In section four, detail analysis has done on telomeric genes on *Saccharomyces cerevisiae*. Discussion and future work scope has discussed in section five and section six contains the conclusion about the current research topic.

Chapter Two: Sources & Nature of Data

1.1. Sources of data

For the present research, we have used data from three different data sources, which are Gene Ontology (GO) database, Database of Essential Genes (DEG) and Sacchromyces Genome Database (SGD).

1.1.1. Gene Ontology (GO) database

A set of annotation maps describing the entire Gene Ontology assembled using data from the Gene Ontology web database[1]. The whole annotation data was obtained in an R[2] package named "GO.db" under Bioconductor package.

1.1.2. Database of Essential Genes (DEG)

The list of essential gene of yeast (*Saccharomyces Cerevisiae*) was taken from the DEG[3] (Database of Essential Genes) (12). Total Number of essential gene was 1110. The essential feature of a gene was determined by the effect of gene deletion (12). So without knowing of actual function of gene it is possible to determine that the gene is essential or not. Information based on SGD, it has determined that, among 1110 essential gene, 13 genes are uncharacterized. And the rest 1093 genes are characterized.

1.1.3. Saccharomyces Genome Database (SGD)

The total list of yeast (*Saccharomyces Cerevisiae*) gene list has taken from Saccharomyces Genome Database (SGD) (14). A total of 7156[4] chromosomal gene with verified and unverified functional profiling was considered for the analysis.

1.1.4. Organization of GO, DEG and SGD data

The list of all chromosomal genes including essential, non-essential, characterized, uncharacterized are summarized in the table 1. In table 1, the 1st column indicated by "PI" is the primary SGD id, 2nd column indicated by "SI" is the systematic name of the gene, 3rd column indicated by "L" is the length of the gene, 4th and 5th column indicated by "START" and "STOP" are the start and stop position of a gene respectively, 6th column indicated by "ORF" is the strand position of a gene where 1 indicate the "Watson" strand (5'-3') and 0 indicate the "Crick" strand (3'-5'). In 7th column the property characterized and uncharacterized are indexed by the Q (Qualifier) with value 1 and 0 where 1 indicate that gene function is verified and 0 indicate that the gene is dubious, uncharacterized or RNA genes. The 8th column indicates the chromosome name for the corresponding gene, 9th column indicates the essential property of a gene which is indexed by 0 and 1 where 1 means the gene is essential and 0 means the gene is non-essential. The next 10th - 12th column indicate the GO (Gene Ontology) annotated terms in the category of molecular function (MF), biological process (BP) and cellular component (CC) respectively.

[1] http://www.geneontology.org/
[2] R is a free software environment for statistical computing and graphics
[3] http://tubic.tju.edu.cn/deg/
[4] http://www.yeastgenome.org/

Table 1: List of all chromosomal genes organization used in the present research

Col 1	*Col 2*	*Col 3*	*Col 4*	*Col 5*	*Col 6*	*Col 7*	*Col 8*	*Col 9*	*Col 10*	*Col 11*	*Col 12*
PI	**SI**	**L**	**START**	**STOP**	**ORF**	**Q**	**CHR**	**E**	**MF**	**BP**	**CC**
S000002143	YAL069W	315	335	649	1	0	1	0	GOXX	GOXX	GOXX
S000028594	YAL068W-A	255	538	792	1	0	1	0	GOXX	GOXX	GOXX
S000002142	YAL068C	363	1807	2169	-1	1	1	0	GOXX	GOXX	GOXX
S000028593	YAL067W-A	228	2480	2707	1	0	1	0	GOXX	GOXX	GOXX

1.2. Nature and overview of data

1.2.1. Nature of data

There are 7156 chromosomal gene in sixteen (16) chromosome and one (1) mitochondrial chromosome, which are active. The data of all 7156 genes has downloaded from SGD database and list of 1110 essential genes has downloaded from DEG database on 11 January, 2013. In these 7156 chromosomal genes 786 genes are dubious where nine genes are from mitochondrial chromosome. According to SGD glossary, a dubious open reading frame (ORF) is one that is unlikely to encode an expressed protein. Dubious ORFs may meet some or all of the following criteria:

1. The ORF is not conserved in other *Saccharomyces* species;
2. There is no well-controlled, small-scale, published experimental evidence that a gene product is produced;
3. A phenotype caused by disruption of the ORF can be ascribed to mutation of an overlapping gene;
4. The ORF does not contain an intron. Many ORFs classified as "Dubious" are small and overlap a larger ORF of the class "Verified" or "Uncharacterized"; however, overlap with another ORF does not mandate that an ORF be classified as "Dubious.

Distribution of all chromosomal genes in *Saccharomyces Cerevisiae* gene including chromosomal genes and mitochondrial gene are summarized in table 2

Table 2: Distribution of all chromosomal genes in *Saccharomyces Cerevisiae* genome

<table>
<tr><th rowspan="2">Chr</th><th colspan="4">Essential</th><th colspan="4">Non- Essential</th><th rowspan="2">Total</th></tr>
<tr><th>Unchartered</th><th>Verified</th><th>No qua</th><th>Dubious</th><th>Unchartered</th><th>Verified</th><th>No qua</th><th>Dubious</th></tr>
<tr><td>1</td><td>0</td><td>12</td><td>0</td><td>0</td><td>17</td><td>63</td><td>10</td><td>25</td><td>127</td></tr>
<tr><td>2</td><td>0</td><td>72</td><td>0</td><td>0</td><td>56</td><td>274</td><td>25</td><td>54</td><td>481</td></tr>
<tr><td>3</td><td>0</td><td>17</td><td>0</td><td>0</td><td>27</td><td>115</td><td>19</td><td>24</td><td>202</td></tr>
<tr><td>4</td><td>0</td><td>167</td><td>0</td><td>0</td><td>81</td><td>490</td><td>53</td><td>98</td><td>889</td></tr>
<tr><td>5</td><td>1</td><td>46</td><td>0</td><td>0</td><td>40</td><td>188</td><td>33</td><td>49</td><td>357</td></tr>
<tr><td>6</td><td>0</td><td>27</td><td>0</td><td>0</td><td>26</td><td>72</td><td>17</td><td>16</td><td>158</td></tr>
<tr><td>7</td><td>0</td><td>107</td><td>0</td><td>0</td><td>70</td><td>343</td><td>55</td><td>63</td><td>638</td></tr>
<tr><td>8</td><td>0</td><td>42</td><td>0</td><td>0</td><td>46</td><td>190</td><td>20</td><td>43</td><td>341</td></tr>
<tr><td>9</td><td>0</td><td>38</td><td>0</td><td>0</td><td>38</td><td>134</td><td>24</td><td>31</td><td>265</td></tr>
<tr><td>10</td><td>3</td><td>65</td><td>0</td><td>0</td><td>46</td><td>239</td><td>36</td><td>45</td><td>434</td></tr>
<tr><td>11</td><td>0</td><td>68</td><td>0</td><td>0</td><td>33</td><td>212</td><td>23</td><td>35</td><td>371</td></tr>
<tr><td>12</td><td>1</td><td>93</td><td>0</td><td>0</td><td>87</td><td>320</td><td>65</td><td>77</td><td>643</td></tr>
<tr><td>13</td><td>2</td><td>80</td><td>0</td><td>0</td><td>65</td><td>306</td><td>45</td><td>52</td><td>550</td></tr>
<tr><td>14</td><td>2</td><td>73</td><td>0</td><td>0</td><td>63</td><td>256</td><td>24</td><td>41</td><td>459</td></tr>
<tr><td>15</td><td>1</td><td>96</td><td>0</td><td>0</td><td>66</td><td>366</td><td>42</td><td>69</td><td>640</td></tr>
<tr><td>16</td><td>0</td><td>97</td><td>0</td><td>0</td><td>61</td><td>298</td><td>35</td><td>55</td><td>546</td></tr>
<tr><td>Mito Chr</td><td>0</td><td>0</td><td>0</td><td>0</td><td>2</td><td>17</td><td>27</td><td>9</td><td>55</td></tr>
<tr><td>Total</td><td>10</td><td>1100</td><td></td><td></td><td>824</td><td>3883</td><td></td><td></td><td></td></tr>
<tr><td rowspan="3"></td><td colspan="4" rowspan="3">1110</td><td colspan="3">4707</td><td>553</td><td rowspan="3"></td></tr>
<tr><td colspan="4">5260</td><td>786</td></tr>
<tr><td colspan="4">6046</td><td>7156</td></tr>
</table>

In all 7156 chromosomal genes, 1110 genes are essential genes and rest 6046 genes are non-essential. In 6046 non-essential genes, 824 genes are uncharacterized including 2 genes from mitochondrial chromosome, 3883 genes are characterized including 17 from mitochondrial chromosome and 553 genes has non qualifies (RNA genes or genes not present systematic sequence of S288C or transposable element genes) including 27 from mitochondrial chromosome. In 786 dubious genes, 9 genes reside in the mitochondrial chromosome and rest 777 resides in 16 chromosomes. In rest of the 6324 chromosomal genes, genes only with feature "ORF" are considered for the current study. So the number of studied genes was 5798 in 16 chromosomes.

Table 3: Distribution of Genes with "No Qualifier" (RNA, Not in systematic sequence of S288C and Transposable element gene) in *Saccharomyces Cerevisiae* genome

	Mitochondrial chromosome	ncRNA	Not in systematic sequence of S288C	Pseudogene	rRNA	snoRNA	snRNA	Transposable element gene	tRNA	Total
Genes with "No Qualifier"	27	14	19	21	25	77	6	89	275	553

In 553 genes with "No qualifier" 27 are mitochondrial genes and rests 526 are chromosomal genes. In these 526 chromosomal genes, 337 are RNA genes, 19 genes are not in systematic sequence of S288C and 89 genes are transposable element gene. These five types of RNA genes are found in all 16 chromosomes which are ncRNA, rRNA, snoRNA, snRNA and tRNA. A non-coding RNA (ncRNA) is a functional part of RNA that is not translated into a protein. ncRNA is also known as non-protein coding RNA (npcRNA), non-messenger RNA (nmRNA), functional RNA (fRNA) and short bacterial ncRNA (sRNA). Ribosomal ribonucleic acid (rRNA) is the RNA component of the ribosome, and is essential for protein synthesis. Small nucleolar RNA (snoRNA) is a class of small RNA that works as a guide for chemical modification of rRNA, tRNA and snoRNA. Small nuclear ribonucleic acid (snRNA) is also known as U-RNA. snRNA has the function in processing of pre-mRNA (hnRNA) in the nucleus, in the regulation of transcription factors (7SK RNA) or RNA polymerase II (B2 RNA) and maintaining the telomeres. Transfer RNA (tRNA) is an adaptor molecule which function is to serves as the physical link between the nucleotide sequence of nucleic acids (DNA and RNA) and the amino acid sequence of proteins.

Figure 2: Graphical representation of distribution of Essential and Non-essential genes in *Saccharomyces Cerevisiae* genome

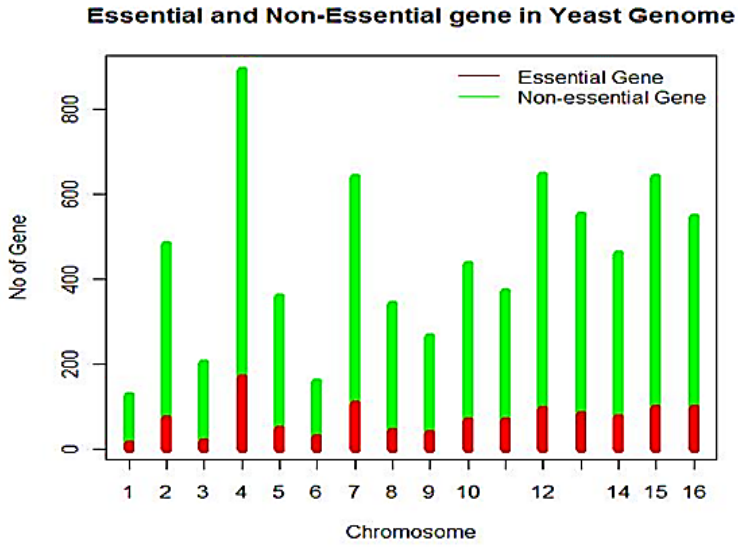


Figure 2 is the graphical representation of 5798 genes in 16 chromosome including 1110 essential genes and 4688 non-essential genes . In figure 1 , red and green color define essential and non-essential genes respectively

1.2.2. Summary of data

Total 5798 chromosomal gens can be categorized into the essential and non-essential group where 4688 genes from non-essential and 1110 genes from essential gene category. Again, there are 397 chromosomal genes, which are RNA genes. These three categories of genes (Non-essential, Essential and RNA genes) can be categorized based on the criteria shorter gene, longer gene and telomeric cutoff point.

The longer and shorter gene can be defined as the gene that is greater or smaller than the median of the respective group of genes and the cut off for the telomeric region is defined from location properties of respective category. (See section: 4.2)

Table 4: Summary table of all genes (Essential, non-essential and RNS genes) that is divided into 2 category, longer gene and shorter gene

	Non-essential Genes Threshold length : 1179 bp			**Essential Genes** Threshold length : 1359 bp				**RNA Genes** Threshold length : 82 bp			
	In cutoff (a^5/u^6) (a% u%)	Out cutoff (a/u) (a%/u%)	Total		In cutoff (a/u) (a%/u%)	Out cutoff (a/u) (a%/u%)	Total		In cutoff	Out cutoff	Total
Long gene Cutoff: 55282 bp	339 (292 / 47) (87%/13%)	2007 (1815/192) (91%/9%)	2346	**Long gene Cutoff: 70126 bp**	106 (106/0)	449 (447/2)	555	**Long gene Cutoff: 54025 bp**	41	148	204
Short gene Cutoff: 74195 bp	478 (315/163) (68%/32%)	1864 (1444/420) (77%/23%)	2342	**Short gene Cutoff: 80627 bp**	113 (111/2)	442 (436/6)	555	**Short gene Cutoff: no cut off**			193
Total			**4688**				**1110**				**397**

[5] a = Annotated / Verified
[6] u = Un-annotated / Unverified

In 4688 non-essential genes, 2346 genes and 2342 genes are longer and shorter gene respectively. For longer genes, 339 genes are in telomeric region and for shorter genes 478 genes are in telomeric region. In essential gene category, there are 555 genes are both in shorter and longer gene group where 106 genes and 113 genes are in telomeric region for longer and shorter gene group respectively. For 397 RNA genes, there are 204 genes are long and rest 193 genes are short. The overall gene distribution is described in figure 3.

Figure 3: Graphical representation of all genes from different category in 16 chromosomes

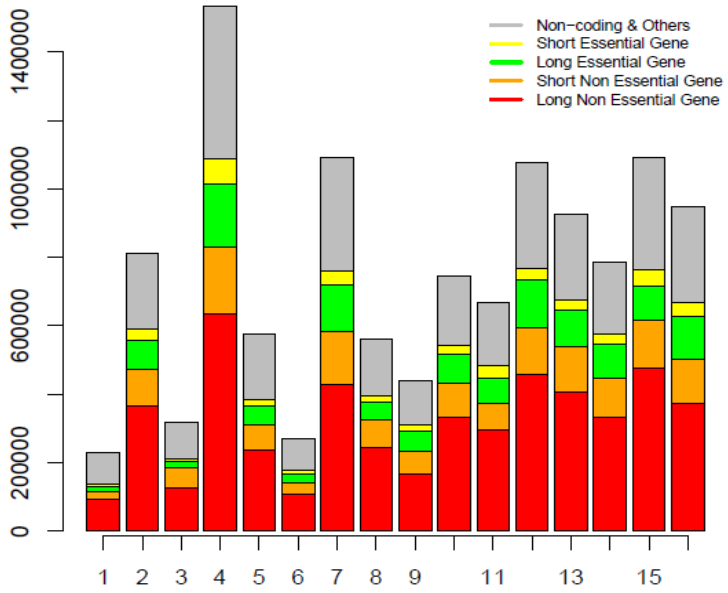
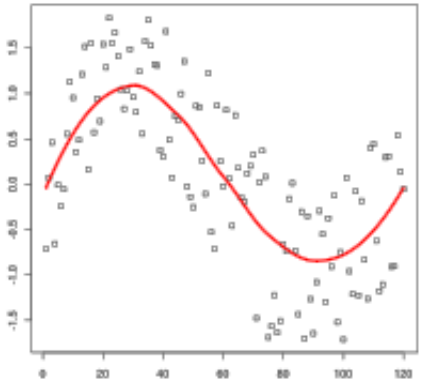


Figure 3 , is a graphical representation of chromosomal genes in different category . Red , orange , green , yellow and ash color indicate the long non-essential , short non-essential , long essential , short essential and non-coding region respectively

Chapter Three: Methods of Analysis

3.1. Local regression and LOESS method

Figure 4: Local Regression : LOESS method



Local Regression is a very popular method for smoothing of data. At each point in the dataset, a low degree polynomial is fitted to a subset of the data, with explanatory variable values near the point whose response is being estimated. The polynomial is fitted using weighted least squares, giving more weights to points near the point whose response is being estimated and less weights to points further away. The value of the regression function for the point is then obtained by evaluating the local polynomial using the explanatory variable values for that data point. The "LOESS" fit is complete after regression function values have been computed for each of the n data points. Some parameters of this method, such as the degree of the polynomial model and the weights, are flexible (15). (Appendix 1)

3.2. Random relocation (RR) test

Consider the sequence pattern of essential property indicators $\{X_i\} = \{1,0\}$ (The gene is essential corresponds to 1 or not to 0) of successive positions $i = 1,2,\ldots\ldots\ldots N$ of a chromosome of "length" N genes with exactly n essential genes. Assume a "symmetric" probability distribution having equal probabilities for all patterns in which gene position indicators are arbitrarily rearranged. For such distributions the conditional distribution, given the number of essential genes in each position probability equivalence class, are uniform over all coherent rearrangements. This fact can be straightforwardly used to construct randomization tests for null-hypotheses that correspond to distributions with such symmetries (16) (9).

3.3. Base-pair specific analysis

The idea behind the "Base pair specific analysis" is that, percentages of genes from different categories are counted in each base pair position from either telomere or centromere up to a specific base pair position. In the present study, for *Saccharomyces Cerevisiae*, 110000 base pair position are considered from both telomeric end and 55000 base pair position are considered on both side of the centromeres. In *Saccharomyces Cerevisiae*, there are 376 genes[7] with intron. As there are a few numbers of genes with intron, so we have counted introns as coding region and total gene length including intron and axon regions is counted as a coding region.

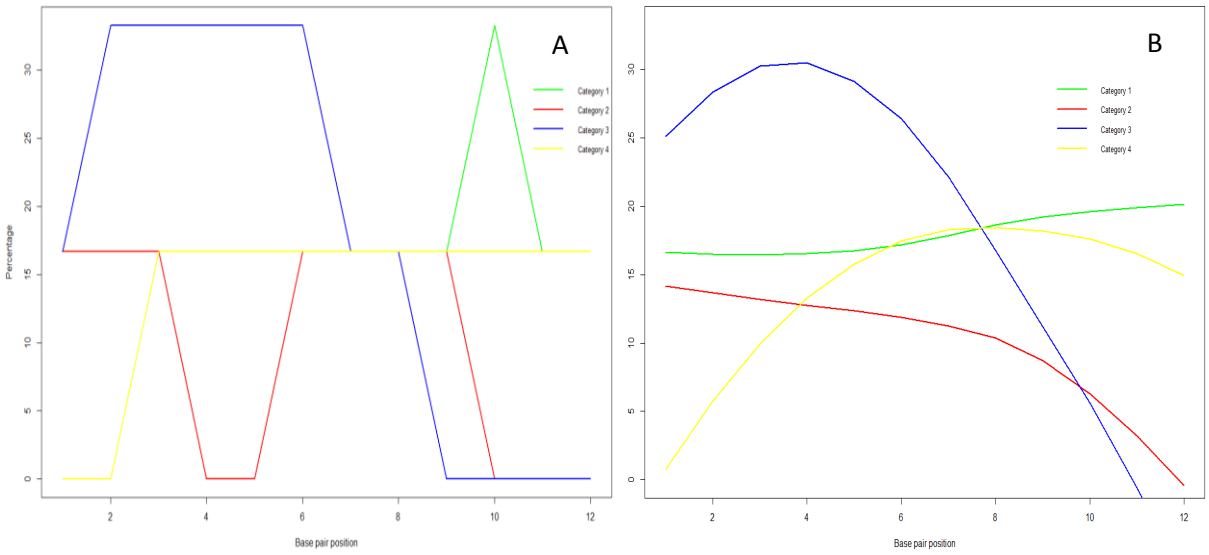
[7] 376 genes with intron in *Saccharomyces Cerevisiae* genome (Source : SGD)

Table 5: Base-pair specific analysis

Chromosomes	Base pair position											
	1	2	3	4	5	6	7	8	9	10	11	12
Chrom 1 – Tel 1												
Chrom 1 – Tel 2												
Chrom 2 – Tel 1												
Chrom 2 – Tel 2												
Chrom 3 – Tel 1												
Chrom 3 – Tel 2												
% of gene in Category 1	16.7	16.7	16.7	16.7	16.7	16.7	16.7	16.7	16.7	33.3	16.7	16.7
% of gene in Category 2	16.7	16.7	16.7	0.0	0.0	16.7	16.7	16.7	16.7	0.0	0.0	0.0
% of gene in Category 3	16.7	33.3	33.3	33.3	33.3	33.3	16.7	16.7	0.0	0.0	0.0	0.0
% of gene in Category 4	0.0	0.0	16.7	16.7	16.7	16.7	16.7	16.7	16.7	16.7	16.7	16.7

Table 5, explain an example of base pair specific analysis. For this illustration let us consider an organism with 3 chromosomes. Observation has been conducted up to 12 base pair position on each of 3 chromosome at each of total 6 telomere and every colored segment present a gene and non-color segment present non-coding region. The four color (Green, red, blue and yellow) define four categories of genes. Then in every base pair position, the percentage of existence of different category genes is counted.

Figure 5: Percentage of gene category in every base pair position before and after smoothing (LOESS method)



In figure 5A, the raw data has been plotted with linear interpretation but in figure 5B the same figure has plotted but with smoothing using LOESS method.

Percentages of different category in every base pair position are plotted in figure 5A. But for large data points it will be very difficult to understand a pattern or trend of every category from each base pair position. So LOESS method is used on each category for smoothing the curve to observe a better trend in the dataset (Figure 5B).

3.4. Gene Ontology (GO) term specific analysis

3.4.1. GO Ontology (GO)

The Gene Ontology (GO) project is a bioinformatics effort to categorize, represent and explain genes and gene products (Proteins), attribute (Function, characteristics etc.) with the help of a controlled vocabulary of terms for any specific species. The principle aims of GO project are

1. To maintain and develop a controlled vocabulary of terms for genes and gene product attributes like function , characteristics etc.
2. Annotate genes or gene products and then arrange these annotation data based on similarity or dissimilarity.
3. Arrange data obtained from the controlled vocabulary of terms in such a way that can be used for any future analysis.

The Gene Ontology project provides a controlled vocabulary list that represents the attribute of gene product. This controlled vocabulary list or ontology can be classified into 3 main broad categories of domains which are

(a) Cellular Component:
A cellular component is a component of a cell that could be a part of some larger object. This cellular component can be a cell anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a product group of genes (e.g. ribosome, proteasome or a protein dimer). The cellular component ontology describes locations, at the levels of subcellular structures and macromolecular complexes.

(b) Molecular Function:
A molecular function is a catalytic or binding activity that occurs at molecular level. Go molecular function term only focus on the activity but not on the molecule or complexes

(c) Biological Process :
A biological process is a series of events that occurred by one or more molecular function.

3.4.2. Over representation tests for GO terms

3.4.2.1. What is over representation of GO term

Any GO terms that have a statistically defined sub-statistically larger than expected subset in the selected genes, may be defined as over represented. These GO terms give insight into the functional characteristics of the gene list. The common test is for over representation, but one can also test for under representation.

3.4.2.2. Basic concept: The Urn model

Suppose there are N balls (Genes) in an urn (Genome), n are white (Genes that share a specific GO term) and m are black (Genes that do not share the specific GO term). Drawing K balls (A list of genes) out of the urn (Genome) without replacement, how many white balls (Genes that share a specific GO term) do we expect to get? What is the probability of getting x white balls (Genes that share a specific GO term)?

3.4.2.3. Hyper-geometric distribution model

The hyper geometric distribution describes the probabilities associated with simple random sampling (SRS) without replacement from a finite population and all element of this finite population has an equal chance of being drawn (17). As in hyper geometric distribution the drawn element are not replaced, so each selection influences the number of individuals of a certain type/category left in the population. So each selection depends on previous selections, unlike binomial distribution which is based on independent trial with replacement.

A random variable X follows the hyper geometric distribution with parameter N, n and K if its probability mass function is as

$$P\ (X = x) = \frac{\binom{n}{x}\binom{N-n}{K-x}}{\binom{N}{K}} \qquad 1.$$

Where , N = Population size (List of genes in reference category)

n = Number of success states in the population (Genes from reference category with specific GO term)

K= Number of draws (List of observe genes from reference category)

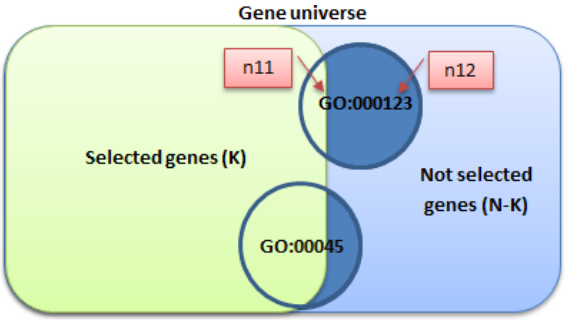
x = Number of successes (Genes from observed category with specific GO term)

3.4.2.4. Testing a GO term for over representation

A specific GO term of interest can be observed to be or not in a list of interested genes by comparing with reference gene list using the hyper geometric distribution model.

	Selected genes (Observed genes)	**Not selected genes**	
In GO term	n11	n12	n
Not in GO term	n21 or K-n11	n22	m
	K	N-K	N = n+m

In the figure the GO terms GO:000123 and GO:000456 can be tested for over representation in the selected genes categories by the hyper geometric distribution model describe in equation (1).

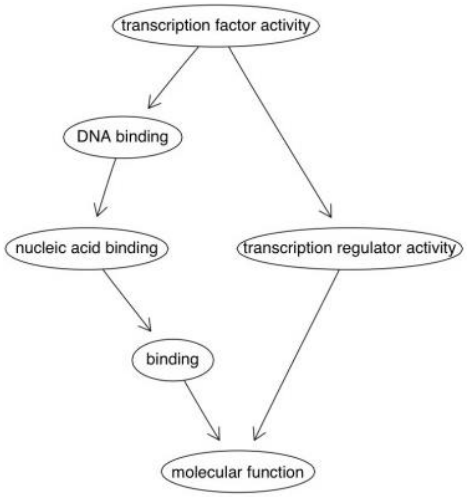


The significance level (α) for hyper geometric distribution can be 0.05 or 0.01 or any specific value of interest. It is known that, p value is the probability of observing a test statistic at least as large as the one calculated assuming the null hypothesis is true. The p-value obtained from the hyper-geometric distribution model is used to accept or reject the null hypothesis (H_0: There is no significant effect of the GO term vs. H_1: There is significant effect of the GO term) by comparing with the significance level (α) and reject the null hypothesis if the p value is smaller.

3.4.3. Gene Ontology (GO) hyper-geometric testing problems

3.4.3.1. Problems of applying hyper-geometric distribution for GO terms

There are some problems regarding the application of hyper-geometric distribution model for GO terms. The problems are



1. **Problem regarding multiple testing:**
   As many tests are performed simultaneously, p-values must be interpreted with much care from biological significant point of view rather than statistical significant point of view. To overcome this problem adjusted p value for multiple comparisons was calculated.
2. **Problem regarding biological significant term:** Hyper geometric test results of significant GO terms often include directly related terms. But are the both term biologically significant?

3.4.3.2. Solution for problems of applying hyper-geometric distribution for GO terms

3.4.3.2.1. Adjusted p-value

The problem of multiple testing can be solved by adjusted P value. There is nothing special about the significance level (α) which could be 0.05, 0.01 or any other value. Any value of significance level can be set as a threshold for the p-value. The adjusted P value is the smallest family wise significance level at which a particular comparison will be declared statistically significant as part of the multiple comparison testing.

As for example, two multiple comparisons test was performed. For the first test significance level was 5% and for the second test significance level is 1%. If a particular comparison is statistically significant by the first calculations (5% significance level) but is not for the second (1% significance level), then its adjusted P value must be between 0.01 and 0.05, say it could be 0.0323. Each comparison will have a unique adjusted P value. But these P values are computed from all the comparisons, and really can't be interpreted for just one comparison. The adjustment methods include

(a) Bonferroni correction
(b) Holm (1979)
(c) Hochberg (1988)
(d) Hommel (1988)
(e) Benjamini & Hochberg - “BH” (1995)
(f) Benjamini & Yekutieli - “BY” (2001)

In Bonferroni correction the p-value are multiplied by the number of comparisons. Holm, Hochberg, Hommel, Benjamini & Hochberg and Benjamini & Yekutieli methods are less conservative. The first four methods Bonferroni, Holm, Hochberg and Hommel are designed to give strong control of the family-wise error rate. Hochberg's and Hommel's methods are valid when the hypothesis tests are independent or when they are non-negatively associated

(18). Hommel's method is more powerful than Hochberg's, but the difference is usually small and the Hochberg p-values are faster to compute. The "BH" and "BY" method of Benjamini, Hochberg, and Yekutieli control the false discovery rate, the expected proportion of false discoveries amongst the rejected hypotheses. The false discovery rate is a less stringent condition than the family-wise error rate, so these methods are more powerful than the others.

3.4.3.2.2. Biologically relevant term from statistical significant term

At a given significance level (say α=0.05), statistically significant terms form all given set of terms are selected by hyper-geometric distribution model with enrichment probabilities (i.e. p-values) corrected by different methods like Bonferroni correction, Benjamini & Hochberg - "BH" and Benjamini & Yekutieli "BY". But when significant GO term include directly related GO term, in that case it is necessary to differentiate between biologically significant term and statistically significant term. The concept of redundancy (19) is very useful to handle this situation and can be used as a practical solution of the problems regarding independence and biological significant term.

3.4.3.2.2.1. What and why is redundancy

List of genes connected to specific GO term can easily found by using Gene Ontology (GO) database based on the assumption that, the GO term in which the frequencies of interesting genes are significantly higher than expected by chance are likely to be interesting GO term (20). According to Gene Ontology's "true path rule" each gene annotated by a term is also annotated by its ancestor terms (20) (21). For this reason many redundant ancestor-offspring relation of GO terms are found in a list of observed gene. The way to address this problem is to select more specific GO terms from a group of terms by assuming that specific terms might have more biological meaning. In this situation the biological relevance of a term with a specific characteristic is observed from the data rather than the prior assumption (22) (23).

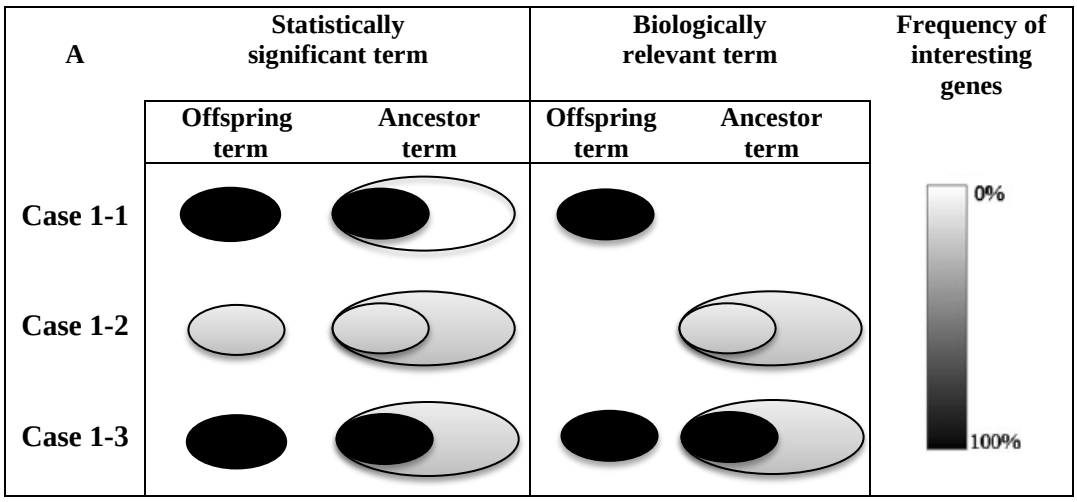
All the terms connected with a list of genes are not significant. Statistically significant GO term can easily be found by using traditional enrichment analysis tool like hyper-geometric distribution model. Then by applying redundancy treatment on statistically significant terms can be more appropriate to get the actual effect of a specific GO term on a list of interested gene (23).

3.4.3.2.2.2. Local redundancy

Local redundancy treatment is applicable for GO terms with ancestor-offspring relationship. In the statistically significant GO term by applying Hyper-geometric distribution model, an ancestor term and an off-spring term both can be detected as statistically significant in several situations. But to define them biologically significant, local redundancy treatment can be applied. There are three rules for this local redundancy treatment (23). They are as follows

1. If the frequency of interesting genes in the remaining genes of ancestor term is not significantly less or equal or higher than the frequency of interesting gene of offspring term, then the offspring term is selected as biologically significant term (Case 1-1).
2. If the frequency of interesting genes in the remaining genes of ancestor term is same as the frequency of interesting gene of offspring term, then the ancestor term is selected as biologically significant term (Case 1-2).

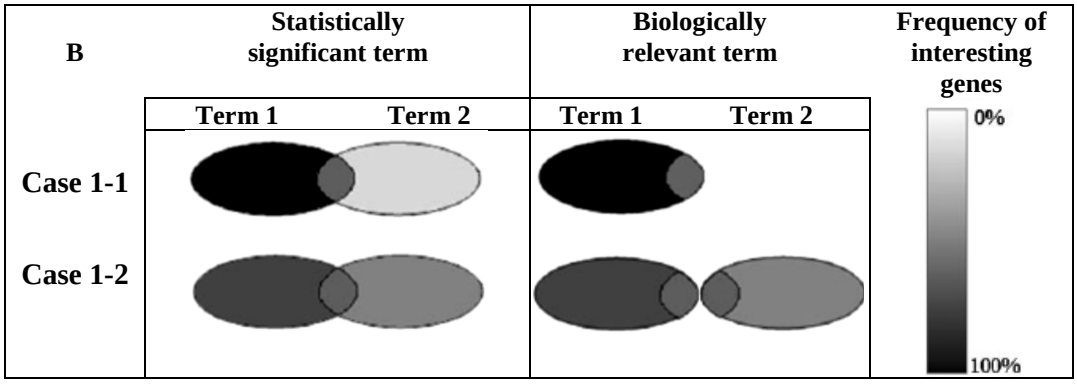
3. If the frequency of interesting genes in the remaining genes of ancestor term is very close to the frequency of interesting gene of offspring term, then the both offspring term and ancestor term are selected as biologically significant term (Case 1-3).



3.4.3.2.2.3. Global redundancy

Genes with multiple annotations of term with no ancestor-offspring relationship is another problem of gene enrichment analysis. Global redundancy treatment applied between GO terms which share multiple function genes but have no ancient-offspring relationship. For a pair of terms with overlapping genes, global redundancy treatment can identify whether their significance might be simply introduced by the overlapping genes or not (23). There are two rules for this global redundancy treatment. They are as follows

1. If the frequency of interesting genes in non-overlapping genes of any term is not significantly higher, then the term can deleted and the other term can be define as biologically significant (Case 2-1).
2. If the frequency of interesting genes in non-overlapping genes of both term is very close to each other, then the both term are define as biologically significant.



3.5. Progressive alignment approach

The most widely used approach to multiple sequence alignment (MSA) that uses a heuristic search which is known as progressing alignment approach. In this approach, MSA is created

by combining pairwise alignment beginning with the most similar pair and progressing to the most distantly related pair. The progressive alignment approach are conducted in two step

Step 1: In the first step pairwise sequence alignment are done and then the sequences are visualized by a guide tree.

Step 2: In the second step the MSA is developed by adding the sequences sequentially to the growing MSA according to the guide tree. The primary guide tree is constructed by a clustering method such as neighbor-joining, UPGMA or distances based method.

3.5.2. Guide tree

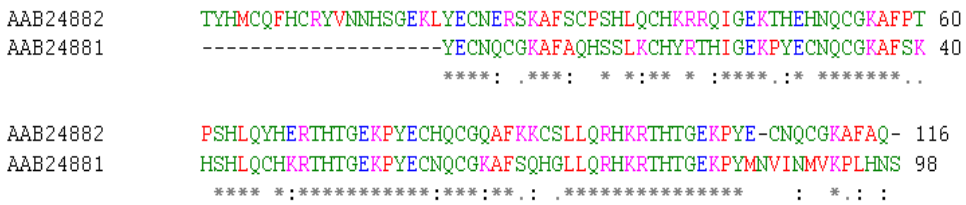
A phylogeny is the evolutionary history of a group of entities. The main aim of phylogeny reconstruction is to describe evolutionary relationships in terms of relative regency of common ancestry. These relationships are represented as a branching diagram or tree with branches joined by nodes and leading to terminals at the tips of the tree (24).

3.5.3. Multiple sequence alignment (MSA)

Phylogenetic methodology relies on the assumption that the characters used to generate trees are homologous. For gene family phylogenies careful alignment of sequence data full fills this requirement. Sequence alignment can be achieved automatically or manually. Automatic alignments may fail to correctly identify regions of conservation within a gene, whereas manual alignments allow this but are more labor intensive.

Sequence alignment is a technique to arrange the DNA, RNA or protein sequences to identify the similar sequence region that may be a consequence of functional, structural or evolutionary similarity or relationships between the sequences. Nucleotide or amino acid residues in the aligned sequence are generally represented as rows within a matrix. Gaps are inserted between the nucleotide or amino acid residues so that identical or similarity increase in successive columns (25).

Figure 6: Protein sequence alignment output by ClustalW (Human zinc finger proteins with GenBank accession number)



Multiple sequence alignment is an extension of pairwise sequence alignment to align more than two sequences at a time. Global multiple sequence alignment methods try to align all of the sequences in a given query set. Multiple sequence alignments are used to identify the conserved sequence regions in a group of sequences that are hypothesized to be evolutionarily related. These kind of conserved sequence motifs can be used to obtain structural and functional information to locate the catalytic active or action sites of enzymes. Sequence alignments are also used to construct phylogenetic trees for obtaining evolutionary

relationships (26) (27). Often they are used to compare genes between species, but we will use them to identify functional gene families within *Saccharomyces Cerevisiae*.

3.5.4. Method of analysis

Once data are aligned, there are many different types of phylogenetic analysis that can be implemented. The type of phylogenetic analysis used will be determined by compromise between the length of computational time and the degree of rigor required. The main techniques are distance, parsimony and likelihood (including Bayesian analysis). For the present research analysis distance method is used for phylogenetic analysis.

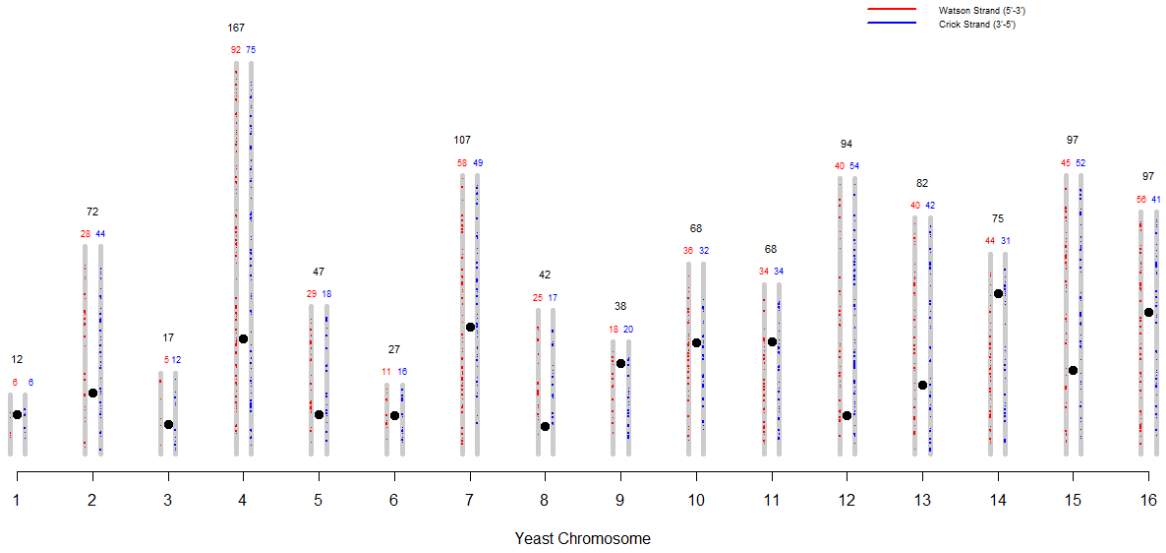
Distance methods [e.g. neighbor joining (NJ), distance and minimum evolution] calculate pairwise distances between sequences and group sequences. This approach is computationally simple and fast. However, distance methods do not allow an analysis of which characters contribute to particular groupings. As with other methods, the outcome may depend on the order in which entities are added to the starting tree, but because only one tree is outputted it is not possible to examine conflicting tree topologies. Although distance methods are often useful for making an initial tree, they should be used for final trees with caution. Instead, parsimony and likelihood are preferred because they have the potential to rigorously explore the relationship between the tree and the entities included. Parsimony and likelihood use different criteria to choose the best trees.

4.1. Random phenomenon of gene distribution in yeast genome

4.1.1. Genome wide essential gene distribution

For the present analysis a special category of genes from *Saccharomyces Cerevisiae* was chosen which possess the essential characteristics. There are 1110 essential genes in yeast and their chromosomal distribution is represented graphically by figure 7. In table 6, the chromosomal proportions of these 1110 genes having the essential characteristics is shown.

Figure 7: Essential genes distribution in 16 chromosomes



The graphical representation of all 1110 essential gene are presented in figure 7 . In the figure 7, the red color is the genes on Watson strand (5'-3') and blue color is the genes are on crick strand (3'-5'). The big black dot presenting the centromere of the chromosome.

Table 6: Percentage distribution of essential genes in 16 chromosomes

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Essential gene Frequency	12	72	17	167	47	27	107	42	38	68	68	94	82	75	97	97
Total gene frequency	92	402	159	738	225	125	520	278	210	353	313	501	453	394	529	456
% of Essential gene Frequency	13.04	17.91	10.69	22.62	17.09	21.60	20.57	15.10	18.09	19.26	21.72	18.76	18.10	19.03	18.33	21.27

In table 6, the percentages of essential genes in each chromosome are counted by comparing with the total gene frequency in the corresponding chromosome.

4.1.2. Telomeric essential genes distribution is not a random phenomenon

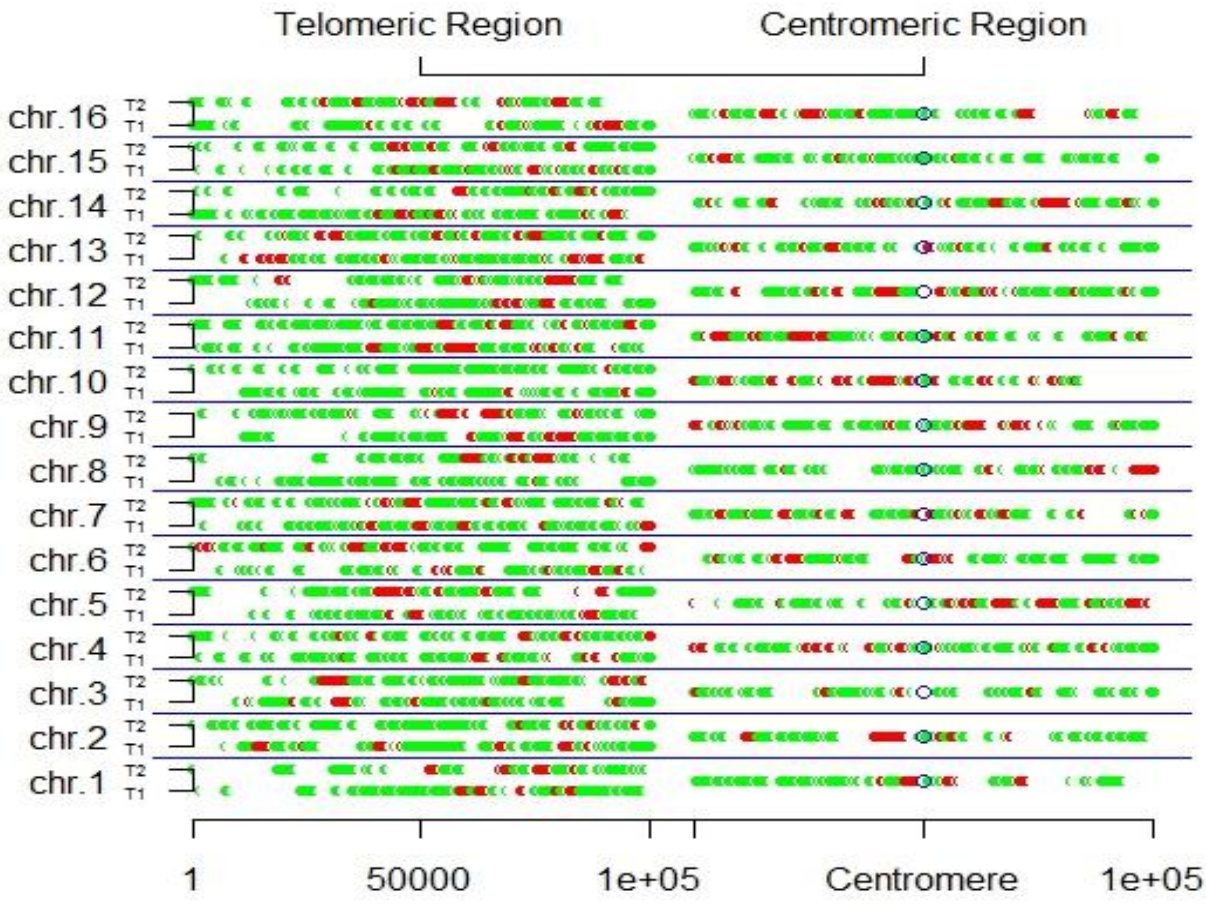
Highly pleiotropic genes[8] in *S.Cerevisiae* are less likely to be located in the telomeric regions near the chromosomal endpoints of the five of the smallest chromosomes in yeast (28). The most possible explanation can be the endpoints of the chromosomes are hot spots of genetic churning (29), and in

[8] Pleiotropic Gene: A gene that produce many effects in the phenotype.

this chromosomal end points mutation and recombination occur most frequently. So the selection likes to exclude the highly pleiotropic and essential genes from the chromosomal endpoints. But, at these chromosomal end regions many genes from larger gene families are to be found (9).

To find the telomeric effect, we have plotted the essential and non-essential genes according to gene's base pair position for both telomeric and centromeric region. For telomere we have plotted the gene position up to 100000 base pair from the telomeric end points and for centromeric region we have plotted 50000 base pair for both left & right arm of telomere from the end of centromere.

Figure 8: Essential and non-essential gene position in telomeric and centromeric region



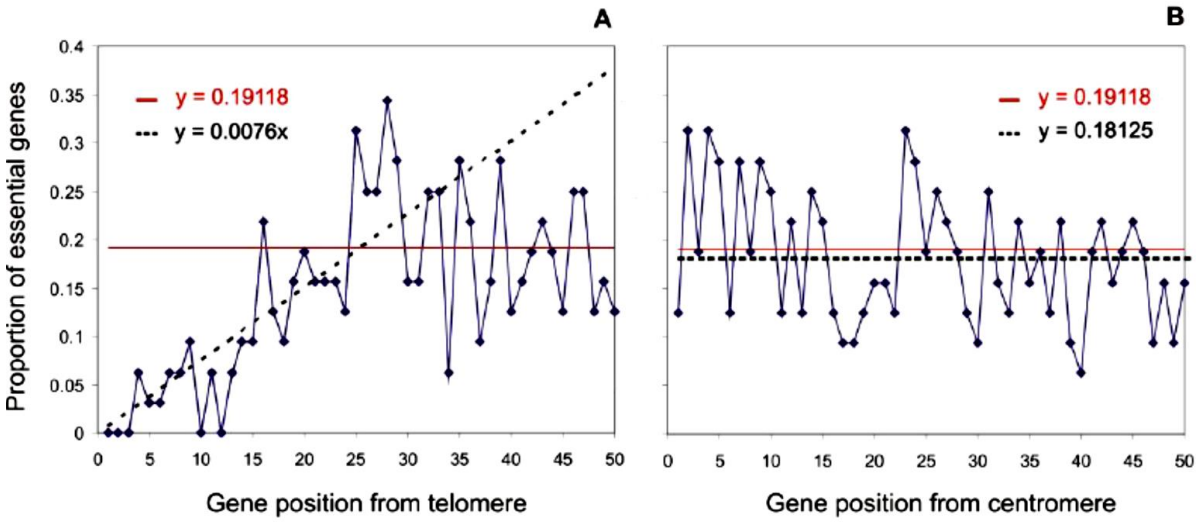
In figure 8, the red color assigns the essential genes, the green spot define the non-essential gene and the white space define non-coding region. For telemetric region 100000 base pair and for centromeric region 50000 base pair are considered on both side of centromere.

In figure 8, till 50000 base pair position from telomeric end points (T1 and T2) for both telomeres, there are less essential genes. But after 50000 base pair position number of essential genes increase. But in the centromeric region we don't observe any specific pattern of essential gene's position. We can explain this situation by the following possible explanation.

1. As telomeric end points are the hotspot of genetic mutation and recombination, so selection includes the less essential genes in that region (29) (9).
2. Due to capping effect, there are fewer genes in the telomeric region (2) (3).

The telomeric effect can be observed very clearly, by plotting the percentage of essential genes as a function of gene order distance to the nearest telomere using the aggregate data from all 32 telomeric regions from 16 chromosomes (9).

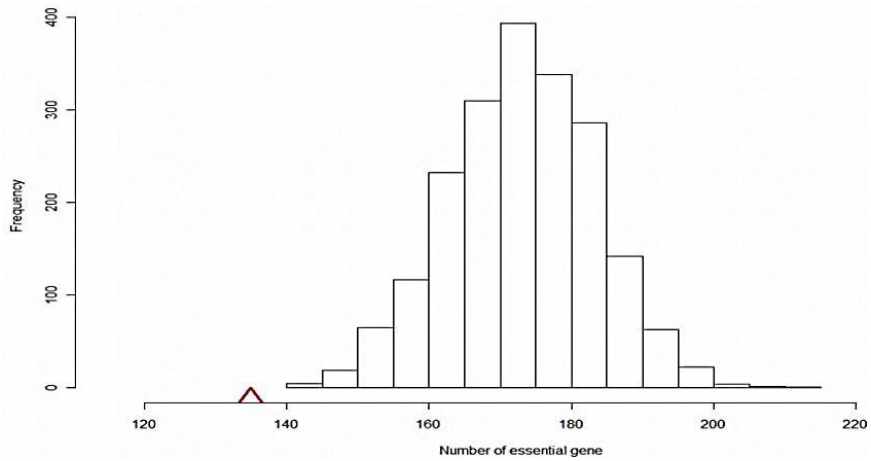
Figure 9: Proportion of essential gene in certain gene order position measured from (A) Telomeric end region (B) Centomeric region



In figure 9A, in first gene position at telomeric end we have found no essential gene. Moving to the centomeric region from the telomeric region, the percentage of essential gene in every gene order position is roughly gradually increasing up to 20 % till 25th gene position. Till this 25th gene position there is an increasing tendency (positive correlation between gene position and percentage of essential genes) of essential gene percentage and after 25th gene position this telomeric effect seems to have vanished. But in Figure 9B, the same analysis was done for centromeric region and we have not observed any centromeric effect other than the percentage of essential gene is approximately follow 20% up to 50th gene order position from centromere (9).

In figure 8 and figure 9, the same telomeric effect is observed by both base pair specific gene position analysis and gene order distance specific analysis. But to define this telomeric effect phenomenon as a random or non-random event, random relocation (RR test) test can be used (9). For RR test, the test statistics counting the total number of essential genes with gene order distance from telomere end to 25th gene position. For this RR test the null hypothesis (H_0) is, there is no telomeric effect against the alternative hypothesis (H_1), there is telemetric effect. We have done 2000 randomized simulations for RR test.

Figure 10: Random relocation test result for telomere up to 25th gene order position.



In figure 10, in 2000 simulation the mean of the number of essential genes up to 25th gene position is 171 with minimum value 140 and maximum value 215.

From figure 10, the null hypothesis was rejected, as the actual number of essential gene up to 25th gene position from telomeric end of all 16 chromosomes is 135. But in 2000 simulations, the expected average is 171. So there is a big difference between the observed case and expected case. Not a single simulation is equal or less that the observed case. So it can be concluded that the telomeric effect of essential gene distribution clearly is not a truly random phenomenon.

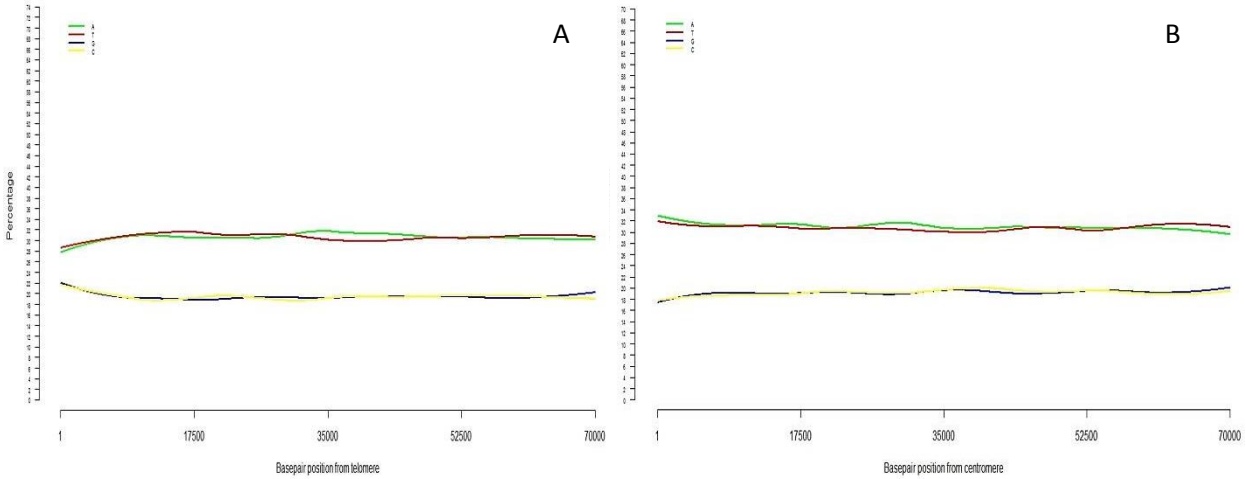
4.2. Base pair specific analysis

The base pair specific analysis was conducted on 5798 yeast gene by considering the essential property of the genes. Among 5798 genes, 1110 gene have this essential property and rest 4688 genes are non-essential. Also gene length was considered for the analysis where longer genes are defined as longer than median length and shorter genes are define as shorter than median length of all genes.

4.2.1. Base pair wise nucleotides frequency

To understand the telomeric effect on nucleotide frequency, the base pair specific analysis has conducted in both telomeric and centromeric region till 70000 base pair in telomeric region and till 35000 base pair on both side of centromere.

Figure 11: Base pair wise nucleotide frequency in telomeric and centromeric region



In figure 11, all 5798 genes were considered to observe percentage of nucleotide frequency in every base pair position. In the telomeric position (Figure 11A) the percentage of A and G is about 28 % and 22% respectively. But in the centromeric region (Figure 11B) the percentage of A and G is about 33% and 18 % respectively. T and C is complementary of A and G, which is clearly observed from the figure.

From figure 11, it has observed that frequency of Adenine (A) and Guanine (G) differs between very end of telomeric region and centromeric region. At telomeric region the relative frequency of A and G is about 28% and 22 % respectively. Whereas in centomeric region the relative frequency of A and G is about 33% and 18% respectively. In the region close to telomeric end the relative frequency of A is lower and the relative frequency of G is higher than the region close to centromere. In other region of chromosome except region close to telomere end and centromere, the relative frequency of A and G is always same.

4.2.2. Coding, non-coding, essential and non-essential genes at telomeric and centromeric region

The frequency of base pair in coding region, non-coding region, essential gene and in non-essential gene at every pair position is observed by base pair specific position analysis.

Figure 12: Frequency of base pair in coding, non-coding, essential gene and non-essential gene region at every base pair position from centroeric and telomeric region

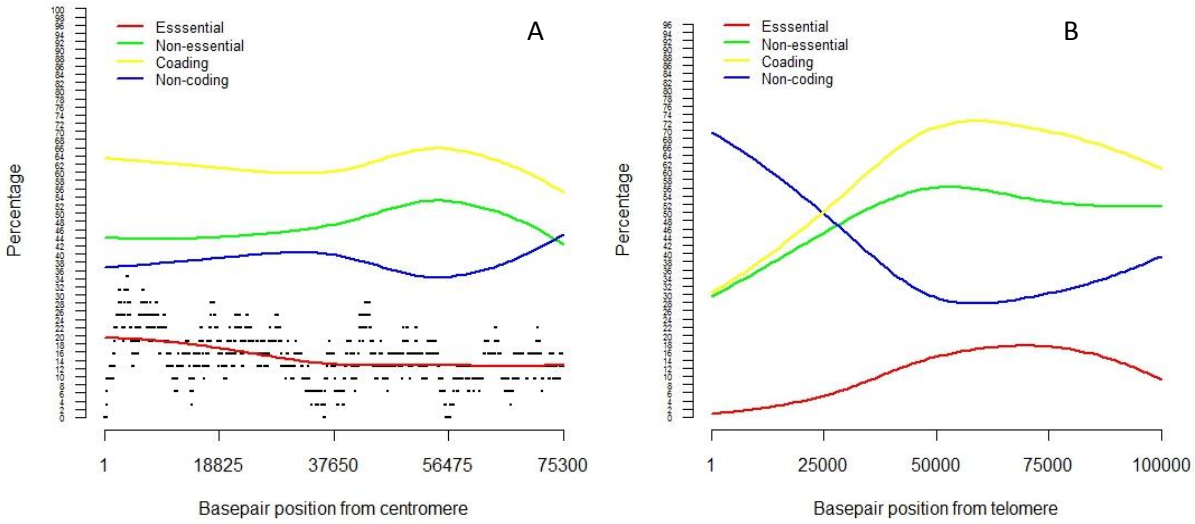


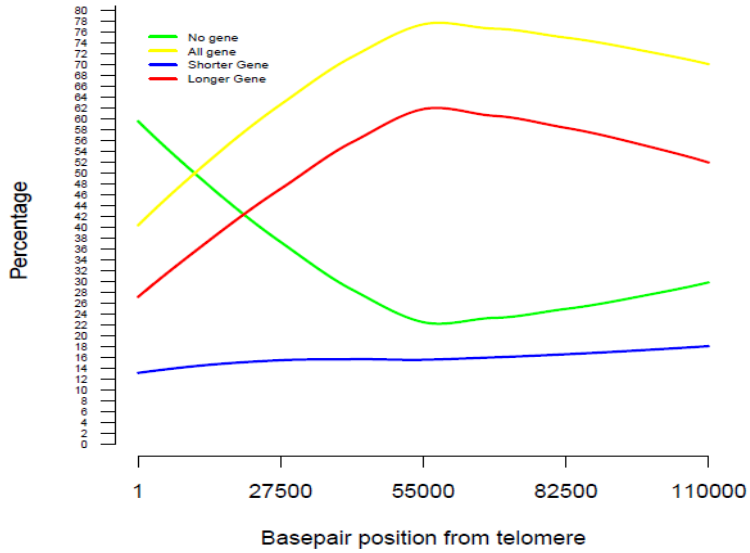
Figure 12 explain the percentage of coding (5798 genes), non-coding (5798 genes), essential gene (1110 genes) and non-essential gene (4688 genes) frequency for every base pair position in both centromeric and telomeric region.

In centromeric region (Figure 12A), till 75300 base pair the observation was conducted and no significant trend was found except almost straight line. But in telomeric region (Figure 12B), increasing trend is observed in per base pair frequency for essential gene, non-essential and coding region. Visa versa, decreasing trend was observed only in non-coding region which support the previous results (9).

4.2.3. Shorter and longer gene at telomeric region and cutoff point for telomeric region

The relative frequency of shorter and longer gene at telomere has observed by applying base pair specific analysis where the longer and shorter genes are defined by the median of overall gene's length. Also the telomeric cutoff point has observed where the fitted line (By using the LOESS method in base pair specific analysis) shows a clear signal of changing the direction of its trend.

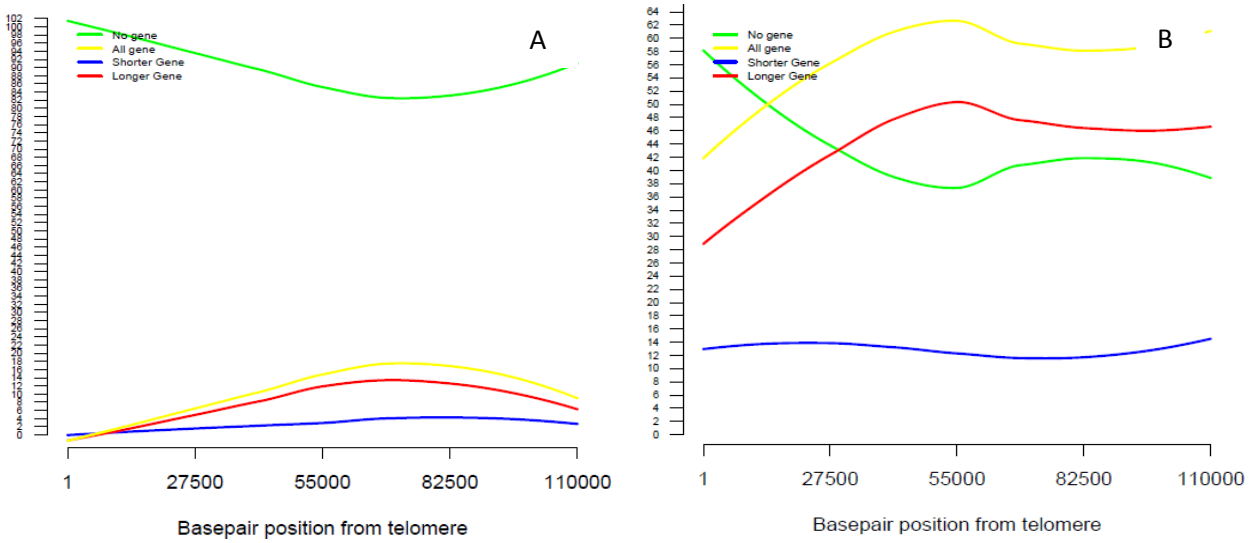
Figure 13: Frequency of base pair containing shorter gene and longer gene. Threshold level = 1206 base pair (Median)



In figure 13, all 5798 yeast genes has consider to count the percentage of shorter or longer gene in every base pair position at both telomere of all 16 chromosomes from base pair position 1 to 110000. The median of length of all genes which is 1206 has taken as the threshold of shorter and longer gene.

In figure 13, it has observed that in all 5798 genes, the frequency of long genes increase sharply compare with shorter gene from telomeric end points. Also the frequency of non-coding region at every base pair position decrease form telomeric region to centromeric region.

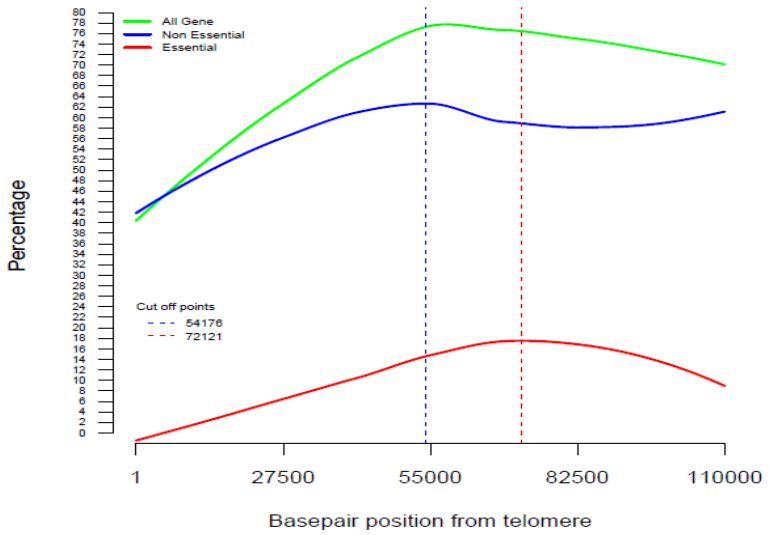
Figure 14: Frequency of base pair containing shorter gene and longer gene in essential (Left side) and non-essential (right side) genes. Threshold level to define shorter and longer gene for essential and non-essential genes are 1359 base pair and 1179 base pair respectively



In figure 14A, essential genes (1110 essential genes) has consider to count the percentage of shorter or longer gene in every base pair position at both telomere of all 16 chromosomes from base pair position 1 to 110000. The median of length of all essential genes which is 1359 has taken as the threshold of shorter and longer gene. In figure 14B, non-essential gene (4688 non-essential genes) has consider to count the percentage of shorter or longer gene in every base pair position at both telomere of all 16 chromosomes from base pair position 1 to 110000. The median of length of all non-essential genes which is 1179 has taken as the threshold of shorter and longer gene.

From figure 14, for both essential and non-essential gene, it has observed that shorter gene has a flat tendency to be in telemetric region and longer gene has a tendency to avoid the telemetric region. So the gene groups that are shorter and longer could be interesting for further research about what function they have common.

Figure 15: Percentage of coding region, essential gene and non-essential gene at every base pair position from telomeric end

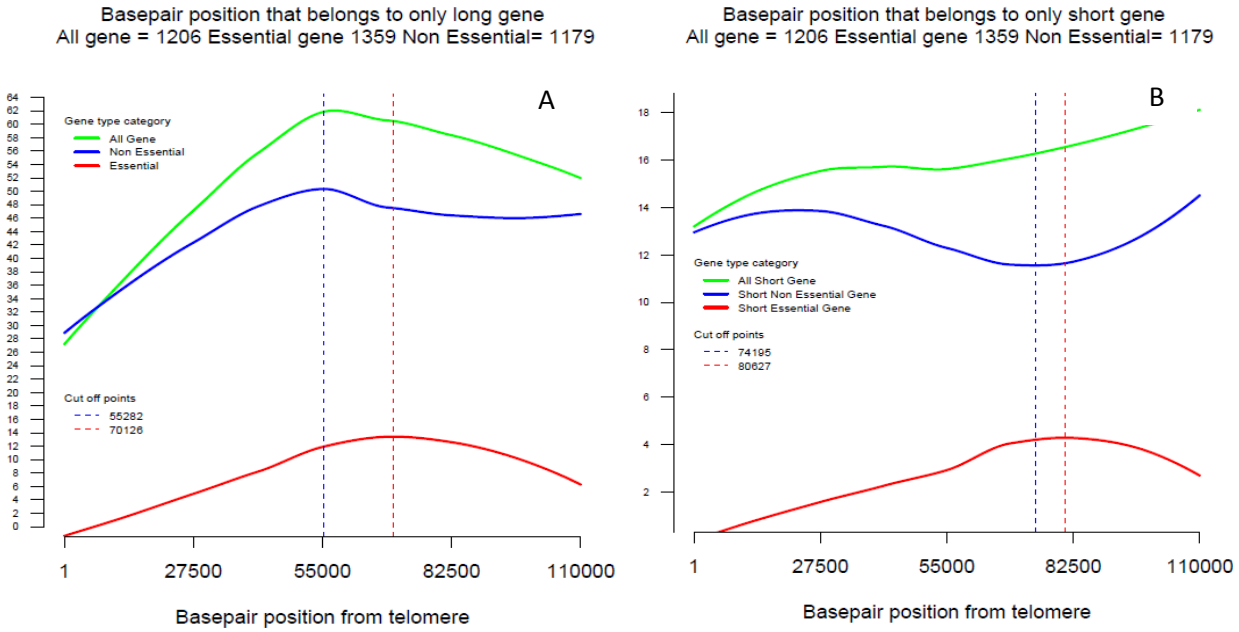


In figure 15, percentage of coding at 1 to 110000 base pair position has observed for 3 categories. a) All genes (Both essential and non-essential), b) Only non-essential and c) Only essential. In figure 15, the blue dotted line represents the base pair position at which the non-essential gene trend reaches its maximum[9] which is 54179 base pair from telomeric end point. The red dotted line indicates the essential gene trend maxima which is 72121 base pair from telomeric end.

In figure 15, the percentage of essential, non-essential and all genes at every base pair position are observed from telomeric end points to110000 base pair position. The dotted line for each curve define the points at which the corresponding curve reach its maximum where the essential gene reaches its maximum at 72121 base pair and non-essential gene reaches its maximum at 54179 base pair counted from telomeric end point. The points at which each curve reaches its maximum can be defined as the "Telomeric cutoff points" for corresponding gene category. So for essential gene the telomeric cut off point is 72121 base pair and for non-essential gene the telomeric cut off point is 54179 base pair.

[9] Cutoff : The base pair at which maxima or minima of any smooth line occur is define as cutoff point.

Figure 16: Percentage of coding region, essential gene and non-essential gene in every base pair position for longer gene (Left figure) and shorter gene (Right figure) separately



In figure 16A and 16B explain the percentage of coding, non-essential and essential genes in longer and shorter gene group respectively. The blue and red dotted line represents the relative maxima for corresponding color line which is defined as cutoff point. For longer gene (Figure 16A), the cut off for non-essential and essential genes are 55282 base pair and 70126 base pair. Similarly for shorter gene (Figure 16B), the cut off for non-essential and essential genes are 74195 base pair and 80627 base pair.

In figure 16A, percentage of longer gene at 1 to 110000 base pair position has observed for 3 categories, a) All genes (Both essential and non-essential), b) Only non-essential and c) Only essential. In that figure up to 55282 base pair, longer non-essential gene (blue line) has a tendency to avoid the telomeric region. From the both telomeric end point to 55282 base pair, in all 16 chromosome there are 339 non-essential genes which are long. Among them 292 are annotated and rest 47 are un-annotated. Genes that only started within that cut off point are considered as telomeric gene. In that figure up to 70126 base pair, longer essential gene (red line) has a tendency to avoid the telomeric region. From the both telomeric end-point to 70126 base pair, in all 16 chromosome there are 106 essential genes which are long. All of them are annotated. Genes that only started within that cut off point are considered as telomeric gene.

In figure 16B, percentage of shorter gene at 1 to 110000 base pair position has observed for 3 categories, a) All genes (Both essential and non-essential), b) Only non-essential and c) Only essential. In that figure, up to 74195 base pair, shorter non-essential gene has a tendency to conserve into telomeric region. From the both telomeric end point to 74195 base pair, in all 16 chromosome there are 478 genes which are short. Among them 325 are annotated and rest 163 are un-annotated. Gene that only started within that cut off point are considered as telomeric gene. In that figure, up to 80627 base pair, shorter essential gene has a tendency to avoid the telomeric region. From the both telomeric end point to 80627 base pair, in all 16 chromosome there are 113 essential genes which are short. Among them 111 are annotated and rest 2 are un-annotated. Genes that only started within that cut off point are considered as telomeric genes.

By comparing figure 15 (Coding region not considering length) and figure 16A (Longer genes), It has observed that by discarding only the shorter genes, the telomeric region avoiding tendency of all genes, essential genes and non-essential genes in figure 15 is almost same as figure 16A (Only longer genes are considered).

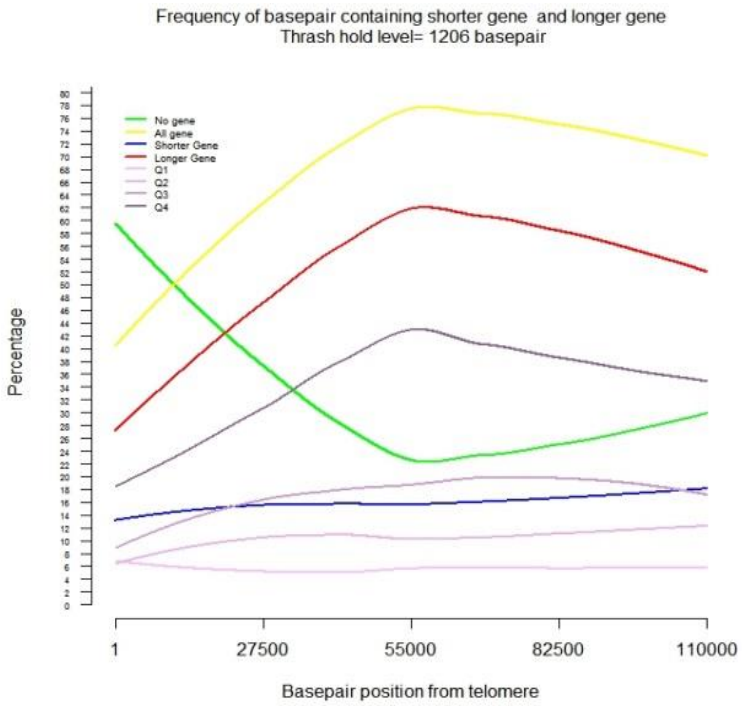
By comparing figure 16A and figure 16B, it has observed that non-essential shorter gene has a flat tendency in the telomeric region weather non-essential longer gene avoid the telomeric region sharply. But for essential gene, there is a decreasing tendency of keeping both longer and shorter gene in telomeric region.

In essential gene groups, for both shorter and longer gene, there is a lower tendency to keep genes in the telomeric region. But in non-essential genes, there could be a special reason or function to have the flat tendency of shorter gene and lower tendency of longer gene in the tolomeric region.

4.2.4. Gene distribution on quartile based on length of genes

Quartile is the set of values of 3 points that divides the whole data set into four equal parts. First quartile (Q1 or lower quartile) splits the lowest 25% of data which is define as 25th percentile. The second quartile (Q2 or median) divides the data set into half which is defined as 50th percentile. Third quartile (Q3 or upper quartile) splits the data set into highest 25% or lowest 75% which is define as 75th percentile. The difference between the upper and lower quartiles is called the interquartile range.

Figure 17: Percentage of all longer gene in every base pair position based on quartile of gene length

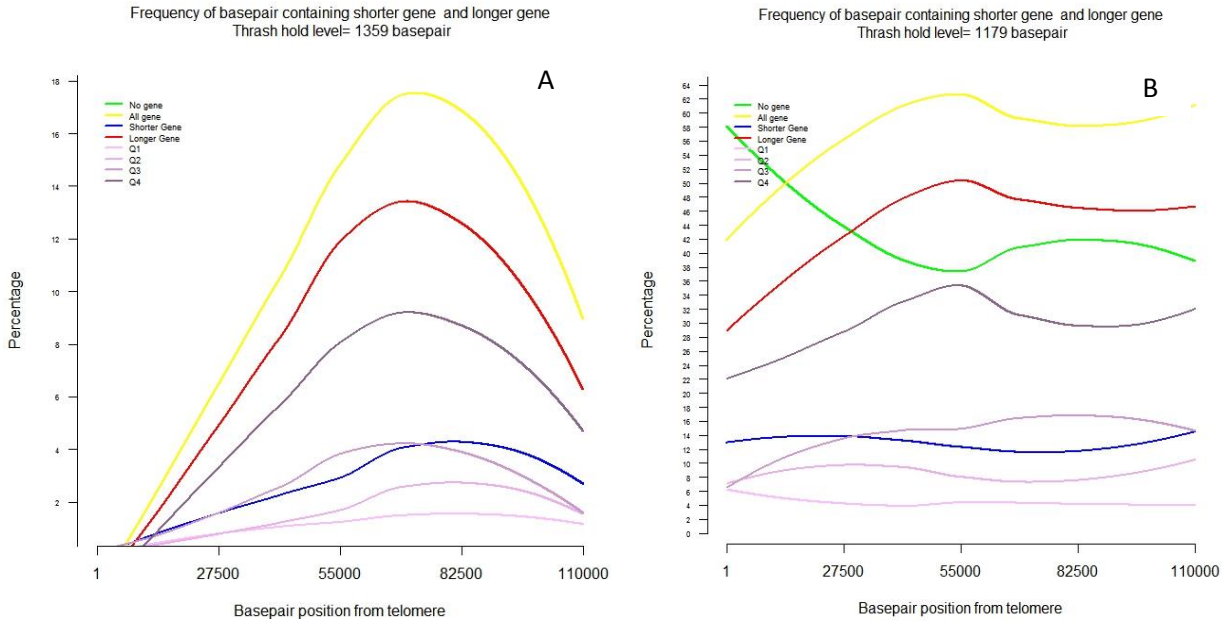


Quartile

Q1	Q2	Q3	Q4
25% Data	**25% Data**	**25% Data**	**25% Data**
lower 25% data	**Upper 75% data**		
Lower 50% data		**Upper 50% data**	
Lower 75% data			**Upper 75% data**

Figure 17 explains the length in different quartile group for all 5798 genes.

Figure 18: Percentage of longer essential and non-essential gene in every base pair position based on quartile of gene length



In figure 18A, explains the length in different quartile group for all 1100 essential genes and figure 18B explains the length in different quartile group for all 4688 non-essential genes.

The entire gene in essential and non-essential group, are categorized based on the quartile of length. From figure 17 and 18, it has observed that really long gene (Gene length greater than 3rd quartile) avoid the telomeric region sharply for both essential and non-essential genes group.

4.2.5. RNA genes in chromosome

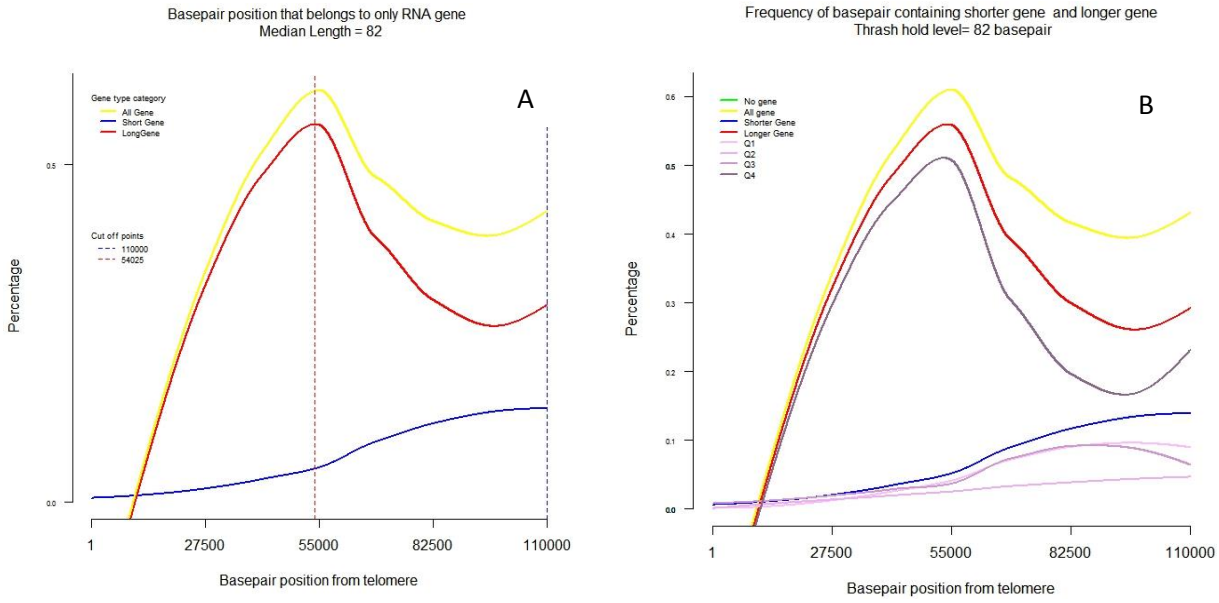
Distribution of 397 RNA genes in 16 yeast chromosome can be explained in the following table

Table 7: Distribution of RNA genes in 16 yeast chromosome

ncRNA	Not in systematic sequence of S288C	Pseudo gene	rRNA	snoRNA	snRNA	Transposable element gene	tRNA	Total
14	19	21	25	77	6	89	275	526
14	0	0	25	77	6	0	275	397

In table 7, 397 RNA genes in 16 yeast chromosomes are categorized in 5 different categories which are ncRNA, rRNA, snoRNA, snRNA and tRNA.

Figure 19: Percentage of base pair specific gene frequency for RNA genes



In figure 19, for 397 RNA genes the median length in 82 which define the shorter and longer gene category. In figure 19A, long 204 RNA genes has telomeric cutoff point on 54025 base pair and 193 short genes has no specific cut of points as is has an increasing tendency till 110000 bp. In right side figure, quartile specific analysis based on gene length has observed for 204 long RNA genes.

4.3. GO term specific analysis

4.3.1. GO term comparison groups

Overrepresentation of the GO term in a list of genes can be observed based on a reference group. As we have categorized chromosomal genes in 3 different category which are essential genes, non-essential genes and RNA genes, so over representation of GO term can be observed in this 3 categories considering the other feature like longer/shorter genes and annotated /un-annotated gene in the telomeric region. The comparison groups for GO terms overrepresentation in 3 categories can be summarized in the following table.

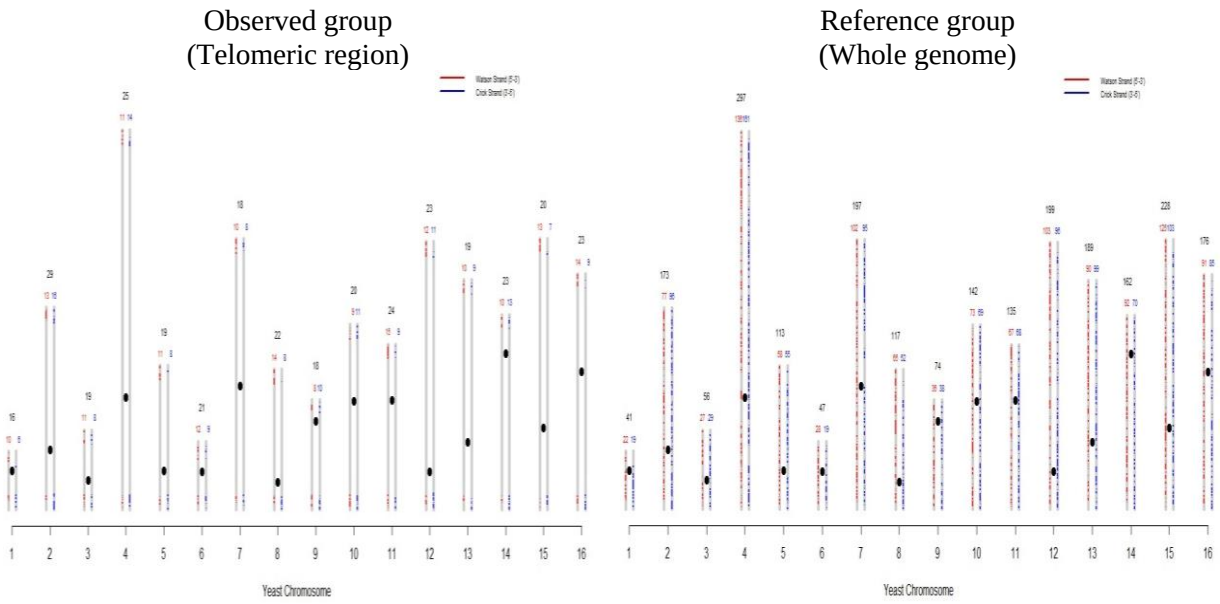
Table 8: Comparison groups for GO term overrepresentation analysis

		Observed group	Reference Group
Essential Gene	**1**	All long genes	All genes
	2	All short genes	All genes
	3	All long under cut off	All long genes
	4	All short under cut off	All short genes
	5	All long annotated under cut off	All long annotated
	6	All long un-annotated under cut off	All long un-annotated
	7	All short annotated under cut off	All short annotated
	8	All short un-annotated under cut off	All short un-annotated
Non-essential genes	**9**	All long genes	All genes
	10	All short genes	All genes
	11	**All long under cut off**	**All long genes**
	12	All short under cut off	All short genes
	13	All long annotated under cut off	All long annotated
	14	All long un-annotated under cut off	All long un-annotated
	15	All short annotated under cut off	All short annotated
	16	All short un-annotated under cut off	All short un-annotated
RNA genes	**17**	All long genes	All genes
	18	All short genes	All genes
	19	All long under cut off	All long genes

Among all the chromosomal gene (5798) considered for analysis, 4688 genes were non-essential gene, 1110 genes were essential gene and 397 genes were RNA genes. So further investigation on non-essential gene could be more interesting and informative because of their abundance. Among these non-essential genes 2346 genes were longer and 2342 genes were shorter. It has observed that shorter non-essential gene showed a flat tendency in the telomeric region but long non-essential genes showed an increasing tendency in the telomeric region in base pair specific analysis. Therefore investigation on telomeric long non-essential gene could be more interesting and informative.

For GO term over representation analysis by hyper geometric distribution model, telometric long non-essential genes were selected as observed group and all long non-essential genes were selected as reference group (Group 11 in table 8).

Figure 20: Long non-essential gene distribution in telomeric region and whole genome



Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Telomeric region gene frequency (Observed group)	16	29	19	25	19	21	18	22	18	20	24	23	19	23	20	23
Whole genome gene frequency (Reference group)	41	173	56	297	113	47	197	117	74	142	135	199	189	162	228	176

In figure 20, the distribution of 2346 long non-essential genes distribution has observed in both tomeric region and whole genome . In telomeric region there are 339 long essential gene in which 292 gene are annotated and rest 47 in un-annotated .

4.3.2. Telomeric long non-essential vs. all long non-essential genes

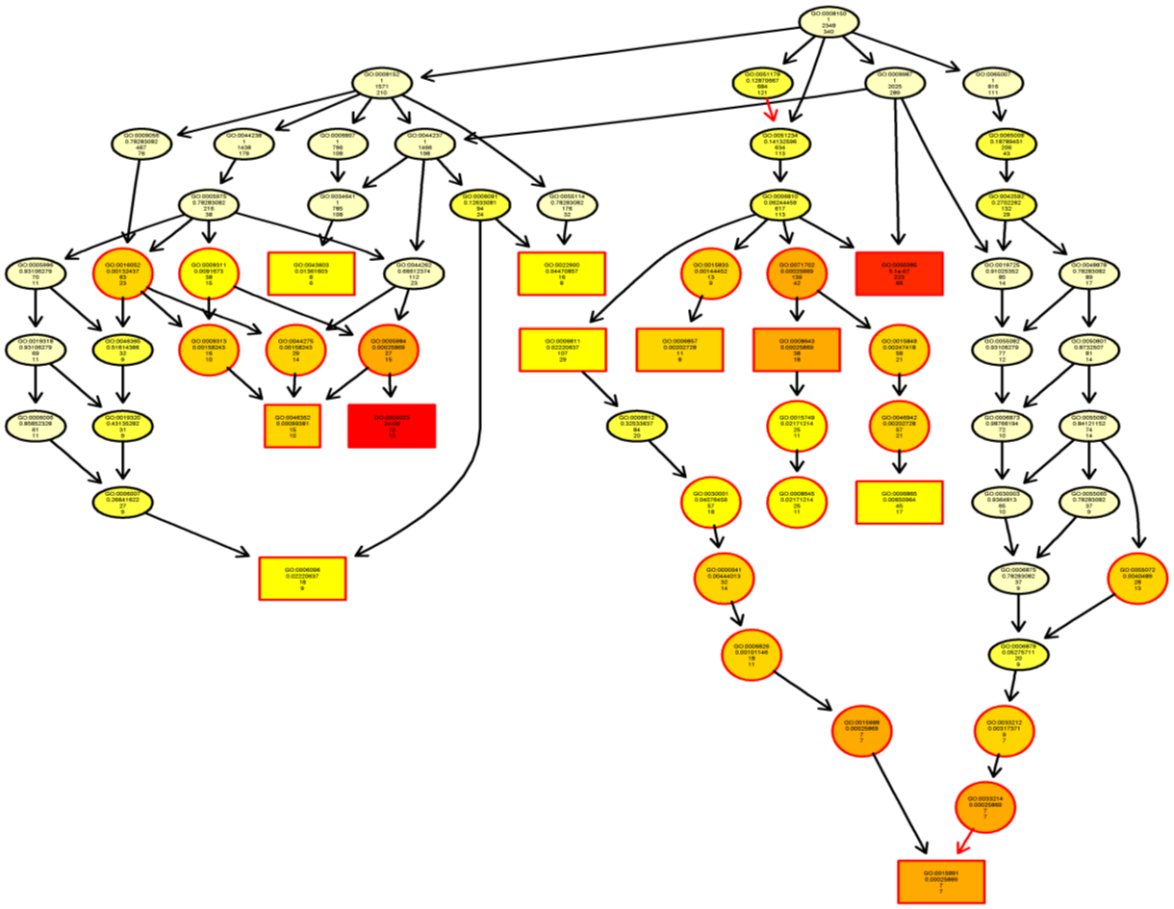
Based on discussion of previous section, 339 long non-essential telomeric genes are consider as observed gene category and 2346 all long non-essential genes are consider as reference gene category. As the GO terms are divided into 3 broad categories like molecular function (MF), biological process (BP) and cellular component (CC). So, the over representation of GO terms can be conducted for these three categories separately. Explanation of some symbols and definitions related to GO term over-representation analysis in GO DAG plot are summarized in the following table 9.

Table 9: Different symbols and terms used in GO DAG plot

Term	Symbol	Definition
Local	Circle	Local represents terms removed after applying local redundancy
Global	Box	Global represents terms removed after applying global redundancy
Final	Rectangular	Final represents the remained terms with evidence that their significance should not be simply due to the overlapping genes
Adjusted p value	Different color shads	The different color shades represent the adjusted p values of the terms

4.3.2.1. Overrepresentation of GO term in biological process (BP)

Figure 21: Over-representation of GO terms in biological process category of 339 telomeric long non-essential genes considering adjusted p value 0.05



In figure 21, 339 telomeric long non-essential genes are consider as observed group and all 2246 long non-essential genes are consider as reference group to observe over represented GO terms in biological process category. Increasing color intensity (More red) refers as lower adjusted p value. Circle, box and rectangular shape represent the GO terms that are removed after local redundancy, global redundancy and final redundancy respectively. The red border on every types of shape indicates that the adjusted p value of that GO term is less than 0.05.

Table 10: Significant GO terms in biological process (BP) category of 339 telomeric long non-essential genes.

GO[10]	C[11]	E[12]	O[13]	pvalue[14]	adjstp[15]	Final result[16]
GO:0000023	13	1.88	13	1.00861541341146e-11	1.64908620092774e-08	Final
GO:0055085	223	32.29	66	6.2434668546274e-10	5.1040341536579e-07	Final
GO:0005984	27	3.91	15	6.41007532276028e-07	0.00025869302185004	Local

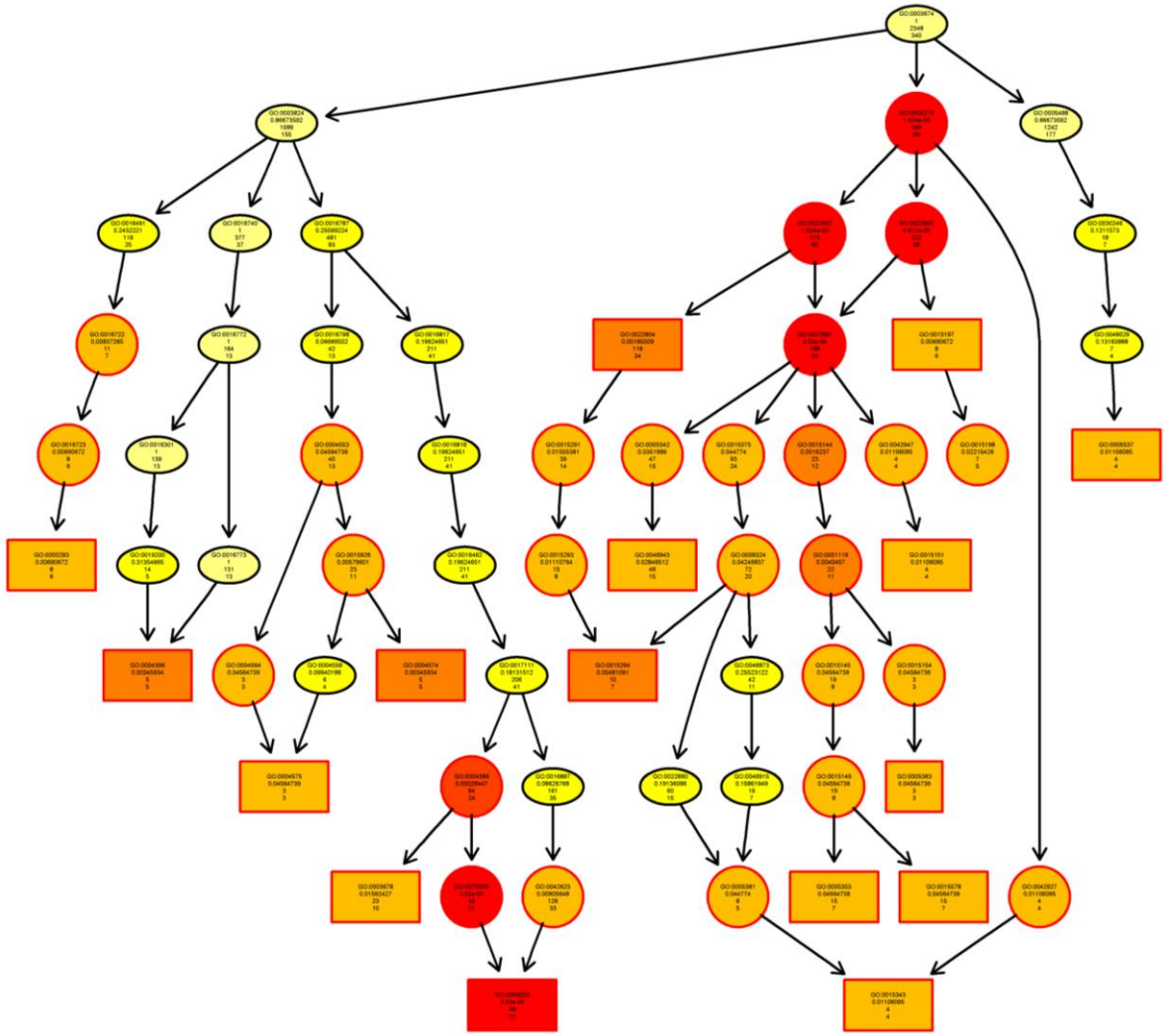
[10] GO = Significant GO terms.
[11] C = Genes in reference category.
[12] E = Expected number of genes in observed category.
[13] O = Genes in observed category.
[14] pvalue = P value by applying Hyper geometric distribution model.
[15] adjstp = Adjusted p value by applying Benjamini & Hochberg (BH) method.
[16] Final result = Result of after applying redundancy method.

GO:0008643	38	5.5	18	1.07511242575242e-06	0.00025869302185004	Final
GO:0015688	7	1.01	7	1.2657762537005e-06	0.00025869302185004	Local
GO:0015891	7	1.01	7	1.2657762537005e-06	0.00025869302185004	Final
GO:0033214	7	1.01	7	1.2657762537005e-06	0.00025869302185004	Local
GO:0071702	139	20.13	42	6.88777796908902e-07	0.00025869302185004	Local
GO:0046352	15	2.17	10	5.4705134582278e-06	0.000993809944911384	Global
GO:0006826	18	2.61	11	6.18628273274524e-06	0.00101145722680385	Local
GO:0016052	63	9.12	23	8.91013887527325e-06	0.00132437064191561	Local
GO:0015833	13	1.88	9	1.06020088829206e-05	0.00144452371029793	Local
GO:0009313	16	2.32	10	1.27245604373272e-05	0.00158242767487917	Local
GO:0044275	29	4.2	14	1.35498394179256e-05	0.00158242767487917	Local
GO:0006857	11	1.59	8	1.98388347434353e-05	0.00202728092534479	Final
GO:0046942	57	8.25	21	1.88823222110335e-05	0.00202728092534479	Local
GO:0015849	58	8.4	21	2.572543583923e-05	0.00247418162336124	Local
GO:0033212	9	1.3	7	3.49399082453239e-05	0.00317370833228359	Local
GO:0055072	28	4.05	13	4.7051421603661e-05	0.00404889864852557	Local
GO:0000041	32	4.63	14	5.43134786850086e-05	0.00444012688249945	Local
GO:0006865	45	6.52	17	8.36100227197267e-05	0.00650963748317872	Final
GO:0009311	38	5.5	15	0.000123352001786148	0.00916729649637963	Local
GO:0043603	8	1.16	6	0.000191540445263483	0.0136160273045998	Final
GO:0008645	25	3.62	11	0.000331989942407129	0.0217121422334262	Local
GO:0015749	25	3.62	11	0.000331989942407129	0.0217121422334262	Local
GO:0006096	18	2.61	9	0.000366710722029207	0.0222063715006575	Final
GO:0006811	107	15.49	29	0.000356079212699867	0.0222063715006575	Final
GO:0030001	57	8.25	18	0.000698109183987383	0.0407645898506918	Local
GO:0022900	16	2.32	8	0.000792996025432702	0.0447085690200851	Final

In the biological process (BP) category, for 339 telomeric long non-essential genes, there are 29 GO terms which are statistically significant (Adjusted p value is less than 0.05). From the result of redundancy (Final, global and local), biologically significant term can be identified. Go terms GO:0000023 and GO:0055085 can be classified as biologically significant terms as the adjusted p value is very small and the redundancy is final.

4.3.2.2. Overrepresentation of GO term in molecular function (MF)

Figure 22: Over-representation of GO terms in molecular function category of 339 telomeric long non-essential genes considering adjusted p value 0.05



In figure 22, 339 telomeric long non-essential genes are consider as observed group and all 2246 long non-essential genes are consider as reference group to observe over represented GO terms in molecular function category. Increasing color intensity (More red) refers as lower adjusted p value. Circle, box and rectangular shape represent the GO terms that are removed after local redundancy, global redundancy and final redundancy respectively. The red border on every types of shape indicates that the adjusted p value of that GO term is less than 0.05.

Table 11: Significant GO terms in molecular function (MF) category of 339 telomeric long non-essential genes.

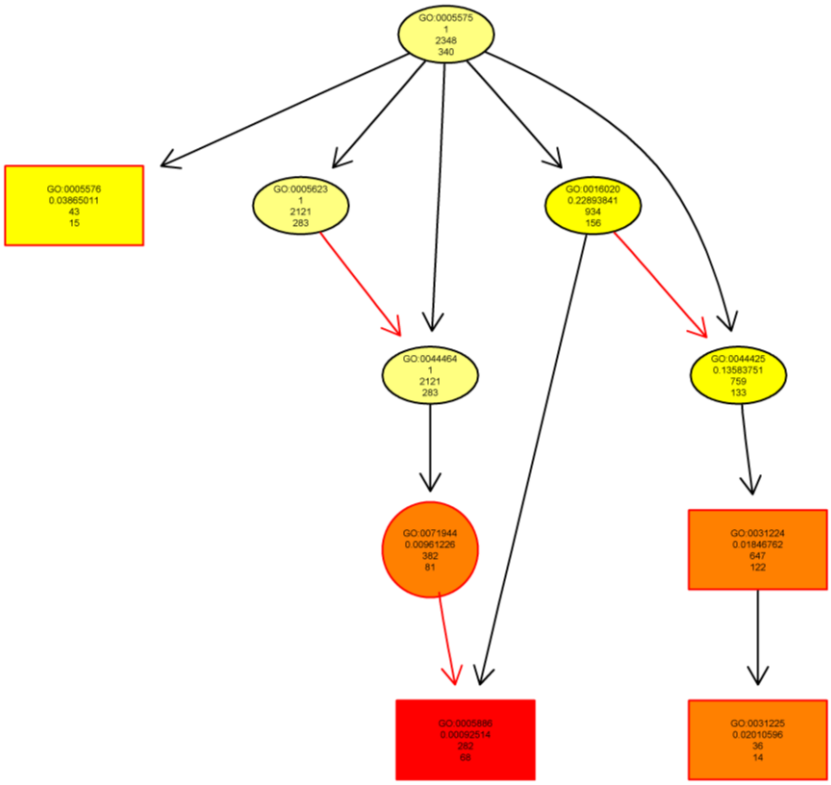
GO	C	E	O	pvalue	adjstp	Final Result
GO:0005215	260	37.65	69	4.15599523684307e-08	1.62398833926658e-05	Local
GO:0022857	215	31.13	60	5.29849376595948e-08	1.62398833926658e-05	Local
GO:0008026	46	6.66	21	2.87899833861793e-07	3.52965196314558e-05	Final
GO:0022891	188	27.22	53	2.55127469928063e-07	3.52965196314558e-05	Local
GO:0070035	46	6.66	21	2.87899833861793e-07	3.52965196314558e-05	Local
GO:0022892	222	32.15	59	4.71025810244718e-07	4.81231369466687e-05	Local

GO:0004386	64	9.27	24	3.24846762711495e-06	0.000284472950774495	Local
GO:0015144	23	3.33	12	2.11901649337376e-05	0.00162369638804764	Local
GO:0022804	118	17.09	34	2.72361726758596e-05	0.00185508598336688	Final
GO:0004396	5	0.72	5	6.20761607990072e-05	0.00345933514270831	Final
GO:0004574	5	0.72	5	6.20761607990072e-05	0.00345933514270831	Final
GO:0051119	22	3.19	11	7.91981232155514e-05	0.00404570412759442	Local
GO:0015294	10	1.45	7	0.000102025730430944	0.00481090559647451	Final
GO:0015926	23	3.33	11	0.000132372084973964	0.00579600629207428	Local
GO:0000293	8	1.16	6	0.000191540445263483	0.00690672311450089	Final
GO:0015197	8	1.16	6	0.000191540445263483	0.00690672311450089	Final
GO:0016723	8	1.16	6	0.000191540445263483	0.00690672311450089	Local
GO:0016722	11	1.59	7	0.000245852797884138	0.00837265361683203	Local
GO:0042623	126	18.25	33	0.000280706662767138	0.00905648338296082	Local
GO:0005537	4	0.58	4	0.000433055121764836	0.0110609495684102	Final
GO:0015151	4	0.58	4	0.000433055121764836	0.0110609495684102	Final
GO:0015343	4	0.58	4	0.000433055121764836	0.0110609495684102	Final
GO:0042927	4	0.58	4	0.000433055121764836	0.0110609495684102	Local
GO:0042947	4	0.58	4	0.000433055121764836	0.0110609495684102	Local
GO:0015293	15	2.17	8	0.00045300307084517	0.0111076352971236	Local
GO:0015291	39	5.65	14	0.000659705014873002	0.0155538143891212	Local
GO:0003678	23	3.33	10	0.000701395450102726	0.0159242744782582	Final
GO:0015198	7	1.01	5	0.00101194027352303	0.0221542638453435	Local
GO:0046943	46	6.66	15	0.00134663681479319	0.028465116119594	Final
GO:0005342	47	6.81	15	0.00172262177311366	0.0351989048972891	Local
GO:0008324	72	10.43	20	0.00214918364736139	0.0424983734139526	Local
GO:0005381	8	1.16	5	0.00237927332062227	0.0447739998996544	Local
GO:0015075	93	13.47	24	0.00241034583472854	0.0447739998996544	Local
GO:0004553	40	5.79	13	0.00289905030224624	0.0456473862800452	Local
GO:0004564	3	0.43	3	0.00301339543186485	0.0456473862800452	Local
GO:0004575	3	0.43	3	0.00301339543186485	0.0456473862800452	Final
GO:0005353	15	2.17	7	0.00283499575268487	0.0456473862800452	Final
GO:0005363	3	0.43	3	0.00301339543186485	0.0456473862800452	Global
GO:0015145	19	2.75	8	0.0031275533829721	0.0456473862800452	Local
GO:0015149	19	2.75	8	0.0031275533829721	0.0456473862800452	Local
GO:0015154	3	0.43	3	0.00301339543186485	0.0456473862800452	Local
GO:0015578	15	2.17	7	0.00283499575268487	0.0456473862800452	Final

In the molecular function (MF) category, for 339 telomeric long non-essential genes, there are 42 GO terms which are statistically significant (Adjusted p value is less than 0.05). From the result of redundancy (Final, global and local), biologically significant term can be identified.

4.3.2.3. Overrepresentation of GO term in cellular component (CC)

Figure 23: Over-representation of GO terms in cellular component (CC) category of 339 telomeric long non-essential genes considering adjusted p value 0.05



In figure 23, 339 telomeric long non-essential genes are consider as observed group and all 2246 long non-essential genes are consider as reference group to observe over represented GO terms in cellular component category. Increasing color intensity (More red color) refers as lower adjusted p value. Circle, box and rectangular shape represent the GO terms that are removed after local redundancy, global redundancy and final redundancy respectively. The red border on every types of shape indicate that the adjusted p value of that GO term is less than 0.05

Table 12: Significant GO terms in cellular component (CC) category of 339 telomeric long non-essential genes

GO	C	E	O	pvalue	adjstp	Final Result
GO:0005886	282	40.83	68	2.89105820328039e-06	0.000925138625049725	Final
GO:0071944	382	55.32	81	6.00765957367022e-05	0.00961225531787235	Local
GO:0031224	647	93.69	122	0.000173133966152705	0.0184676230562885	Final
GO:0031225	36	5.21	14	0.000251324507299366	0.0201059605839493	Final
GO:0005576	43	6.23	15	0.000603907989446539	0.0386501113245785	Final

In the cellular component (CC) category, for 339 telomeric long non-essential genes, there are 5 GO terms which are statistically significant (Adjusted p value is less than 0.05). From the result of redundancy (Final, global and local), biologically significant term can be identified.

4.3.2.4. Most significant GO term and cross validation

In 339 observed telomeric non-essential gene category both statistically and biologically significant GO terms have found in three different GO term broad categories named biological process (BP), molecular function (MF) and cellular component (CC). For this over representation study all 2246 non-essential genes are consider as reference category.

The group of genes under a significant GO term from a specific GO category (MF or BP or CC) can be treated as observed group to find the over representation of GO term in another GO category.

To conduct this study top 2 statistically significant GO term from each category has consider. The considered GO terms are summarized in the following table

Table 13: Top 2 statistically significant GO terms from each GO category

GO category		Term	C	E	O	pvalue	adjstp	Final Result
Biological process (BP)	**A**	GO:0000023	13	1.88	13	1.00861541341146e-11	1.64908620092774e-08	Final
	B	GO:0055085	223	32.29	66	6.2434668546274e-10	5.1040341536579e-07	Final
Molecular function (MF)	**C**	GO:0005215	260	37.65	69	4.15599523684307e-08	1.62398833926658e-05	Local
	D	GO:0008026	46	6.66	21	2.87899833861793e-07	3.52965196314558e-05	Final
Cellular component (CC)	**E**	GO:0005886	282	40.83	68	2.89105820328039e-06	0.000925138625049725	Final
	F	GO:0071944	382	55.32	81	6.00765957367022e-05	0.00961225531787235	Local

In table 13, top 2 statistically significant GO term from each GO category (BP, CC and MF) has stated. It is noted that here observed group is 339 telomeric non-essential genes and the reference group is all 2246 telomeric non-essential gene.

Case A: Considering GO term GO:0000023 (BP)

GO:0000023 (BP) to molecular function (MF)

By considering the 13 genes (Which sharing the GO term GO:0000023 significantly) as observed group, then the significant GO terms in molecular function (MF) category are describe in the following table.

Table 14: Significant GO term in MF category by considering genes that share GO term GO:0000023 (BP)

GO	C	E	O	pvalue	adjstp	Final Result
	11	0.42	7	5.42498690414561e-09	2.38699423782407e-07	Local
GO:0015926	23	3.33	11	0.000132372084973964	0.00579600629207428	Local
	13	0.5	7	2.73178492049553e-08	3.85112827294698e-07	Local
GO:0004553	40	5.79	13	0.00289905030224624	0.0456473862800452	Local
	5	0.19	5	3.50102570267907e-08	3.85112827294698e-07	Final
GO:0004574	5	0.72	5	6.20761607990072e-05	0.00345933514270831	Final
GO:0016798	13	0.5	7	2.73178492049553e-08	3.85112827294698e-07	Local
	4	0.15	4	1.30704959577788e-06	9.58503036903779e-06	Final
GO:0015151	4	0.58	4	0.000433055121764836	0.0110609495684102	Final
	4	0.15	4	1.30704959577788e-06	9.58503036903779e-06	Local
GO:0042947	4	0.58	4	0.000433055121764836	0.0110609495684102	Local
GO:0043167	57	2.18	9	2.44229912004235e-05	0.000134326451602329	Local
GO:0043169	57	2.18	9	2.44229912004235e-05	0.000134326451602329	Final
	3	0.11	3	4.40475713771038e-05	0.000193809314059257	Global
GO:0005363	3	0.43	3	0.00301339543186485	0.0456473862800452	Global
	3	0.11	3	4.40475713771038e-05	0.000193809314059257	Local

GO:0015154	3	0.43	3	0.00301339543186485	0.0456473862800452	Local
GO:0005352	2	0.08	2	0.00135346173867779	0.00496269304181856	Local
GO:0032450	2	0.08	2	0.00135346173867779	0.00496269304181856	Global
	3	0.11	2	0.00397229007327915	0.0124843402303059	Local
GO:0004564	3	0.11	2	0.00397229007327915	0.0124843402303059	Local
	3	0.11	2	0.00397229007327915	0.0124843402303059	Final
GO:0004575	3	0.43	3	0.00301339543186485	0.0456473862800452	Final
	11	0.42	3	0.00606714163463462	0.0177969487949282	Local
GO:0051119	22	3.19	11	7.91981232155514e-05	0.00404570412759442	Local
GO:0004558	4	0.15	2	0.00777231100983722	0.0204686588349148	Local
	12	0.46	3	0.00790834545894437	0.0204686588349148	Local
GO:0015144	23	3.33	12	2.11901649337376e-05	0.00162369638804764	Local
GO:0016787	83	3.17	7	0.0192345266863455	0.0470177318999557	Local

In table 14, 13 genes that share the GO term GO:0000023 from 339 observed group (Telomeric long non-essential) of genes are consider to study over representation of GO terms in molecular function category. Here the reference group is same as earlier study which was all 2346 long non-essential genes. The shaded GO term is common between present table and table 11, where 339 telomeric non-essential genes are in observed category. For every shaded GO term, the first row is the result by considering the 13 genes (Gene group those share GO term GO: GO:0000023) and the second row is the result by considering all 339 telomeric long non-essential genes.

GO:0000023 (BP) to cellular component (CC)

While considering the 13 genes (Which sharing significant GO term GO:0000023) as observed group, then there is no significant GO term in cellular component (CC) category. Here the reference gene category is 2346 long non-essential genes.

Case B: Considering GO term GO:0055085 (BP)

GO:0055085 (BP) to molecular function (MF)

Again by considering the 66 genes (Which sharing the GO term GO:0055085 that is significant) as observed group and 2346 long non-essential genes as reference group, then the significant GO terms in molecular function (MF) category are describe in the following table

Table 15 : Significant GO term in MF category by considering genes that share GO term GO:0055085 (BP)

GO	C	E	O	pvalue	adjstp	Final Result
GO:0005215	69	13.39	59	0	0	Local
	260	37.65	69	4.15599523684307e-08	1.62398833926658e-05	Local
GO:0022804	34	6.6	31	0	0	Local
	118	17.09	34	2.72361726758596e-05	0.00185508598336688	Final
GO:0022857	60	11.65	56	0	0	Final
	215	31.13	60	5.29849376595948e-08	1.62398833926658e-05	Local
GO:0022891	53	10.29	50	0	0	Local
	188	27.22	53	2.55127469928063e-07	3.52965196314558e-05	Local
GO:0022892	59	11.45	52	0	0	Local
	222	32.15	59	4.71025810244718e-07	4.81231369466687e-05	Local
GO:0015075	24	4.66	22	1.77635683940025e-15	4.9737991503207e-14	Local
	93	13.47	24	0.00241034583472854	0.0447739998996544	Local
GO:0008324	20	3.88	18	2.6336710590158e-12	6.32081054163792e-11	Local
	72	10.43	20	0.00214918364736139	0.0424983734139526	Local
GO:0005342	15	2.91	14	4.09649758559283e-10	7.64679549310662e-09	Local
	47	6.81	15	0.00172262177311366	0.0351989048972891	Local
GO:0046943	15	2.91	14	4.09649758559283e-10	7.64679549310662e-09	Local

	46	6.66	15	0.00134663681479319	0.028465116119594	Final
GO:0015144	12	2.33	12	1.2024866657967e-09	2.02017759853846e-08	Local
	23	3.33	12	2.11901649337376e-05	0.00162369638804764	Local
GO:0015291	14	2.72	13	2.35445485330388e-09	3.59589468504593e-08	Local
	39	5.65	14	0.000659705014873002	0.0155538143891212	Local
GO:0051119	11	2.14	11	7.19305626173394e-09	1.00702787664275e-07	Local
	22	3.19	11	7.91981232155514e-05	0.00404570412759442	Local
GO:0022890	15	2.91	13	1.49956872474988e-08	1.93790419813831e-07	Local
GO:0005275	10	1.94	10	4.23876532762435e-08	4.74741716693927e-07	Local
GO:0015171	10	1.94	10	4.23876532762435e-08	4.74741716693927e-07	Local
GO:0015145	8	1.55	8	1.40897277345342e-06	1.3923966231775e-05	Local
	19	2.75	8	0.0031275533829721	0.0456473862800452	Local
GO:0015149	8	1.55	8	1.40897277345342e-06	1.3923966231775e-05	Local
	19	2.75	8	0.0031275533829721	0.0456473862800452	Local
GO:0005353	7	1.36	7	7.95233785688687e-06	6.67996379978497e-05	Local
	15	2.17	7	0.00283499575268487	0.0456473862800452	Final
GO:0005355	7	1.36	7	7.95233785688687e-06	6.67996379978497e-05	Local
GO:0015578	7	1.36	7	7.95233785688687e-06	6.67996379978497e-05	Local
	15	2.17	7	0.00283499575268487	0.0456473862800452	Final
GO:0046873	11	2.14	9	9.66533139179493e-06	7.73226511343594e-05	Local
GO:0016820	9	1.75	8	1.0711588193324e-05	7.82411659338449e-05	Local
GO:0042626	9	1.75	8	1.0711588193324e-05	7.82411659338449e-05	Local
GO:0015399	10	1.94	8	4.52380626443372e-05	0.000292307481701871	Local
GO:0015405	10	1.94	8	4.52380626443372e-05	0.000292307481701871	Local
GO:0043492	10	1.94	8	4.52380626443372e-05	0.000292307481701871	Local
GO:0015293	8	1.55	7	5.37558934411431e-05	0.00033448111474489	Local
	15	2.17	8	0.00045300307084517	0.0111076352971236	Local
GO:0015294	7	1.36	6	0.000262162071349326	0.00151873199954092	Local
	10	1.45	7	0.000102025730430944	0.00481090559647451	Final
GO:0046915	7	1.36	6	0.000262162071349326	0.00151873199954092	Local
GO:0008509	6	1.16	5	0.00123732727851678	0.00598218937146641	Local
GO:0015077	6	1.16	5	0.00123732727851678	0.00598218937146641	Local
GO:0015078	4	0.78	4	0.00131750599252534	0.00598218937146641	Local
GO:0015151	4	0.78	4	0.00131750599252534	0.00598218937146641	Local
	4	0.58	4	0.000433055121764836	0.0110609495684102	Final
GO:0015197	6	1.16	5	0.00123732727851678	0.00598218937146641	Final
	8	1.16	6	0.000191540445263483	0.00690672311450089	Final
GO:0015238	4	0.78	4	0.00131750599252534	0.00598218937146641	Local
GO:0042947	4	0.78	4	0.00131750599252534	0.00598218937146641	Local
	4	0.58	4	0.000433055121764836	0.0110609495684102	Local
GO:0072349	4	0.78	4	0.00131750599252534	0.00598218937146641	Local
GO:0005381	5	0.97	4	0.00561508506338182	0.0230081534304426	Local
	8	1.16	5	0.00237927332062227	0.0447739998996544	Local
GO:0008028	5	0.97	4	0.00561508506338182	0.0230081534304426	Local
GO:0015103	5	0.97	4	0.00561508506338182	0.0230081534304426	Local
GO:0015198	5	0.97	4	0.00561508506338182	0.0230081534304426	Local
	7	1.01	5	0.00101194027352303	0.0221542638453435	Local
GO:0000099	3	0.58	3	0.00704761142033394	0.023679974372322	Local
GO:0005354	3	0.58	3	0.00704761142033394	0.023679974372322	Local
GO:0005363	3	0.58	3	0.00704761142033394	0.023679974372322	Local
	3	0.43	3	0.00301339543186485	0.0456473862800452	Global
GO:0015154	3	0.58	3	0.00704761142033394	0.023679974372322	Local
	3	0.43	3	0.00301339543186485	0.0456473862800452	Local
GO:0015174	3	0.58	3	0.00704761142033394	0.023679974372322	Local
GO:0015295	3	0.58	3	0.00704761142033394	0.023679974372322	Local
GO:0015298	3	0.58	3	0.00704761142033394	0.023679974372322	Local
GO:0015299	3	0.58	3	0.00704761142033394	0.023679974372322	Local
GO:0015665	3	0.58	3	0.00704761142033394	0.023679974372322	Local

In table 15, 66 genes that share the GO term GO:0055085 in observed group (Telomeric long non-essential) of 339 genes, are consider to study over representation of GO terms in

molecular function category. Here the reference group is same as earlier study which was all 2346 long non-essential genes. The shaded GO term is common between present table and table 11, where 339 telomeric non-essential genes are in observed category. For every shaded GO term, the first row is the result by considering the 66 genes (Gene group those share GO term GO:0055085) in observed group and the second row is the result by considering all 339 telomeric long non-essential genes in observed group.

GO:0055085 (BP) to cellular component (CC)

While considering the 66 genes (Which sharing GO term GO:0055085 that is significant) as observed group, then the significant GO term in cellular component (CC) category are given in following table. Here the reference gene category is 2346 long non-essential genes.

Table 16: Significant GO term in CC category by considering genes that share GO term GO:0055085 (BP)

GO	C	E	O	pvalue	adjstp	Final Result
GO:0005886	68	13.2	44	0	0	Global
	282	40.83	68	2.89105820328039e-06	0.000925138625049725	Final
GO:0016020	156	30.28	66	0	0	Local
GO:0016021	109	21.16	66	0	0	Final
GO:0031224	122	23.68	66	0	0	Local
	647	93.69	122	0.000173133966152705	0.0184676230562885	Final
GO:0044425	133	25.82	66	0	0	Local
GO:0071944	81	15.72	44	0	0	Local
	382	55.32	81	6.00765957367022e-05	0.00961225531787235	Local
GO:0005887	5	0.97	5	0.000243111224811243	0.002505104237338	Local
GO:0031226	5	0.97	5	0.000243111224811243	0.002505104237338	Local
GO:0044459	7	1.36	6	0.000262162071349326	0.002505104237338	Local
GO:0005773	25	4.85	12	0.000717164218116961	0.00616761227580586	Global
GO:0031090	52	10.09	18	0.00361086115167597	0.0282303690040121	Local
GO:0005774	19	3.69	9	0.0041981848963486	0.030086991757165	Local
GO:0044437	20	3.88	9	0.00644406727018354	0.0426299834796757	Local

In table 16, 66 genes that share the GO term GO:0055085 in observed group (Telomeric long non-essential) of 339 genes, are consider to study over representation of GO terms in cellular component category. Here the reference group is same as earlier study which was all 2346 long non-essential genes. The shaded GO term is common between present table and table 12, where 339 telomeric non-essential genes are in observed category. For every shaded GO term, the first row is the result by considering the 66 genes (Gene group those share GO term GO:0055085) in observed group and the second row is the result by considering all 339 telomeric long non-essential genes in observed group.

Case C: Considering GO term GO:0005215 (MF)

GO:0005215 (MF) to biological process (BP)

By considering the 69 genes (Which sharing the GO term GO:0005215 significantly) as observed group, then the significant GO terms in biological process (BP) category are describe in the following table.

Table 17: Significant GO term in BP category by considering genes that share GO term GO:0005215 (MF)

GO	C	E	O	pvalue	adjstp	Final Result
GO:0006810	113	22.93	69	0	0	Local
GO:0051179	121	24.56	69	0	0	Local
GO:0051234	113	22.93	69	0	0	Local
GO:0055085	66	13.39	59	0	0	Final
	223	32.29	66	6.2434668546274e-10	5.1040341536579e-07	Final
GO:0071702	42	8.52	36	0	0	Global
	139	20.13	42	6.88777796908902e-07	0.00025869302185004	Local
GO:0015849	21	4.26	19	1.14175335852451e-12	5.3662407850652e-11	Local
	58	8.4	21	2.572543583923e-05	0.00247418162336124	Local
GO:0046942	21	4.26	19	1.14175335852451e-12	5.3662407850652e-11	Local
	57	8.25	21	1.88823222110335e-05	0.00202728092534479	Local
GO:0006865	17	3.45	15	1.04752206908643e-09	4.30793450911794e-08	Local
	45	6.52	17	8.36100227197267e-05	0.00650963748317872	Final
GO:0006811	29	5.89	20	2.78427070288245e-09	1.01780562360925e-07	Local
	107	15.49	29	0.000356079212699867	0.0222063715006575	Final
GO:0008643	18	3.65	15	5.30191124337165e-09	1.74432879906927e-07	Local
	38	5.5	18	1.07511242575242e-06	0.00025869302185004	Final
GO:0046034	9	1.83	9	3.76872164031816e-07	1.12719038151334e-05	Final
GO:0006820	8	1.62	8	2.0511730894901e-06	5.62363288701869e-05	Final
GO:0006200	7	1.42	7	1.10167844969977e-05	0.000278809392270173	Local
GO:0015833	9	1.83	8	1.54455804933784e-05	0.000362971141594392	Final
	13	1.88	9	1.06020088829206e-05	0.00144452371029793	Local
GO:0003333	8	1.62	7	7.3776064348996e-05	0.00142778383357763	Local
GO:0006812	20	4.06	12	6.64022485533655e-05	0.00142778383357763	Final
GO:0006857	8	1.62	7	7.3776064348996e-05	0.00142778383357763	Local
	11	1.59	8	1.98388347434353e-05	0.00202728092534479	Final
GO:0008645	11	2.23	8	0.000198206953694213	0.00343210988238927	Local
	25	3.62	11	0.000331989942407129	0.0217121422334262	Local
GO:0015749	11	2.23	8	0.000198206953694213	0.00343210988238927	Local
	25	3.62	11	0.000331989942407129	0.0217121422334262	Local
GO:0015698	5	1.01	5	0.000305721234465262	0.00502911430695356	Local
GO:0030001	18	3.65	10	0.000719381749078485	0.0112703140688963	Local
	57	8.25	18	0.000698109183987383	0.0407645898506918	Local
GO:0015718	4	0.81	4	0.00158034361200488	0.0216638770145669	Local
GO:0072337	4	0.81	4	0.00158034361200488	0.0216638770145669	Local
GO:1901264	4	0.81	4	0.00158034361200488	0.0216638770145669	Local
GO:0009141	17	3.45	9	0.00219709332368212	0.0258158465532649	Local
GO:0009144	17	3.45	9	0.00219709332368212	0.0258158465532649	Local
GO:0009199	17	3.45	9	0.00219709332368212	0.0258158465532649	Local
GO:0009205	17	3.45	9	0.00219709332368212	0.0258158465532649	Local
GO:0006163	18	3.65	9	0.00367480489828587	0.0390003487592275	Local
GO:0009150	18	3.65	9	0.00367480489828587	0.0390003487592275	Local
GO:0009259	18	3.65	9	0.00367480489828587	0.0390003487592275	Local

In table 17, 69 genes that share the GO term GO:0005215 in observed group (Telomeric long non-essential) of 339 genes, are consider to study over representation of GO terms in biological process (BP) category. Here the reference group is same as earlier study which was all 2346 long non-essential genes. The shaded GO terms are common between present table and table 10, where 339 telomeric non-essential genes are in observed category. For every shaded GO term, the first row is the result by considering the 69 genes (Gene group those share GO term GO:0005215) in observed group and the second row is the result by considering all 339 telomeric long non-essential genes in observed group.

GO:0005215 (MF) to cellular component (CC)

By considering the 69 genes (Which sharing the GO term GO:0005215 significantly) as observed group, then the significant GO terms in cellular component (CC) category are describe in the following table.

Table 18: Significant GO term in CC category by considering genes that share GO term GO:0005215 (MF)

GO	C	E	O	pvalue	adjstp	Final Result
GO:0005886	68	13.8	45	0	0	Global
	282	40.83	68	2.89105820328039e-06	0.000925138625049725	Final
GO:0016020	156	31.66	69	0	0	Local
GO:0016021	109	22.12	68	0	0	Final
GO:0031224	122	24.76	68	0	0	Local
	647	93.69	122	0.000173133966152705	0.0184676230562885	Final
GO:0044425	133	26.99	69	0	0	Local
GO:0071944	81	16.44	45	0	0	Local
	382	55.32	81	6.00765957367022e-05	0.00961225531787235	Local
GO:0005887	5	1.01	5	0.000305721234465262	0.00380827118413091	Local
GO:0031226	5	1.01	5	0.000305721234465262	0.00380827118413091	Local
GO:0044459	7	1.42	6	0.000342744406571782	0.00380827118413091	Local
GO:0031090	52	10.55	20	0.000760080082230052	0.00760080082230052	Local
GO:0005773	25	5.07	12	0.00112741910235514	0.0102492645668649	Global
GO:0010008	6	1.22	5	0.00154229518282456	0.0118638090986505	Global
GO:0044440	6	1.22	5	0.00154229518282456	0.0118638090986505	Local
GO:0005768	10	2.03	6	0.00606918644122068	0.0404612429414712	Local
GO:0005774	19	3.86	9	0.00583905576345067	0.0404612429414712	Local

In table 18, 69 genes that share the GO term GO:0005215 in observed group (Telomeric long non-essential) of 339 genes, are consider to study over representation of GO terms in cellular component category. Here the reference group is same as earlier study which was all 2346 long non-essential genes. The shaded GO terms are common between present table and table 12, where 339 telomeric non-essential genes are in observed category. For every shaded GO term, the first row is the result by considering the 69 genes (Gene group those share GO term GO:0005215) in observed group and the second row is the result by considering all 339 telomeric long non-essential genes in observed group.

Case D: Considering GO term GO:0008026 (MF)

GO:0008026 (MF) to biological process (BP)

By considering the 21 genes (Which sharing the GO term GO:0008026 significantly) as observed group, then the significant GO terms in biological process (BP) category are describe in the following table.

Table 19: Significant GO term in BP category by considering genes that share GO term GO:0008026 (MF)

GO	C	E	O	pvalue	adjstp	Final Result
GO:0000722	7	0.43	7	1.18735288268113e-09	9.1426171966447e-08	Final
GO:0006312	10	0.62	7	1.27299286667437e-07	4.90102253669632e-06	Local
GO:0000723	11	0.68	7	3.37115470427918e-07	5.19157824458994e-06	Local
GO:0032200	11	0.68	7	3.37115470427918e-07	5.19157824458994e-06	Local
GO:0060249	11	0.68	7	3.37115470427918e-07	5.19157824458994e-06	Local
GO:0006310	16	0.99	8	4.8279784881089e-07	6.19590572640642e-06	Local
GO:0006259	33	2.04	8	0.000278847042016417	0.00306731746218059	Local
GO:0051276	36	2.22	8	0.000542918027335193	0.00522558601310123	Local

GO:0042592	29	1.79	7	0.00079838931932526	0.00682595530653383	Local
GO:0006996	50	3.09	8	0.00563251133579667	0.0433703372856344	Local

In table 19, 21 genes that share the GO term GO:0005215 in observed group (Telomeric long non-essential) of 339 genes, are consider to study over representation of GO terms in biological process (BP) category. Here the reference group is same as earlier study which was all 2346 long non-essential genes. There is no GO term is common between table 19 and table 10.

GO:0008026 (MF) to cellular component (CC)

While considering the 21 genes (Which sharing significant GO term GO:0008026) as observed group, then there is no significant GO term in cellular component (CC) category. Here the reference gene category is 2346 long non-essential genes.

Case E: Considering GO term GO:0005886 (CC)

GO:0005886 (CC) to biological process (BP)

By considering the 68 genes (Which sharing the GO term GO:0005886 significantly) as observed group, then the significant GO terms in biological process (BP) category are describe in the following table.

Table 20: Significant GO term in BP category by considering genes that share GO term GO:0005886 (CC)

GO	C	E	O	pvalue	adjstp	Final Result
GO:0006810	113	22.6	55	0	0	Local
GO:0051179	121	24.2	55	0	0	Local
GO:0051234	113	22.6	55	0	0	Local
GO:0055085	66	13.2	44	0	0	Final
	223	32.29	66	6.2434668546274e-10	5.1040341536579e-07	Final
GO:0006811	29	5.8	23	5.20028464734423e-13	5.89712279008836e-11	Final
	107	15.49	29	0.000356079212699867	0.0222063715006575	Final
GO:0071702	42	8.4	28	1.00186525742174e-12	9.46762668263544e-11	Local
	139	20.13	42	6.88777796908902e-07	0.00025869302185004	Local
GO:0008643	18	3.6	15	4.19207057955617e-09	3.3955771694405e-07	Final
	38	5.5	18	1.07511242575242e-06	0.00025869302185004	Final
GO:0006812	20	4	14	7.18786825393281e-07	5.09440162497488e-05	Local
GO:0030001	18	3.6	13	1.10933507313504e-06	6.98881096075075e-05	Local
	57	8.25	18	0.000698109183987383	0.0407645898506918	Local
GO:0006820	8	1.6	8	1.81335591975529e-06	0.000102817280650125	Local
GO:0000041	14	2.8	11	2.22258529358754e-06	0.000114564169224012	Local
	32	4.63	14	5.43134786850086e-05	0.00444012688249945	Local
GO:0048878	17	3.4	12	4.65496790236219e-06	0.000219947233386613	Local
GO:0055072	13	2.6	10	9.90343163487228e-06	0.000431941979767122	Global
	28	4.05	13	4.7051421603661e-05	0.00404889864852557	Local
GO:0006826	11	2.2	9	1.2703296183747e-05	0.000450173058511534	Local
	18	2.61	11	6.18628273274524e-06	0.00101145722680385	Local
GO:0008645	11	2.2	9	1.2703296183747e-05	0.000450173058511534	Local
	25	3.62	11	0.000331989942407129	0.0217121422334262	Local
GO:0015749	11	2.2	9	1.2703296183747e-05	0.000450173058511534	Local
	25	3.62	11	0.000331989942407129	0.0217121422334262	Local
GO:0015833	9	1.8	8	1.36984838754461e-05	0.000456884726904585	Final
	13	1.88	9	1.06020088829206e-05	0.00144452371029793	Local
GO:0050801	14	2.8	10	2.91055474880286e-05	0.00086857081187959	Local
GO:0055080	14	2.8	10	2.91055474880286e-05	0.00086857081187959	Local
GO:0006857	8	1.6	7	6.64996261052764e-05	0.00188526440008459	Local
	11	1.59	8	1.98388347434353e-05	0.00202728092534479	Final

GO:0015849	21	4.2	12	0.000110490523052897	0.0028476421168633	Local
	58	8.4	21	2.572543583923e-05	0.00247418162336124	Local
GO:0046942	21	4.2	12	0.000110490523052897	0.0028476421168633	Local
	57	8.25	21	1.88823222110335e-05	0.00202728092534479	Local
GO:0015698	5	1	5	0.000283567521822792	0.00699055586406622	Local
GO:0055082	12	2.4	8	0.000445428448881624	0.0105232471048284	Final
GO:0003333	8	1.6	6	0.00105609522724981	0.0239522397540257	Local
GO:0072337	4	0.8	4	0.0014887294895698	0.032465754637926	Global
GO:0006865	17	3.4	9	0.00196037609508393	0.0396976159254496	Local
	45	6.52	17	8.36100227197267e-05	0.00650963748317872	Final
GO:0019725	14	2.8	8	0.00190353147758171	0.0396976159254496	Local
GO:0006875	9	1.8	6	0.00266567843393062	0.04875611845286	Local
GO:0006879	9	1.8	6	0.00266567843393062	0.04875611845286	Local
GO:0055065	9	1.8	6	0.00266567843393062	0.04875611845286	Local
GO:0009987	289	57.8	65	0.00277163706269623	0.0491099442046488	Local

In table 20, 68 genes that share the GO term GO:0005886 in observed group (Telomeric long non-essential) of 339 genes, are consider to study over representation of GO terms in biological process (BP) category. Here the reference group is same as earlier study which was all 2346 long non-essential genes. The shaded GO terms are common between present table and table 10, where 339 telomeric non-essential genes are in observed category. For every shaded GO term, the first row is the result by considering the 68 genes (Gene group those share GO term GO:0005886) in observed group and the second row is the result by considering all 339 telomeric long non-essential genes in observed group.

GO:0005886 (CC) to molecular function (MF)

By considering the 68 genes (Which sharing the GO term GO:0005886 significantly) as observed group, then the significant GO terms in molecular function (MF) category are describe in the following table.

Table 21: Significant GO term in MF category by considering genes that share GO term GO:0005886 (CC)

GO	C	E	O	pvalue	adjstp	Final Result
GO:0005215	69	13.8	45	0	0	Local
	260	37.65	69	4.15599523684307e-08	1.62398833926658e-05	Local
GO:0022857	60	12	42	0	0	Local
	215	31.13	60	5.29849376595948e-08	1.62398833926658e-05	Local
GO:0022891	53	10.6	39	0	0	Local
	188	27.22	53	2.55127469928063e-07	3.52965196314558e-05	Local
GO:0022892	59	11.8	43	0	0	Local
	222	32.15	59	4.71025810244718e-07	4.81231369466687e-05	Local
GO:0015075	24	4.8	19	1.04980579784808e-10	4.1572309594784e-09	Local
	93	13.47	24	0.00241034583472854	0.0447739998996544	Local
GO:0015144	12	2.4	12	1.77874315276938e-09	5.86985240413895e-08	Final
	23	3.33	12	2.11901649337376e-05	0.00162369638804764	Local
GO:0051119	11	2.2	11	1.02667808699053e-08	2.90403230320178e-07	Local
	22	3.19	11	7.91981232155514e-05	0.00404570412759442	Local
GO:0022804	34	6.8	21	1.55431433279674e-08	3.84692797367193e-07	Local
	118	17.09	34	2.72361726758596e-05	0.00185508598336688	Final
GO:0008324	20	4	15	5.70016480683222e-08	1.25403625750309e-06	Local
	72	10.43	20	0.00214918364736139	0.0424983734139526	Local
GO:0022890	15	3	12	4.83364091974359e-07	9.57060902109231e-06	Final
GO:0015145	8	1.6	8	1.81335591975529e-06	2.99203726759623e-05	Local
	19	2.75	8	0.0031275533829721	0.0456473862800452	Local
GO:0015149	8	1.6	8	1.81335591975529e-06	2.99203726759623e-05	Local
	19	2.75	8	0.0031275533829721	0.0456473862800452	Local
GO:0005353	7	1.4	7	9.8991396929593e-06	0.000130668643947063	Local

	15	2.17	7	0.00283499575268487	0.0456473862800452	Final
GO:0005355	7	1.4	7	9.8991396929593e-06	0.000130668643947063	Local
GO:0015578	7	1.4	7	9.8991396929593e-06	0.000130668643947063	Local
	15	2.17	7	0.00283499575268487	0.0456473862800452	Final
GO:0046873	11	2.2	9	1.2703296183747e-05	0.000157203290273869	Global
GO:0015291	14	2.8	10	2.91055474880286e-05	0.000338994023684098	Final
	39	5.65	14	0.000659705014873002	0.0155538143891212	Local
GO:0008509	6	1.2	6	5.33276235069602e-05	0.000555729971283059	Final
GO:0015197	6	1.2	6	5.33276235069602e-05	0.000555729971283059	Final
	8	1.16	6	0.000191540445263483	0.00690672311450089	Final
GO:0005342	15	3	10	7.33057552678718e-05	0.000691168549668506	Local
	47	6.81	15	0.00172262177311366	0.0351989048972891	Local
GO:0046943	15	3	10	7.33057552678718e-05	0.000691168549668506	Local
	46	6.66	15	0.00134663681479319	0.028465116119594	Final
GO:0015103	5	1	5	0.000283567521822792	0.0024411464922136	Local
GO:0015198	5	1	5	0.000283567521822792	0.0024411464922136	Local
	7	1.01	5	0.00101194027352303	0.0221542638453435	Local
GO:0016722	7	1.4	6	0.00031389852639141	0.00248607632901997	Final
	11	1.59	7	0.000245852797884138	0.00837265361683203	Local
GO:0046915	7	1.4	6	0.00031389852639141	0.00248607632901997	Local
GO:0005275	10	2	7	0.000703543014306085	0.00515931543824462	Local
GO:0015171	10	2	7	0.000703543014306085	0.00515931543824462	Final
GO:0000293	6	1.2	5	0.00143476701340206	0.00979599547081406	Local
	8	1.16	6	0.000191540445263483	0.00690672311450089	Final
GO:0016723	6	1.2	5	0.00143476701340206	0.00979599547081406	Local
	8	1.16	6	0.000191540445263483	0.00690672311450089	Local
GO:0015294	7	1.4	5	0.00423693823092863	0.027963792324129	Local
	10	1.45	7	0.000102025730430944	0.00481090559647451	Final
GO:0005381	5	1	4	0.0063093773605577	0.0402986037867879	Local
	8	1.16	5	0.00237927332062227	0.0447739998996544	Local
GO:0005354	3	0.6	3	0.00771848981515422	0.0436645995257296	Local
GO:0005363	3	0.6	3	0.00771848981515422	0.0436645995257296	Local
	3	0.43	3	0.00301339543186485	0.0456473862800452	Global
GO:0015154	3	0.6	3	0.00771848981515422	0.0436645995257296	Local
	3	0.43	3	0.00301339543186485	0.0456473862800452	Local
GO:0015295	3	0.6	3	0.00771848981515422	0.0436645995257296	Local

In table 21, 68 genes that share the GO term GO:0005886 in observed group (Telomeric long non-essential) of 339 genes, are consider to study over representation of GO terms in molecular function (MF) category. Here the reference group is same as earlier study which was all 2346 long non-essential genes. The shaded GO terms are common between present table and table 11, where 339 telomeric non-essential genes are in observed category. For every shaded GO term, the first row is the result by considering the 68 genes (Gene group those share GO term GO:0005886) in observed group and the second row is the result by considering all 339 telomeric long non-essential genes in observed group

Case F: Considering GO term GO:0071944 (CC)

GO:0071944 (CC) to biological process (BP)

By considering the 81 genes (Which sharing the GO term GO:0071944 significantly) as observed group, then the significant GO terms in biological process (BP) category are describe in the following table.

Table 22: Significant GO term in BP category by considering genes that share GO term GO:0071944 (CC)

GO	C	E	O	pvalue	adjstp	Final Result
GO:0055085	66	15.72	44	0	0	Final
	223	32.29	66	6.2434668546274e-10	5.1040341536579e-07	Final
GO:0006810	113	26.92	57	2.33146835171283e-15	4.90385509976932e-13	Local
GO:0051234	113	26.92	57	2.33146835171283e-15	4.90385509976932e-13	Local
GO:0051179	121	28.83	58	2.62012633811537e-14	4.133249298377e-12	Local
GO:0006811	29	6.91	24	2.63167265757147e-12	3.3211708938552e-10	Final
	107	15.49	29	0.000356079212699867	0.0222063715006575	Final
GO:0071702	42	10.01	28	2.28388419287739e-10	2.40188487617606e-08	Local
	139	20.13	42	6.88777796908902e-07	0.00025869302185004	Local
GO:0008643	18	4.29	15	6.7167397310719e-08	6.05466110043767e-06	Final
	38	5.5	18	1.07511242575242e-06	0.00025869302185004	Final
GO:0000041	14	3.34	12	1.02047416716822e-06	6.43919199483147e-05	Local
	32	4.63	14	5.43134786850086e-05	0.00444012688249945	Local
GO:0006812	20	4.76	15	8.3388482818858e-07	6.43919199483147e-05	Local
GO:0030001	18	4.29	14	9.84087993094462e-07	6.43919199483147e-05	Local
	57	8.25	18	0.000698109183987383	0.0407645898506918	Local
GO:0006826	11	2.62	10	3.33927502971232e-06	0.000191552958522589	Local
	18	2.61	11	6.18628273274524e-06	0.00101145722680385	Local
GO:0048878	17	4.05	13	3.65843973071378e-06	0.000192372955840033	Local
GO:0055072	13	3.1	11	4.10264984462838e-06	0.00019913631168927	Global
	28	4.05	13	4.7051421603661e-05	0.00404889864852557	Local
GO:0006820	8	1.91	8	7.89030878056707e-06	0.000355627488609844	Local
GO:0050801	14	3.34	11	1.54039606619083e-05	0.000607493698604009	Local
GO:0055080	14	3.34	11	1.54039606619083e-05	0.000607493698604009	Local
GO:0008645	11	2.62	9	6.17132142508581e-05	0.00204952832591008	Local
	25	3.62	11	0.000331989942407129	0.0217121422334262	Local
GO:0015749	11	2.62	9	6.17132142508581e-05	0.00204952832591008	Local
	25	3.62	11	0.000331989942407129	0.0217121422334262	Local
GO:0015833	9	2.14	8	5.71334406880286e-05	0.00204952832591008	Final
	13	1.88	9	1.06020088829206e-05	0.00144452371029793	Local
GO:0055082	12	2.86	9	0.000198510500389326	0.00626300628728324	Local
GO:0006857	8	1.91	7	0.00022881895463589	0.00687546477977365	Local
	11	1.59	8	1.98388347434353e-05	0.00202728092534479	Final
GO:0000128	5	1.19	5	0.000696984196938644	0.0156620818190483	Final
GO:0001403	5	1.19	5	0.000696984196938644	0.0156620818190483	Final
GO:0015698	5	1.19	5	0.000696984196938644	0.0156620818190483	Local
GO:0015849	21	5	12	0.000719810416406341	0.0156620818190483	Local
	58	8.4	21	2.572543583923e-05	0.00247418162336124	Local
GO:0036267	5	1.19	5	0.000696984196938644	0.0156620818190483	Local
GO:0044182	5	1.19	5	0.000696984196938644	0.0156620818190483	Local
GO:0046942	21	5	12	0.000719810416406341	0.0156620818190483	Local
	57	8.25	21	1.88823222110335e-05	0.00202728092534479	Local
GO:0070783	5	1.19	5	0.000696984196938644	0.0156620818190483	Local
GO:0006875	9	2.14	7	0.000829718253453349	0.0156665762981886	Local
GO:0006879	9	2.14	7	0.000829718253453349	0.0156665762981886	Local
GO:0015688	7	1.67	6	0.000893814178660524	0.0156665762981886	Local
	7	1.01	7	1.2657762537005e-06	0.00025869302185004	Local
GO:0015891	7	1.67	6	0.000893814178660524	0.0156665762981886	Local
	7	1.01	7	1.2657762537005e-06	0.00025869302185004	Final
GO:0033212	7	1.67	6	0.000893814178660524	0.0156665762981886	Local
	9	1.3	7	3.49399082453239e-05	0.00317370833228359	Local
GO:0033214	7	1.67	6	0.000893814178660524	0.0156665762981886	Local
	7	1.01	7	1.2657762537005e-06	0.00025869302185004	Local
GO:0055065	9	2.14	7	0.000829718253453349	0.0156665762981886	Local
GO:0019725	14	3.34	9	0.00116824377961611	0.0199232925658855	Local
GO:0051704	12	2.86	8	0.00163769290798244	0.0271943217088663	Local
GO:0006873	10	2.38	7	0.00222939607637862	0.0351687231048727	Local
GO:0030003	10	2.38	7	0.00222939607637862	0.0351687231048727	Local
GO:0000501	4	0.95	4	0.00304138558664124	0.0426469845593472	Local
GO:0003333	8	1.91	6	0.00288879985073454	0.0426469845593472	Local

GO:0007155	4	0.95	4	0.00304138558664124	0.0426469845593472	Final
GO:0022610	4	0.95	4	0.00304138558664124	0.0426469845593472	Local
GO:0072337	4	0.95	4	0.00304138558664124	0.0426469845593472	Global

In table 22, 81 genes that share the GO term GO:0071944 in observed group (Telomeric long non-essential) of 339 genes, are consider to study over representation of GO terms in biological process (BP) category. Here the reference group is same as earlier study which was all 2346 long non-essential genes. The shaded GO terms are common between present table and table 10, where 339 telomeric non-essential genes are in observed category. For every shaded GO term, the first row is the result by considering the 81 genes (Gene group those share GO term GO:0071944) in observed group and the second row is the result by considering all 339 telomeric long non-essential genes in observed group.

GO:0071944 (CC) to molecular function (MF)

By considering the 81 genes (Which sharing the GO term GO:0071944 significantly) as observed group, then the significant GO terms in molecular function (MF) category are describe in the following table.

Table 23: Significant GO term in MF category by considering genes that share GO term GO:0071944 (CC)

GO	C	E	O	pvalue	adjstp	Final Result
GO:0005215	69	16.44	45	0	0	Local
	260	37.65	69	4.15599523684307e-08	1.62398833926658e-05	Local
GO:0022857	60	14.29	42	0	0	Local
	215	31.13	60	5.29849376595948e-08	1.62398833926658e-05	Local
GO:0022891	53	12.63	39	0	0	Local
	188	27.22	53	2.55127469928063e-07	3.52965196314558e-05	Local
GO:0022892	59	14.06	43	0	0	Local
	222	32.15	59	4.71025810244718e-07	4.81231369466687e-05	Local
GO:0015075	24	5.72	19	3.76388753497992e-09	1.61094386497141e-07	Final
	93	13.47	24	0.00241034583472854	0.0447739998996544	Local
GO:0015144	12	2.86	12	1.72754682781218e-08	6.16158368586344e-07	Final
	23	3.33	12	2.11901649337376e-05	0.00162369638804764	Local
GO:0051119	11	2.62	11	8.11947010737057e-08	2.48223800425329e-06	Local
	22	3.19	11	7.91981232155514e-05	0.00404570412759442	Local
GO:0022804	34	8.1	21	6.1283618890684e-07	1.6393368053258e-05	Local
	118	17.09	34	2.72361726758596e-05	0.00185508598336688	Final
GO:0008324	20	4.76	15	8.3388482818858e-07	1.98279281369285e-05	Local
	72	10.43	20	0.00214918364736139	0.0424983734139526	Local
GO:0022890	15	3.57	12	4.10895286240276e-06	8.79315912554191e-05	Local
GO:0015145	8	1.91	8	7.89030878056707e-06	0.000140710506586779	Local
	19	2.75	8	0.0031275533829721	0.0456473862800452	Local
GO:0015149	8	1.91	8	7.89030878056707e-06	0.000140710506586779	Local
	19	2.75	8	0.0031275533829721	0.0456473862800452	Local
GO:0005353	7	1.67	7	3.55063895124408e-05	0.000506557823710822	Local
	15	2.17	7	0.00283499575268487	0.0456473862800452	Final
GO:0005355	7	1.67	7	3.55063895124408e-05	0.000506557823710822	Local
GO:0015578	7	1.67	7	3.55063895124408e-05	0.000506557823710822	Local
	15	2.17	7	0.00283499575268487	0.0456473862800452	Final
GO:0046873	11	2.62	9	6.17132142508581e-05	0.000825414240605227	Local
GO:0008509	6	1.43	6	0.000158121787962151	0.00178095066441581	Local
GO:0015197	6	1.43	6	0.000158121787962151	0.00178095066441581	Final
	8	1.16	6	0.000191540445263483	0.00690672311450089	Final
GO:0015291	14	3.34	10	0.000158023319707712	0.00178095066441581	Final
	39	5.65	14	0.000659705014873002	0.0155538143891212	Local

GO:0005342	15	3.57	10	0.000381139494902261	0.00388399294805161	Local
	47	6.81	15	0.00172262177311366	0.0351989048972891	Local
GO:0046943	15	3.57	10	0.000381139494902261	0.00388399294805161	Local
	46	6.66	15	0.00134663681479319	0.028465116119594	Final
GO:0015103	5	1.19	5	0.000696984196938644	0.00648498339760304	Local
GO:0015198	5	1.19	5	0.000696984196938644	0.00648498339760304	Local
	7	1.01	5	0.00101194027352303	0.0221542638453435	Local
GO:0016722	7	1.67	6	0.000893814178660524	0.00765104936933409	Final
	11	1.59	7	0.000245852797884138	0.00837265361683203	Local
GO:0046915	7	1.67	6	0.000893814178660524	0.00765104936933409	Local
GO:0005275	10	2.38	7	0.00222939607637862	0.0176700281609268	Local
GO:0015171	10	2.38	7	0.00222939607637862	0.0176700281609268	Final
GO:0005537	4	0.95	4	0.00304138558664124	0.0224433281221112	Final
	4	0.58	4	0.000433055121764836	0.0110609495684102	Final
GO:0048029	4	0.95	4	0.00304138558664124	0.0224433281221112	Local
GO:0000293	6	1.43	5	0.00339129624182077	0.0234108837338595	Local
	8	1.16	6	0.000191540445263483	0.00690672311450089	Final
GO:0016723	6	1.43	5	0.00339129624182077	0.0234108837338595	Local
	8	1.16	6	0.000191540445263483	0.00690672311450089	Local

In table 23, 81 genes that share the GO term GO:0071944 in observed group (Telomeric long non-essential) of 339 genes, are consider to study over representation of GO terms in molecular function (MF) category. Here the reference group is same as earlier study which was all 2346 long non-essential genes. The shaded GO terms are common between present table and table 11, where 339 telomeric non-essential genes are in observed category. For every shaded GO term, the first row is the result by considering the 81 genes (Gene group those share GO term GO:0071944) in observed group and the second row is the result by considering all 339 telomeric long non-essential genes in observed group.

4.3.3. Progressive alignment approach

By applying the progressing alignment approach it is possible to observe the similarity of the gene's product (Proteins) in the observed gene category that share a specific GO term. Also chromosome-wise gene distribution of genes those share a specific GO term describes the cells backup system to reserve the same information in different chromosome in case of extreme condition like aneuploidy[17] (30).

To conduct this study top 2 statistically significant GO term from each category has consider. The considered GO terms are summarized in table 13, which is again given below.

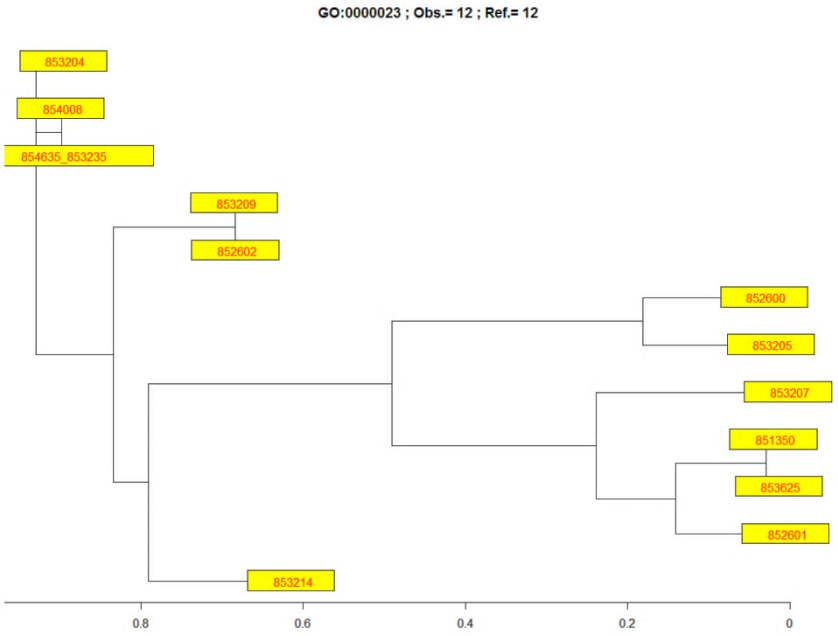
GO category		Term	C	E	O	pvalue	adjstp	Final Result
Biological process (BP)	**A**	GO:0000023	13	1.88	13	1.00861541341146e-11	1.64908620092774e-08	Final
	B	GO:0055085	223	32.29	66	6.2434668546274e-10	5.1040341536579e-07	Final
Molecular function (MF)	**C**	GO:0005215	260	37.65	69	4.15599523684307e-08	1.62398833926658e-05	Local
	D	GO:0008026	46	6.66	21	2.87899833861793e-07	3.52965196314558e-05	Final
Cellular component (CC)	**E**	GO:0005886	282	40.83	68	2.89105820328039e-06	0.000925138625049725	Final
	F	GO:0071944	382	55.32	81	6.00765957367022e-05	0.00961225531787235	Local

[17] Aneuploidy is an abnormal number of chromosomes, and is a type of chromosome abnormality. An extra or missing chromosome is a common cause of genetic disorders (birth defects). Some cancer cells also have abnormal numbers of chromosomes.

Case A: Considering GO term GO:0000023 (BP)

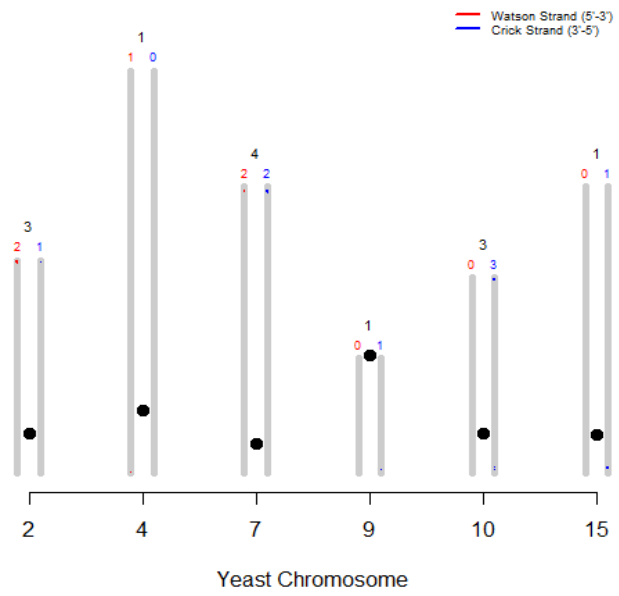
There are 13 genes that share the GO term GO:0000023 both in observed and reference category. In figure 43, the protein sequence similarity in the 13 genes in observed category is presented by guide tree using progressive alignment approach.

Figure 24: Guide tree based on protein sequence alignment of 13 genes that share GO term GO:0000023



In figure 24, in the reference category of 2346 long gene, there are 13 genes which share this GO term. But those 13 genes belong to the observed gene list of 339 telomeric genes. So GO:0000023 (Biological Process: maltose metabolic process) term has a very sharp tendency to reside in the telomeric region.

Figure 25: Gene distribution of 13 genes that share GO term GO:0000023

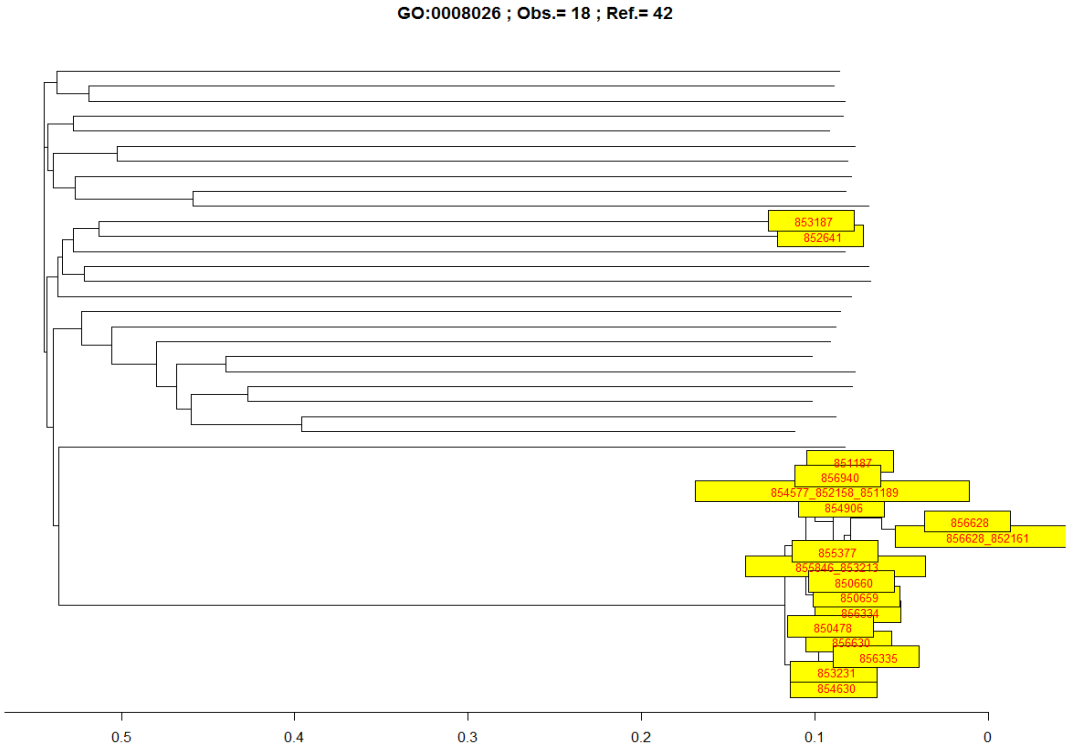


In figure 25, the chromosomal distribution of 13 gene that share the GO term GO:0000023 are presented.

Case D: Considering GO term GO:0008026 (MF)

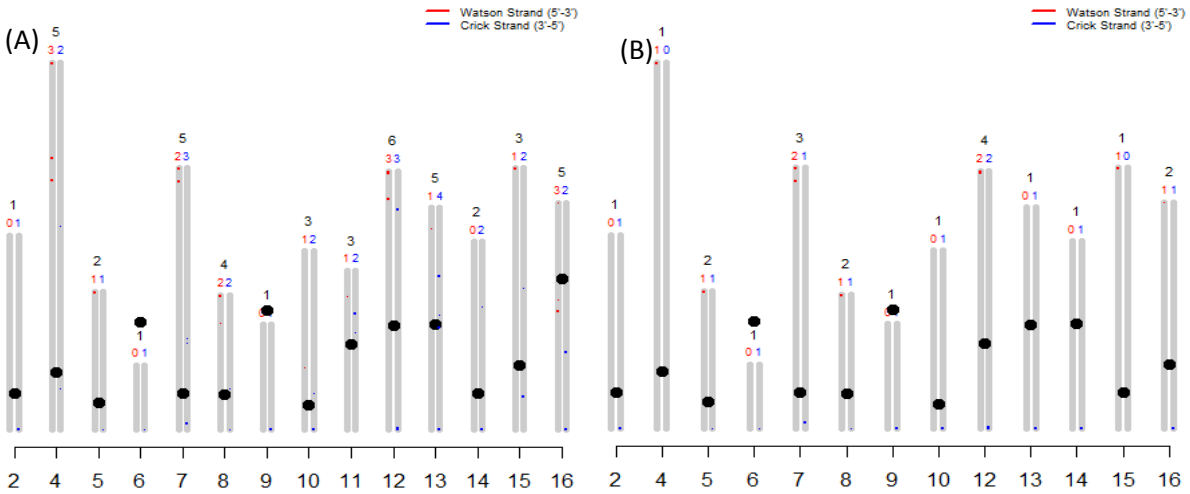
There are 46 genes and 21 genes those share the GO term GO:0000023 in reference and observed category respectively. In figure 26, the protein sequence in between the 21 genes in observed category is presented by guide tree using progressive alignment approach.

Figure 26: Guide tree based on protein sequence alignment of 21 genes that share GO term GO:0008026



In figure 26, in the reference category of 2346 long gene, there are 46 gene which share this GO term GO:0008026 (ATP-dependent helicase activity) are presented .The 21 genes in observed category are defined with label.

Figure 27: Chromosomal distribution of genes that share GO term GO:0008026 in observed and reference group



In figure 27, the chromosomal distribution of 46 gene in reference category (A) and 21 genes in observed category (B) those share the GO term GO:0000023 are presented.

4.4. Cell Cycle data

In the present analysis we have studied 5798 genes in 16 chromosomes where 4688 are non-essential and 1110 are essential genes. For cell cycle data, result was taken from Spellman et al. (1998) (31) where gene expression were measured (log-ratios, with Cy3-labeled common reference) for 6178 *Saccharomyces Cerevisiae* genes in 77 conditions. There are four main time courses: alpha (alpha factor arrest), cdc15, cdc28, and elu (elutriation), corresponding to different synchronization methods. For details on experimental procedures and analysis, refer to Spellman et al. (1998) (31) and the Yeast Cell Cycle Analysis Project website (32). Form cell cycle data it has observed that, there are 800 genes that are regulated. Among 5798 studied genes, there are 790 cell cycle regulate genes. Summary of 790 cell cycle regulated genes in essential and non-essential gene category are described in the following table 24.

Table 24 : Summary of Cell cycle regulated genes in essential and non-essential gene category

All cell cycle genes			**S**	**S/G2**	**G2/M**	**M/G1**	**G1**	**Total**
			71	118	190	113	298	790
Non-Essential	**All genome**		57	100	162	105	238	662
	Telomeric genes	**All genes**	15	6	26	23	47	117
		Long genes	10	2	18	14	22	66
		Short genes	5	4	8	9	25	51
Essential	**All genome**		14	18	28	8	60	128
	Telomeric genes	**All genes**	1	5	2	1	9	18
		Long genes	0	2	2	1	4	9
		Short genes	1	3	0	0	5	9

Chapter Five: Discussion

5.1. Random phenomenon in yeast genome

From the simulation on all yeast genome by Random Relocation (RR) test, it has observed that there is a big difference between the observed and expected situation for the number of essential yeast genes in the telomeric region. Not a single simulation (Out of 2000) is equal or less than the observed case. So it can be conclude that the telomeric effect of essential gene distribution is not a truly random phenomenon.

5.2. Base pair specific analysis

In the base pair specific analysis, it was observed that for essential genes including short and long by length, avoid the telomeric region which supports the results from RR test of previous section (5.1). It has also been observed that short non-essential genes has a flat tendency in the telomeric region where the long non-essential genes has decreasing tendency to reside in the telomeric region. The tendency of long non-essential genes to reside in the telomeric region is decreasing from centromeric to telomeric region. But the minimum percentage of long non-essential genes to reside in the telomeric region is about 29% per base pair position which increase up to about 48% till 55000 base pair position which is interestingly large.

Though the telomere is the hot location of genetic mutation and evaluation, still there are a large number of non-essential genes whose length is greater than 1179 bp. So there should be some special function about these genes, so that cell wants to keeps them in the telomeric region, though they are under high risk of genetic mutation and evolution. As by RR test it has been showed that distribution of essential genes in the telomeric region is a non-random phenomenon , so the distribution of non-essential genes in the telomeric region will also be a non-random phenomenon. In all genome, if the physical position of one group of genes is non-random, so the physical position of others genes in that location will also be non-random. But in the complex biological system it is hard to make a difference between effect and cause. So it can be suspected that, for better cellular workflow, cell want to have mutation in some long non-essential genes those are located in the telomeric region, so that cell can survive better with new or different environmental situation (33).

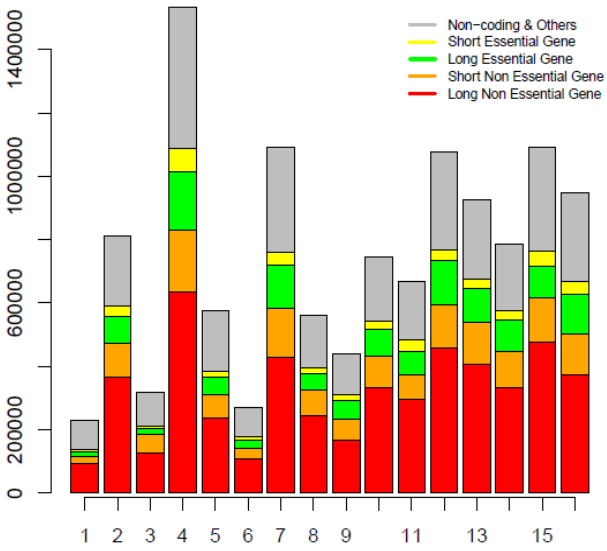
Median of all non-essential genes is used to define the short and long genes. It has also observed that genes whose length is in 3rd quartile (Q3), has a keen tendency to avoid the telomeric region for both essential and non-essential genes. On the other hand, genes whose length is in 1st (Q1) and 2nd (Q3)quantile, has a flat tendency in telomeric region. So it can be said that very longer genes has a higher tendency to avoid the telomeric region other than the shorter essential or non-essential genes.

For longer RNA genes the same telomeric region avoiding phenomenon has observed for 397 RNA genes with median length of 82bp. In 397 RNA genes, 204 genes are longer and rest 193 are shorter.

5.3. GO term specific analysis

Over representation of GO term in a specific list of genes can be identified by Hyper Geometric tests. For over-representation test there should be one observed group for which the over-representation is observed and another is reference group with which observed group is compared. According to table 8, there are different comparison group of genes to count the overrepresentation of GO term. But it has observed form figure 28, majority part of the genome is occupied by long non-essential genes. As our main concern is about the telomeric region, so it will be very informative to test the overrepresentation of GO term in non-essential gene list. From previous section, by base pair specific analysis, it has been observed that up to 55000 base pair position, long non-essential genes show an increasing tendency. So by taking 55000 base pair as a cutoff point, all observed genes are in the telomeric region within 55000 base pair position and reference gene group is all long non-essential gene in the genome.

Figure 28: Distribution of essential genes , non-essential genes and non-coding region by number of base pair in all 16 chromosome



In figure 28, different color represent the proportion of the number of base pair frequency from different genes categories

By using the GO term specific analysis and progressive alignment approach on telomeric long non-essential gene category, it has observed that some GO terms from 3 different GO categories are highly significantly over represented. For biological process category GO:0000041 (Transition metal ion transport), GO:0005984 (Disaccharide metabolic), GO:0006096 (Glycolysis), GO:0009313 (Oligosaccharide catabolic process), GO:0009311 (Oligosaccharide metabolic process), GO:0022900 (Electron transport chain) are significantly over represented. Again for cellular component category GO:0031225 (Anchored to membrane) and for molecular function category GO:0003678 (DNA helicase activity), GO:0004386 (Helicase activity), GO:0008026 (ATP-dependent helicase activity), GO:0070035 (Purine NTP-dependent helicase activity) are significantly over represented GO terms. In progressive alignment approach, the phylogenetic tree analysis for the protein sequences analysis of these long non-essential telomeric genes for those over represented GO terms, indicate that these proteins are very much evolutionary related. The same kind of analysis (Both GO term specific analysis and progressive alignment approach) has also been conducted for telomeric non-essential short genes, centromeric non-essential short & long genes, and centromeric essential short & long genes. But no specific statistically significantly over represented GO term for smalle groups of genes has observed for those groups.

Overrepresentation of a gene family with large sequence similarity may be either dependent on evolutionary forces that facilitate quick evolution in order to adapt to changes in threats or

nutrition conditions (Because of higher mutation rates and more rearrangements) or a result of a late evolution by duplications and specializations of one or a few number of genes that happened to reside in a telomere region. Even if pieces in telomeric regions are translocate and simultaneously duplicated to other chromosomes, it is probably more common that they are still telomeric after such translocation and chromosomes rearrangement processes are quite slow compared to e.g. amino acid mutation time scales. Studies of homologous gene families in more species may help to shed light on this ambiguity.

Similarly we find that many gene families are strategically spread out in different chromosomes, which may either be thought of as resulting from robustness properties (better backup), but on the other hand maybe is just the result of a high percentage of duplications having a simultaneous translocation to another chromosomes telomeric region. And such phenomena may moreover be sequence dependent? However this kind of processes maybe can be discriminated between by studying the flanking regions of the genes in the family.

There is a drawback of hyper geometric random allocation type of models, when one wishes to use models with covariates. One may think of the use of GO-classification in order to predict the localization of a gene in telomeric region or not, as a binary prediction problem, and try to use methods from binary regression and suitable predictive variable selection methods. However there is a basic problem of a clash between the typical hyper geometric framework used here and the typical independence assumptions used in such regression models. Moreover the huge dimensions and the complicated overlap structures of the GO-classification systems (Results in section 4.3.2.4) make it plausible that one cannot really trust that. Identified predictors using such techniques would be biologically (causally) involved and further biological understanding would be necessary to select biologically meaningful predictors in such an approach.

5.4. Cell cycle data and biological underrepresentation

Among 5798 genes in all 16 chromosome, there are 790 genes which are cell cycle regulated (31). Among these 790 cell cycle regulated genes, 245 genes are telomeric where 117 genes are non-essential and 128 genes are essential. In 117 telometic non-essential cell cycle regulated genes, in telomeric region, all cell cycle phase and transition phases (S, G2/M, M/G1, G1) have about 15%-20% of genes compare with all genome except S/G2 transition phase (5% of genes when compare with all genome) (Table: 24). That means, there is an underrepresentation of genes in S/G2 transition phase while cell cycle. This underrepresentation could be a group of biological interest, as less non-essential gene are involve with this transition phase. But it is hard to deduce that, this underrepresentation is the effect of over representation in other group of genes or this underrepresentation is a biological fact.

5.5. Future work

For the future, we suggest that comparison of our different finding corresponding telomeric statistical properties in *Saccharomyces cerevisiae* should be performed with other yeast species, like *Schizosaccharomyces pombe*, which is evolutionary distant enough to be genomically fairly reshuffled.

If the same analysis would have done with more well annotated organism like *Caenorhabditis elegans, Mus musculus* then it is possible to compare the results which might uncover a new analysis in comparative genomics. Again it is possible to make an R package which will contain the whole analysis pipeline discussed in this thesis work and apply the analysis pipeline on other organism easily.

Chapter Six: Conclusion

Prokaryote and Eukaryote are not same. In Eukaryotes the co-localization of genes in telomeric and centromeric region differs from organism to organism. In the thesis report different co-localization characteristics were observed on *Saccharomyces Cerevisiae*. For *Saccharomyces cerevisiae*, a clear signal was found that there is less code near the end of the telomere and genes those reside in this region are generally shorter in length and have weaker phenotypic effect when deleted (Non-essential genes).

In the present research work we have developed an analysis pipeline in which we can use one or more filters (e.g. 1. Essential property 2. Length property) and perform an over representation analysis on a family of gene classes using the traditional hyper geometric testing and also redundancy technique.

After applying this analysis pipeline on *Saccharomyces Cerevisiae* genome in the extreme telomeric region, it was observed that in telomeric region there are some small groups of genes with specific function related to metal ion transport, disaccharide and oligosaccharide metabolic and catabolic process. It has also observed that, in cell cycle data for non-essential genes S/G2 transition phase is underrepresented compared with G2/M, M/G1, G1 and S phase. So it is evident that in biological phenomenon only overrepresentation analysis with conventional hyper geometric model is not enough to understand the effect. Better statistical tool is necessary to explain the underrepresentation situation more clearly.

Appendix

1. Local regression

1.1. LOESS method

Assume that for $i = 1 \dots \dots n$, the ith measurement y_i of the response y and the corresponding measurement x_i of the vector x of p predictors can be modeled as

$$y_i = g(x_i) + \in_i \tag{1.1}$$

Where g is the regression function and $\in_i$ is a random error.

The idea of local regression is that, at a predictor x, the regression function $g(x)$ can be locally approximated by the value of a function in some specified parametric class. Such a local approximation is obtained by fiting a regression surface to the data points within a chosen neighborhood of the point x.

In the loess method, weighted least square is used to fit linear or quadratic functions of the predictors at the centers of neighborhoods. The radius of each neighborhood is chosen so that the neighborhood contains a specified percentage of the data points. The fraction of the data, called the smoothing parameter, in each local neighborhood controls the smoothness of the estimated surface. Data points in a given local neighborhood are weighted by a smooth decreasing function of their distance from the center of the neighborhood.

In a direct implementation, such fiting is done at each point at which the regression surface is to be estimated. A much faster computational procedure is to perform such local fiting at a selected sample of points in predictor space and then to blend these local polynomials to obtain a regression surface.

The underlining principle of local regression is that a smooth function can be well approximated by a low degree polynomial in the neighborhood of any point x. For an example, a local linear approximation is

$$g(x_i) \approx a_0 + a_1(x_i - x) \qquad for\ \ (x - h) \leq x_i \leq (x + h) \tag{1.2}$$

Where h is a fixed parameter known as bandwidth. A local quadratic approximation is

$$g(x_i) \approx a_0 + a_1(x_i - x) + \frac{a_2}{2}(x_i - x)^2 \tag{1.3}$$

The local approximation can be fitted by locally weighted least square. A weighted function $W(.)$ is chosen so that the most weight is given to those observations close to the fitted point x . One common choice for $W(.)$ is bisquare function

$$W(x) = \begin{cases} (1-x^2)^2 & -1 \le x \le 1 \\ 0 & x > 1 \; or \; x < -1 \end{cases} \quad (1.4)$$

In the case of local linear regression, coefficient estimate $\hat{a}_0$, $\hat{a}_1$ are chosen to minimize

$$\sum_{i=1}^{n} W\left(\frac{x_i - x}{h}\right)\left(Y_i - \left(a_o + a_1(x_i - x)\right)\right)^2 \quad (1.5)$$

The local linear regression estimate is define as

$$\hat{g}(x) = \hat{a}_0$$

Each local least square problem defines $\hat{g}(x)$ at one point x ; if x is changed the smoothing weights $W\left(\frac{x_i - x}{h}\right)$ change, and so the estimate $\hat{a}_0$ and $\hat{a}_1$ change. Since (1.5) is a weighted least squares problem, one can obtain the coefficient estimates by solving the normal equations

$$X^T W(Y - X\left(\begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix}\right) = 0 \quad (1.6)$$

Where, X is the design matrix

$$X = \begin{pmatrix} 1 & x_1 - x \\ \vdots & \vdots \\ 1 & x_n - x \end{pmatrix}$$

For local linear regression, W is a diagonal matrix with entries $W\left(\frac{x_i - x}{h}\right)$ and $Y = (Y_1 \dots \dots \dots Y_n)^T$. When $X^T W X$ in invertible, one has the explicit representation

$$\begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = (X^T W X)^{-1} X^T W Y \quad (1.7)$$

This shows that the local regression estimate is a linear estimate.

1.2. Pros and cons of LOESS method

LOESS method is very popular in modern regression methods for applications that fit the general framework of least squares regression but which have a complex deterministic structure. The main advantage of LOESS has over many other methods is that it does not require the specification of a function to fit a model to all of the data in the sample.

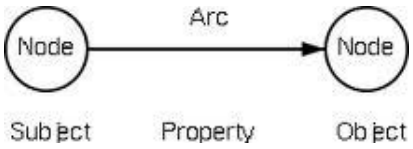
The main disadvantage of LOESS method is that, it makes less efficient use of data than other least squares methods. It requires fairly large, densely sampled data sets in order to produce good models. Another disadvantage of LOESS is, it does not produce a regression function that is easily represented by a mathematical formula. This can make it difficult to transfer the

results of an analysis to other people. In order to transfer the regression function to another person, they would need the data set and software for LOESS calculations.

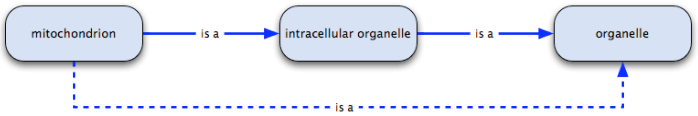
2. Gene Ontology (GO)[18]

2.1. Gene Ontology (GO) relations

The ontology of GO terms is composed as a graph, with terms as nodes in the graph and the relations between the terms as arcs. As each GO term is defined in a in a proper way, so the relation between the GO terms is also defined and can be categorized. The category of the relation is as follows

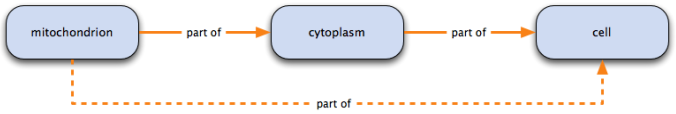


(a) "*is a*" relation



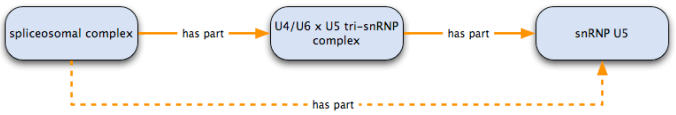
The "*is a*" relation in GO is very simple. If it is said that, A *is a* B, that means node A *is a subtype of* node B. As for example, mitotic cell cycle *is a* cell cycle, or lyase activity *is a* catalytic activity. In the figure mitochondrion *is an* intracellular organelle and intracellular organelle *is an* organelle, therefore mitochondrion *is an* organelle.

(b) "*part of*" relation



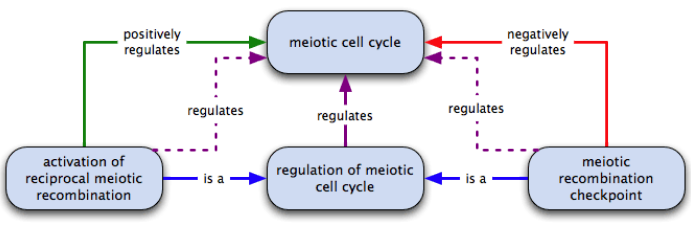
The relation *part of* is used to represent part-whole relationships in the Gene Ontology. For example, mitochondrion is *part of* cytoplasm and cytoplasm is *part of* cell, therefore mitochondrion is *part of* cell.

(c) "has part" relation



The logical complement to the *part of* relation is *has part*, which represents a part-whole relationship from the perspective of the parent. As for example, spliceosomal complex *has part* U4/U6 x U5 tri-snRNP complex and U4/U6 x U5 tri-snRNP complex *has part* snRNP U5, therefore spliceosomal complex *has part* snRNP U5.

(d) "regulates" relation



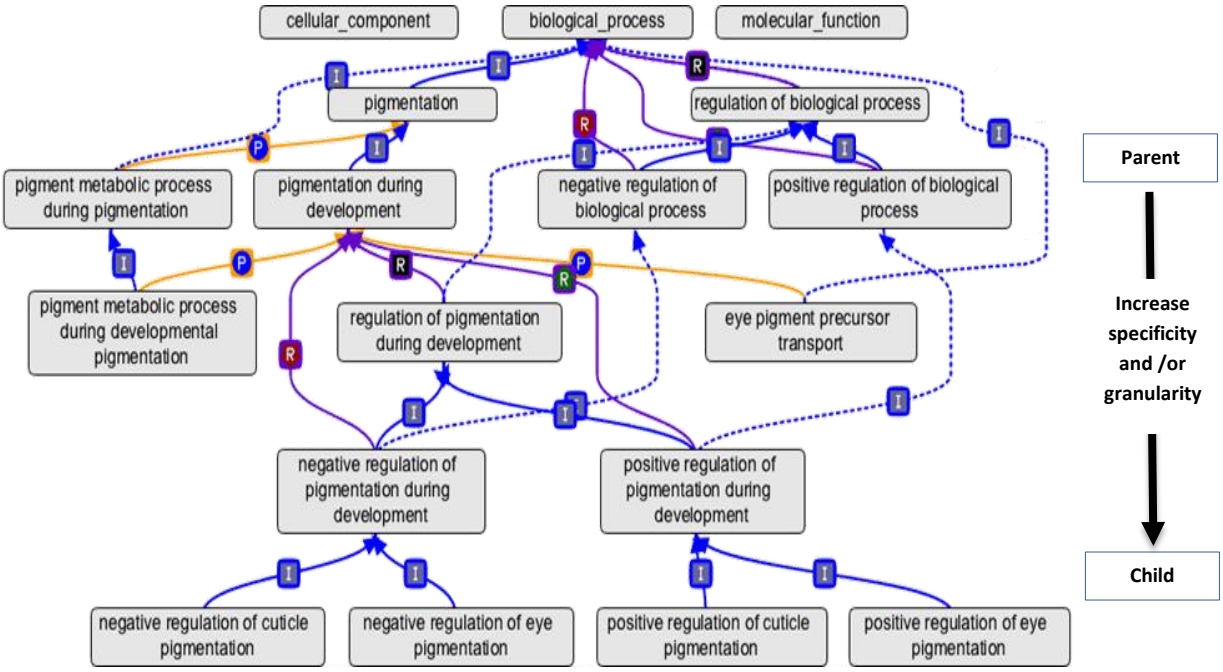
For the "regulates" relation in gene ontology refers, where one process directly affected another process. That means the first one regulate the second one. The *regulates* relation has two sub-relations, *positively regulates* and *negatively regulates*, to represent these more specific forms of regulation. As example in the figure illustrates the relationship between meiotic cell cycle, meiotic recombination checkpoint, which *negatively regulates* the meiotic cell cycle; the activation of reciprocal meiotic

[18] Source : http://www.geneontology.org/

recombination, which *positively regulates*; and regulation of meiotic cell cycle, representing anything that *regulates* the meiotic cell cycle.

2.2. Gene Ontology system

The structure of GO can be described in terms of a graph, where each GO term is a node, and the relationships between the terms are arcs between the nodes. The relationships used in GO are directed, for example, a mitochondrion is an organelle, but an organelle is not a mitochondrion and the graph is acyclic, meaning that cycles are not allowed in the graph. The ontologies resemble a hierarchy, as child terms are more specialized and parent terms are less specialized, but unlike a hierarchy, a term may have more than one parent term. For example, the biological process term hexose biosynthetic process has two parents, hexose metabolic process and monosaccharide biosynthetic process. This is because biosynthetic process is a type of metabolic process and a hexose is a type of monosaccharide. The following diagram is a screenshot from the ontology editing software OBO-Edit, showing a small set of terms from the ontology.



A set of terms under the biological process node pigmentation

In the diagram, relations between the terms are represented by the colored arrows; the letter in the box midway along each arrow is the relationship type. Note that the terms get more specialized going down the graph, with the most general terms—the root nodes, cellular component, biological process and molecular function—at the top of the graph. Terms may have more than one parent, and they may be connected to parent terms via different relations.

3. R Code

```
##############################################
##          Data Construction               ##
##                                          ##
##############################################
##              Function                    ##
##                                          ##
##############################################
## Obtain GO term annotated for every gene  ##
## in MF,BP and CC.                         ##
##############################################
GoTerm<-function(gid,ontology,organism=yeast){
        ###### Organusm List
        yeast<-"org.Sc.sgd.db"
        human<-"org.Hs.eg.db"

        ###### Required Package
        require(organism, character.only = TRUE) ||
stop(paste("package",organism, "is required", sep = " "))

        ###### Database quary
        .sql <- paste("select distinct t1.gene_id,t2.go_id", " from genes
as t1 inner join",
        paste("go", tolower(ontology), "all", sep = "_"), " as t2 on
t1._id=t2._id",
        seq = "")
        organism <- strsplit(organism, ".db")
        organism <- organism[[1]]
        conn <- get(paste(organism, "_dbconn", sep = ""))()
        allAnn <- dbGetQuery(conn, .sql)
        gg<-allAnn$go_id[which(allAnn$gene_id%in%gid)]
        return(gg)
        }


##############################################
##          GO term specific analysis       ##
##                                          ##
##############################################

############################################## 1
##  Over-representation test of significant ##
##  GO term and vizualization of GO map     ##
##  based on observed and reference gene    ##
##  list.                                   ##
##############################################
rm(list=ls())                   ## Clear memory

##############################################
##              Functions                   ##
##  Note : Most of the code of functuons#   ##
##  was obtained "GOFunction" package of    ##
##  Bioconductor                            ##
##                                          ##
##                                          ##
##############################################
##              Local Redundancy            ##
##############################################
```

```
localRedundancy <-  function(sigTerm, generalAnn, sigTermRelation, annRef,
annInterest, ppth, pcth){
     annRef <- unique(annRef[,1])
     allRefnum <- length(annRef)
     annInterest <- unique(annInterest[,1])
     allInterestnum <- length(annInterest)

     sigTermRelationRe <- sigTermRelation
     sigLabel <- array(0,dim=c(nrow(sigTerm),1))
     sigTerm$Label <- sigLabel
     sigTerm$SeLabel <- sigLabel
     La <- sigTerm[,1]
     noRelationTerm <-
setdiff(La,union(sigTermRelation[,1],sigTermRelation[,2]))
     sigTerm[sigTerm[,1] %in% noRelationTerm,7] <- 1
     La <- setdiff(La,noRelationTerm)

     while (length(La) > 0) {

          sigTermRelationRe <- sigTermRelationRe[sigTermRelationRe[,2] %in%
La,]
          if (nrow(sigTermRelationRe) == 0)
                 leafnode <- La
          else
                 leafnode <-
setdiff(sigTermRelationRe[,2],sigTermRelationRe[,1])

          La <- setdiff(La,leafnode)


          for (j in c(1:length(leafnode))){
               node <- leafnode[j];
               genes <- generalAnn[generalAnn[,2]==node,1]
               genes <- intersect(genes,annRef)
               sgenes <- intersect(genes,annInterest)
               childnode <- sigTermRelation[sigTermRelation[,1]==node,2]
               if (length(childnode)==0) {
                   sigTerm[sigTerm[,1]==node,7] <- 1
               }
               else {
                   activeChild <- sigTerm[(sigTerm[,1] %in% childnode) &
sigTerm[,7]==1,1]
                   if (length(activeChild)==0) {
                       sigTerm[sigTerm[,1]==node,7] <- 1
                   }
                   else {
                       allcgenes <-
generalAnn[generalAnn[,2]==activeChild[1],1]
                       if (length(activeChild)>1) {
                           for (k in c(2:length(activeChild))) {
                               cgenes <-
generalAnn[generalAnn[,2]==activeChild[k],1]
                               allcgenes <- union(allcgenes,cgenes)
                           }
                       }
                       allcgenes <- intersect(allcgenes,annRef)
                       allcsiggenes <- intersect(allcgenes,annInterest)
                       extragenes <- setdiff(genes,allcgenes)
                       extrasiggenes <- intersect(extragenes,annInterest)
                       if (length(extragenes)!=0) {
                             fp <- length(extrasiggenes)/length(extragenes)
```

```
                              fc <- length(allcsiggenes)/length(allcgenes)
                              p <- 1-phyper(length(extrasiggenes)-
1,allInterestnum,allRefnum-allInterestnum,length(extragenes),lower.tail =
TRUE,log.p= FALSE)
                              pc <- 1-phyper(length(allcsiggenes)-
1,length(sgenes),length(genes)-length(sgenes),length(allcgenes),lower.tail
= TRUE,log.p= FALSE)
                              if (fp>=fc | p<=ppth) {
                                   sigTerm[sigTerm[,1]==node,7] <- 1
                                   if (pc>pcth)
                                       sigTerm[sigTerm[,1] %in%
activeChild,8] <- sigTerm[sigTerm[,1] %in% activeChild,8]+1
                              }
                         }
                    }
               }
          }
       }
       sigTerm[,7] <- sigTerm[,7]+sigTerm[,8]
       sigTerm[sigTerm[,7]>1,7] <- 0;
       sigTermRedun <- sigTerm[sigTerm[,7]==1,c(1:6)]
       return(sigTermRedun)
    }
##                                          ##
##                                          ##
##############################################
##              Global Redundancy           ##
##############################################
globalRedundancy <- function(generalAnn, sigTermRelation, annRef,
annInterest, sigTermRedun, poth, peth){
     annRef <- unique(annRef[,1])
     allRefnum <- length(annRef)
     annInterest <- unique(annInterest[,1])
     allInterestnum <- length(annInterest)
     sigTermRedun$overlap = array(0,dim=c(nrow(sigTermRedun),1));
     sigTermenv <- new.env(hash=T,parent=emptyenv())
     assign("sigTerm",sigTermRedun,envir=sigTermenv)

     calculateEachTerm <- function (term1) {
         sigTermRedun <- get("sigTerm",sigTermenv)
          gene1 <- generalAnn[generalAnn[,2]==term1,1]
          gene1 <- intersect(gene1, annRef)
          siggene1 <- intersect(gene1, annInterest)
          extrterm <- setdiff(sigTermRedun[,1], term1);
          calculateExtraTerm <- function(term2) {
               gene2 <- generalAnn[generalAnn[,2]==term2,1];
               gene2 <- intersect(gene2, annRef);
               siggene2 <- intersect(gene2, annInterest)
               po <- sigTermRelation[(sigTermRelation[,1]==term1 &
sigTermRelation[,2]==term2) | (sigTermRelation[,1]==term2 &
sigTermRelation[,2]==term1),]
               if (nrow(po)==0){
                    refov <- intersect(gene1,gene2);
                    if (length(refov)>0) {
                         sigov <- intersect(siggene1,siggene2)
                         extra1 <- setdiff(gene1,refov)
                         extrasig1 <- intersect(extra1, annInterest)
                         extra2 <- setdiff(gene2,refov)
                         extrasig2 <- intersect(extra2, annInterest)
                         if(length(extra2)==0){
                               return(0)
```

```
                        }
                        else{
                              pex2 <- 1-phyper(length(extrasig2)-
1,allInterestnum,allRefnum-allInterestnum,length(extra2),lower.tail =
TRUE,log.p= FALSE)
                              po2 <- 1-phyper(length(sigov)-
1,length(siggene2),length(gene2)-length(siggene2),length(refov),lower.tail
= TRUE,log.p= FALSE)
                              if(length(extra1)==0){
                                   if ((po2>poth) | (po2<=poth &
pex2<=peth)){

sigTermRedun[sigTermRedun[,1]==term1,7] <- 1

assign("sigTerm",sigTermRedun,envir=sigTermenv)
                                   }
                              }
                              else{
                                   pex1 <- 1-phyper(length(extrasig1)-
1,allInterestnum,allRefnum-allInterestnum,length(extra1),lower.tail =
TRUE,log.p= FALSE)
                                   po1 <- 1-phyper(length(sigov)-
1,length(siggene1),length(gene1)-length(siggene1),length(refov),lower.tail
= TRUE,log.p= FALSE)
                                   if((po1<=poth) & (pex1>peth)){
                                       if ((po2>poth) | (po2<=poth &
pex2<=peth)){

sigTermRedun[sigTermRedun[,1]==term1,7] <- 1

assign("sigTerm",sigTermRedun,envir=sigTermenv)
                                       }
                                   }
                              }
                        }
                   }
              }
         }
         lapply(extrterm,calculateExtraTerm)
    }
    lapply(sigTermRedun[,1],calculateEachTerm)
    sigTermRedun <- get("sigTerm",sigTermenv)
    sigTermRedun <- sigTermRedun[sigTermRedun[,7]==0,c(1:6)]
    return(sigTermRedun);
}
##                                          ##
##                                          ##
##############################################
##              GO DAG node and edge        ##
##############################################
createGODAG <-      function(sigNodes, ontology = "BP"){

  nodeLabel <- new.env(hash = T, parent = emptyenv())

  isNodeInDAG <- function(node) {
    return(exists(node, envir = nodeLabel, mode = 'logical', inherits =
FALSE))
  }
  setNodeInDAG <- function(node) {
    assign(node, TRUE, envir = nodeLabel)
  }
```

```
  GOParents <- get(paste('GO', ontology, 'PARENTS', sep = ''))

  ROOT <- as.character(revmap(GOParents)$all)

  adjList <- as.list(GOParents)
  edgeEnv <- new.env(hash = T, parent = emptyenv())
  envAddEdge <- function(u, v, type) {
    assign(v, switch(type, is_a = 0, part_of = 1, -1), envir = get(u, envir
= edgeEnv))
  }

  createNodesRelationship <- function(node) {
    if(isNodeInDAG(node))
      return(1)

    ## we put the node in the graph and we get his parents
    setNodeInDAG(node)    # we visit the node
    assign(node, new.env(hash = T, parent = emptyenv()), envir = edgeEnv) #
adj list

    if(node == ROOT)
      return(2)

    adjNodes <- adjList[[node]]

    if(length(adjNodes) == 0)
      cat('\n There are no adj nodes for node: ', node, '\n')

    for(i in 1:length(adjNodes)) {
      x <- as.character(adjNodes[i])
      envAddEdge(node, x, names(adjNodes[i]))
      createNodesRelationship(x)
    }

    return(0)
  }

  ## we start from the most specific nodes
  lapply(sigNodes, createNodesRelationship)

  .graphNodes <- ls(edgeEnv)
  .edgeList <- eapply(edgeEnv,
                      function(adjEnv) {
                        aux <- as.list(adjEnv)
                        return(list(edges = match(names(aux), .graphNodes),
                                    weights = as.numeric(aux)))
                      })

  ## now we can build the graphNEL object
  GOgraph.topo <- new('graphNEL',
                      nodes = .graphNodes,
                      edgeL = .edgeList,
                      edgemode = 'directed')

  require("SparseM") || stop("package SparseM is required")

  GOgraph.topo <- sparseM2Graph(t(graph2SparseM(GOgraph.topo, TRUE)),
                        .graphNodes, edgemode = "directed")
  return(GOgraph.topo)
  }
```

```
##                                            ##
##                                            ##
################################################
##              Vizualive GO nodes            ##
################################################
showSigNodes <-     function(DAG, sigTerm, sigTerm_Local, sigTerm_Global,
dagTermInfo, bmpSize, filename,l_ref,l_obs,sigLav){

  require('Rgraphviz') || stop('package Rgraphviz is required')

  graphAttrs <- getDefaultAttrs(layoutType = 'dot')
  graphAttrs$cluster <- NULL
  graphAttrs$node$shape <- 'ellipse'
  graphAttrs$node$fontsize <- '60'

  nodeAttrs <- list()
  edgeAttrs <- list()

  allTerm <- as.character(dagTermInfo[,1])
  allAp <- round(as.numeric(dagTermInfo[,6]),8)
  C<- as.numeric(dagTermInfo[,3])
  O<-as.numeric(dagTermInfo[,4])

  nodeAttrs$label[allTerm] <- paste(allTerm,allAp,C,O,sep="\\\n")
  nodeAttrs$fontsize[allTerm]<- '60'

  rmLocalTerm <- setdiff(sigTerm, sigTerm_Local)
  nodeAttrs$color[rmLocalTerm] <- rep('red', length(rmLocalTerm))
  nodeAttrs$shape[rmLocalTerm] <- rep('circle', length(rmLocalTerm))

  rmGlobalTerm <- setdiff(sigTerm_Local, sigTerm_Global)
  nodeAttrs$color[rmGlobalTerm] <- rep('red', length(rmGlobalTerm))
  nodeAttrs$shape[rmGlobalTerm] <- rep('box', length(rmGlobalTerm))
  nodeAttrs$height[rmGlobalTerm] <- rep('0.7', length(rmGlobalTerm))
  nodeAttrs$width[rmGlobalTerm] <- rep('0.7', length(rmGlobalTerm))

  nodeAttrs$color[sigTerm_Global] <- rep('red', length(sigTerm_Global))
  nodeAttrs$shape[sigTerm_Global] <- rep('rectangle',
length(sigTerm_Global))
  nodeAttrs$height[sigTerm_Global] <- rep('0.7', length(sigTerm_Global))
  nodeAttrs$width[sigTerm_Global] <- rep('1.1', length(sigTerm_Global))



  dagTermInfo[dagTermInfo[,5]<2.2e-16,5] <- 2.2e-16;
  dagTermInfo[dagTermInfo[,6]<2.2e-16,6] <- 2.2e-16;
  dagTermInfo$colorran <- round(log10(dagTermInfo[,6])-
range(log10(dagTermInfo[,6]))[1] + 1)
  mm <- max(dagTermInfo$colorran)
  colorMap <- heat.colors(mm)
  nodeAttrs$fillcolor[allTerm] <- unlist(lapply(dagTermInfo$colorran,
function(x) return(colorMap[x])))

  weightsList <- edgeWeights(DAG)
  to <- lapply(weightsList, names)
  from <- nodes(DAG)
  edge.names <- paste(rep(from, listLen(to)), unlist(to), sep = "~")
  edge.weights <- unlist(weightsList)
  names(edge.weights) <- edge.names
  ##    0 for a is_a relation,  1 for a part_of relation
  edgeAttrs$color <- ifelse(edge.weights == 0, 'black', 'red')
```

```
  filename <- paste(filename,".bmp",sep="")
  bmp(filename, width = bmpSize, height = bmpSize, res = 800, antialias =
"none");
  plot(DAG, attrs = graphAttrs, nodeAttrs = nodeAttrs, edgeAttrs =
edgeAttrs,main = paste("Ref.=",l_ref,"; Obs.=",l_obs,"; Adj.P=",sigLav))
  dev.off()
}
##                                            ##
##                                            ##
################################################
##          SGD to Entrez id converter        ##
################################################
sgd2entz <-         function(a){
        library (biomaRt)
        ensembl = useMart("ensembl")
        mart <- useMart(biomart = "ensembl", dataset =
"scerevisiae_gene_ensembl")
        aa<- getBM(attributes = "entrezgene",filters = "sgd_gene",values =
a, mart = mart)
        return(aa)}
##                                            ##
##                                            ##
################################################
##  GO Hypergeometric test with redunadancy ##
gohyper <-
function(observedGenes,referenceGenes,ontology="MF",organism=yeast,multiCor
rec="BH",sigLav=0.05,minObs=2){
        ###### Organusm List
        yeast<-"org.Sc.sgd.db"
        human<-"org.Hs.eg.db"

        ###### Required Package
        require(organism, character.only = TRUE) ||
stop(paste("package",organism, "is required", sep = " "))

        ###### Database quary
        .sql <- paste("select distinct t1.gene_id,t2.go_id", " from genes
as t1 inner join",
                paste("go", tolower(ontology), "all", sep = "_"), " as t2
on t1._id=t2._id",
                seq = "")
        organism <- strsplit(organism, ".db")
        organism <- organism[[1]]
        conn <- get(paste(organism, "_dbconn", sep = ""))()
        allAnn <- dbGetQuery(conn, .sql)

        annRef <- allAnn[allAnn[, 1] %in% referenceGenes,]
        annObs <- allAnn[allAnn[, 1] %in% observedGenes,]
        tabRef<-table(annRef)
        tabObs<-table(annObs)
        nGoRef<-ncol(tabRef)
        nGoObs<-ncol(tabObs)

        refGo<-rep( list(list()), nGoRef)
        for(i in 1:nGoRef)  {
            refGo[[i]]<-rownames(tabRef)[which(tabRef[,i]==1)]
                        }
        names(refGo)<-colnames(tabRef)

        obsGo<-rep( list(list()), nGoObs)
        for(i in 1:nGoObs)  {
```

```
            obsGo[[i]]<-rownames(tabObs)[which(tabObs[,i]==1)]
                        }
        names(obsGo)<-colnames(tabObs)


        ###### Summary Table
        nam<-names(obsGo)
        lng<-length(nam)
        summ<-matrix(NA, ncol=6,nrow=lng)
        colnames(summ)<-c("GO","C","E","O","pvalue","adjstp")

        for(k in 1: lng){
            summ[k,1]<-names(obsGo[k])
            summ[k,2]<-length(refGo[[summ[k,1]]])

            ## Expected value
            summ[k,3]<-
round((length(observedGenes)*as.numeric(summ[k,2]))/length(referenceGenes),
2)
            summ[k,4]<-length(obsGo[[summ[k,1]]])

            ## Hypergeometric Distribution
            l_ref<-length(referenceGenes)   #
            l_obs<-length(observedGenes)    # k = Number of balls drawn
from the urn
            n1<-as.numeric(summ[k,4])       # q = Number of white balls
drawn without replacement from an urn which contains both black and white
balls
            n2<-as.numeric(summ[k,2])       # m = Number of white balls in
the urn
            n3<-l_ref-as.numeric(n2)        # n = Number of black balls in
the urn

            summ[k,5]<-1-phyper(n1-1,n2,n3,l_obs,lower.tail=T,log.p=F)
#phyper(q, m, n, k, lower.tail = TRUE, log.p = FALSE)

            }
        p<-as.vector(summ[,5])
        summ[,6]<-p.adjust(p,multiCorrec,n=length(p))
        summ<-data.frame(summ)
        allTerm<-summ[order(as.numeric(as.vector(summ[,6]))),]
        sigTerm<-summ[which((as.numeric(as.vector(summ[,6]))<sigLav) &
(as.numeric(as.vector(summ[,4]))>=minObs)),]

        # All GO term relation
        require("GO.db") || stop("package GO.db is required")
        conn1 <- get("GO_dbconn")()
        .sql <- paste("select distinct t1.go_id parentid,t2.go_id childid
from ",
                      paste("go", tolower(ontology), "offspring", sep = "_"),
                      " as t3 inner join  go_term as t1 on t1._id=t3._id
inner join go_term as t2",
                      " on t2._id=t3._offspring_id", sep = "")
        allTermRelation <- dbGetQuery(conn1, .sql)
        sigTerm<-sigTerm
        sigTermRelation <- allTermRelation[(allTermRelation[, 1] %in%
sigTerm[, 1]) & (allTermRelation[, 2] %in% sigTerm[,1]), ]

        # Local Redundancy
        lrd<-localRedundancy(sigTerm = sigTerm,generalAnn =
allAnn,sigTermRelation = sigTermRelation,annRef = annRef,
```

```
                                annInterest = annObs,ppth = sigLav,pcth =
sigLav)

        # Global Redundancy
        grd<-globalRedundancy(generalAnn = allAnn,sigTermRelation =
sigTermRelation,annRef = annRef,
                                annInterest = annObs,sigTermRedun =
lrd,poth = sigLav,peth = sigLav)

        # All GO term names
        require("GO.db") || stop("package GO.db is required")
        conn2 <- get("GO_dbconn")()
        .sql <- paste("select distinct go_id goid,term name from go_term
where ontology='",toupper(ontology), "'", sep = "")
        allTermName <- dbGetQuery(conn2, .sql)

        # Vizualize GO DAG
        require("graph") || stop("package graph is required")
        sigDAG <- createGODAG(as.character(sigTerm[, 1]), ontology)

        allDAGTerm <- allTerm[allTerm[, 1] %in% nodes(sigDAG), ]
        allDAGTerm <- allDAGTerm[order(allDAGTerm[, 1]), ]

        dagTermName <- allTermName[allTermName[, 1] %in% allDAGTerm[,1], ]
## DAG node terms names
        dagTermName <- dagTermName[order(dagTermName[, 1]), ]

        allDAGTerm$name <- dagTermName[, 2]
        allDAGTerm <- allDAGTerm[, c(1, 7, 2, 4, 5, 6)]

        allDAGTerm<-transform(allDAGTerm, C = as.numeric(as.vector(C)),
                                O = as.numeric(as.vector(O)),
                                pvalue = as.numeric(as.vector(pvalue)),
                                adjstp  = as.numeric(as.vector(adjstp)))

        sigTermID <- as.character(sigTerm[, 1])
        sigTerm_LocalRedunID <- as.character(lrd[, 1])
        sigTerm_GlobalRedunID <- as.character(grd[, 1])

        a<-showSigNodes(DAG = sigDAG, sigTerm = sigTermID, sigTerm_Local =
sigTerm_LocalRedunID,l_ref,l_obs,sigLav,
                        sigTerm_Global = sigTerm_GlobalRedunID,dagTermInfo
= allDAGTerm, bmpSize = 4000, filename = ontology )

        label <- array("", dim = c(nrow(sigTerm), 1))
        rmsigLocalTerm <- setdiff(sigTermID, sigTerm_LocalRedunID)
        label[sigTermID %in% rmsigLocalTerm, 1] <- "Local"
        rmsigGlobalTerm <- setdiff(sigTerm_LocalRedunID,
sigTerm_GlobalRedunID)
        label[sigTermID %in% rmsigGlobalTerm, 1] <- "Global"
        label[sigTermID %in% sigTerm_GlobalRedunID, 1] <- "Final"
        sigTerm$FinalResult <- label

        sigTerm<-sigTerm[order(as.numeric(as.vector(sigTerm$adjstp))),]
        tablename <- paste(ontology, ".xlsx", sep = "")
        #write.csv(sigTerm, tablename, quote = FALSE,row.names = F)
        #require(organism, character.only = TRUE) ||
stop(paste("package",organism, "is required", sep = " "))
        library(xlsx)
        write.xlsx(sigTerm,tablename,sheetName="Sheet1",col.names=TRUE,
row.names=TRUE, append=FALSE)
```

```
return(list(refGo,obsGo,allTerm,sigTerm,lrd,grd,sigDAG,allTermName,dagTermN
ame,allDAGTerm))
        }
##                                            ##
################################################



################################################
##              Reading Gene list             ##
################################################
obs<-scan("input/obs.txt",what="")
ref<-scan("input/ref.txt",what="")


################################################
##              Convert Gene ID               ##
##          SGD -> Entrez gene ID             ##
##          SGD -> Entrez gene ID             ##
################################################
obs<-sgd2entz(obs)
ref<-sgd2entz(ref)


################################################
##          Hypergeometric test               ##
################################################
res1<-gohyper(obs,ref,ontology="BP")
################################################

################################################
## Upto here we will get significant GO     ##
## term in observed gene list by            ##
## Hypergeometric distibution with Global,  ##
## Local and Final redundancy .             ##
################################################



################################################
## A group of genes in obsberved and        ##
## reference category can be annotated by a ##
## specific GO term.                        ##
## Following code is to produce protein     ##
## sequences of that group of genes in      ##
## obsberved and reference category in      ##
## FASTA formate for a specific GO term.    ##
################################################
##                                            ##
##                                            ##
################################################
##          Reading Gene list                 ##
################################################
obs<-scan("input/obs.txt",what="")
ref<-scan("input/ref.txt",what="")

################################################
##          Sefl writen function              ##
##                                            ##
##                                            ##
```

```
##############################################
##      SGD to Entrex gene id converter     ##
##############################################
sgd2entzid <-   function(a){
        library (biomaRt)
        ensembl = useMart("ensembl")
        mart <- useMart(biomart = "ensembl", dataset =
"scerevisiae_gene_ensembl")
        aa<- getBM(attributes = "entrezgene",filters = "sgd_gene",values =
a, mart = mart)
        return(aa)}
##############################################
##                                          ##
##                                          ##
##      Entrez to SGD ID converter          ##
##############################################
entrez2sgdid <- function(a){
        library (biomaRt)
        ensembl = useMart("ensembl")
        mart <- useMart(biomart = "ensembl", dataset =
"scerevisiae_gene_ensembl")
        aa<- getBM(attributes = "sgd_gene",filters = "entrezgene",values =
a, mart = mart)
        return(aa)}
##############################################
##                                          ##
##                                          ##
## Entrez ID to Protein Sequence FASTA      ##
## format( This program in only optimized   ##
## for yeast)                               ##
id2prot<-function(g,filnm){
        ## Loading library
        library(biomaRt)

        ## Select biomaRt database and dataset
        #ensembl = useMart("ensembl")
        mart <- useMart("ensembl", dataset = "scerevisiae_gene_ensembl")

        seq<-
biomaRt::getSequence(id=g,type="entrezgene",seqType="peptide",mart=mart)
        colnames(seq)<-c("pro_seq","entz_id")

        zz <- file(filnm,"w")
        exportFASTA(seq,zz)
        close(zz)
        }
##############################################
##                                          ##
##                                          ##
##  Specific GO term in observe and         ##
##  reference group                         ##
go2fasta<-
function(GO,ontology,observedGenes,referenceGenes,organism=yeast){
        ###### Organusm List
        yeast<-"org.Sc.sgd.db"
        human<-"org.Hs.eg.db"

        ###### Required Package
        require(organism, character.only = TRUE) ||
stop(paste("package",organism, "is required", sep = " "))
```

```
        ###### Database quary
        .sql <- paste("select distinct t1.gene_id,t2.go_id", " from genes
as t1 inner join",
        paste("go", tolower(ontology), "all", sep = "_"), " as t2 on
t1._id=t2._id",
        seq = "")
        organism <- strsplit(organism, ".db")
        organism <- organism[[1]]
        conn <- get(paste(organism, "_dbconn", sep = ""))()
        allAnn <- dbGetQuery(conn, .sql)

        ###### Get the genes in obs and ref catagory
        annObs <- allAnn[allAnn[, 1] %in% observedGenes,]
        annRef <- allAnn[allAnn[, 1] %in% referenceGenes,]

        ###### Get the gene in obs and ref catagory with required GO term
        annObsGO<-annObs[annObs[,2] %in% GO,]
        annRefGO<-annRef[annRef[,2] %in% GO,]

        ## Creat directory to save  temporary working file
        a<-dir()
        if(as.numeric(as.numeric("outf"%in%a)==0)==1) dir.create("outf")

        ##### Writing and Reading FASTA file
        annObsGO_prot<-
id2prot(as.numeric(annObsGO[,1]),"outf/annObsGO_prot.fasta")
        annRefGO_prot<-
id2prot(as.numeric(annRefGO[,1]),"outf/annRefGO_prot.fasta")

        require("seqinr", character.only = TRUE) ||
stop(paste("package",seqinr, "is required", sep = " "))
        annObsGO_protFasta<-read.fasta("outf/annObsGO_prot.fasta",seqtype =
"AA", as.string = FALSE)
        annRefGO_protFasta<-read.fasta("outf/annRefGO_prot.fasta",seqtype =
"AA", as.string = FALSE)

        return(list(allAnn,annObsGO,annRefGO))
        }
###############################################


###############################################
## Interested significant GO term           ##
###############################################
GO<-"GO:0000023"
ontology<-"BP"

obs<-sgd2entzid(obs)
ref<-sgd2entzid(ref)

res2<-go2fasta(GO,ontology,obs,ref)



###############################################
##          Self writen function             ##
##                                           ##
## Chromosome wise distribution of genes     ##
## in observed and reference category        ##
###############################################
##                                           ##
```

```
## data.xls and cent.txt are provided in    ##
## suplymentary data                        ##
fun<-function(a)    # a is a vector
    {
    ## Loading Library
    library (xlsReadWrite)

    ## Loading Data
    data_m<-data.frame(read.xls("func/data.xls", colNames=TRUE))    ## Main
data
    cent<-data.frame(read.table("func/cent.txt",header=T))          ##
Position of centromere


    ## Data set of required gene or genes set
    data<-data_m[which(data_m$pi%in%a),]
    tab<-table(data$chrom)
    n<-dim(tab)

    ## Ploting data
    plot(1:n,1:n,ylim=c(1,max(data_m$sto)+200000),col="white",axes =
FALSE,xlab="Yeast Chromosome",ylab="")
    axis(1,at=1:n,labels=as.numeric(names(tab)))

    legend("topright",bty="n",c("Watson Strand (5'-3')","Crick Strand (3'-
5')"),cex=0.6,col=c("red","blue"),lty=c(1,1),lwd=c(2,2))

    for(i in 1: n)
        {
        c<-as.numeric(names(tab[i]))
        d<-subset(data,data$chrom==c)
        d1<-subset(data_m,data_m$chrom==c)

        ### Line the string : 2 lines
        lines(c(i-0.1,i-0.1),c(0,max(d1$sto)),col="grey80",lwd=5)   ## ORF
= 1
        lines(c(i+0.1,i+0.1),c(0,max(d1$sto)),col="grey80",lwd=5)   ## ORF
= -1

        wg<-sum(as.numeric(d$orf==1))
        cg<-sum(as.numeric(d$orf==-1))
        legend(i-
0.1,max(d1$sto)+100,legend=wg,bty="n",cex=0.6,xjust=0.75,yjust=0,text.col="
red")

legend(i+0.1,max(d1$sto)+100,legend=cg,bty="n",cex=0.6,xjust=0.75,yjust=0,t
ext.col="blue")

legend(i,max(d1$sto)+100,legend=wg+cg,bty="n",cex=0.7,xjust=0.75,yjust=-
0.5,text.col="black")

        ### plot centromere
        lines(c(i,i),c(cent$str[i],cent$sto[i]),lwd=10)

        ## plot genes
        w<-which(d$orf==1)
        c<-which(d$orf==-1)
        if ((length(w)>0)==TRUE) for (j in 1:length(w))
                        {
                        lines(c(i-0.1,i-
0.1),c(d$str[w][j],d$sto[w][j]),col="red",lwd=2)
```

```
                        }
        if ((length(c)>0)==TRUE)for (j in 1:length(c))
                        {

lines(c(i+0.1,i+0.1),c(d$str[c][j],d$sto[c][j]),col="blue",lwd=2)
                        }

        }
return(tab)
    }
################################################

sgdObs<-entrez2sgdid(res2[[2]])
sgdRef<-entrez2sgdid(res2[[3]])

x11()
resObs3<-fun(sgdObs)
x11()
resRef3<-fun(sgdRef)



################################################
## Seqence allignment by clustalx softwars  ##
## and the allignment and tree file were    ##
## saved in .aln and .dnd formate           ##
## respectively                             ##
################################################

################################################
##  Draw Dendrogrrm from clustalx output    ##
################################################

require("ape", character.only = TRUE) || stop(paste("package",ape, "is
required", sep = " "))
obsTree <- read.tree(file="outf/annObsGO_prot.dnd")
refTree <- read.tree(file="outf/annRefGO_prot.dnd")
x11()
plot.phylo(refTree,show.tip.label = F)
axisPhylo()
tipObs<-obsTree$tip.label
tipRef<-refTree$tip.label
tiplabels(tipObs,match(tipObs,refTree$tip.label),cex=0.8,col="red")
title(main = paste(GO,"; Obs.=",length(tipObs),"; Ref.=",length(tipRef)))
```

Bibliography

1. *Human cancers express mutator phenotypes: origin, consequences and targeting.* **Loeb, Lawrence A.** June 2011, Nature Reviews Cancer, Vol. 11, pp. 450-457.

2. *The remaking of chromosomes.* **Muller, H. J.** 198, 1938, The Collecting Net, Vol. 13, pp. 181-195.

3. *The Behavior in Successive Nuclear Divisions of a Chromosome Broken at Meiosis.* **McClintock, Barbara.** 8, August 1939, National Accademy of Science, Vol. 25, pp. 405-416.

4. *The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity.* **Greider, C. W. and Blackburn, E. H.** 1987, Cell, Vol. 51, pp. 887-898.

5. *Molecular basis for telomere repeat divergence in budding yeast.* **Forstemann, K. and Lingner, J.** 2001, Mol. Cell Biol., Vol. 21, pp. 7277-7286.

6. *Organization of DNA sequences and replication origins at yeast telomeres.* **Chan, C. S. and Tye, B. K.** 1983, Cell, Vol. 33, pp. 563-573.

7. *Mitotic recombination among subtelomeric Y' repeats in Saccharomyces cerevisiae.* **Louis, E. J. and Haber, J. E.** 1990, Genetics, Vol. 124, pp. 547-559.

8. *Unusual DNA sequences associated with the ends of yeast chromosomes.* **Walmsley, R. W., et al., et al.** 1984, Nature, Vol. 310, pp. 157-160.

9. *Testing of Chromosomal Clumping of Gene Properties.* **Luciano, Daniel, Anders, Olle.** 1, February 2009, Statistical Applications in Genetics and Molecular Biology, Vol. 8.

10. *Essential Bacillus subtilis genes. Proc. Natl Acad. Sci. USA, 100, 4678–4683.* **Kobayashi K., Ehrlich,S.D., Albertini,A., Amati,G., Andersen,K.K., Arnaud,M., Asai,K., Ashikaga,S., Aymerich,S., Bessieres,P. et al. (2003).**

11. *An estimation of minimal genome size required for life. .* **(1995), Itaya M.**

12. *Functional profiling of the Saccharomyces cerevisiae genome.* **Giaever G, et al.** 418, July 25, 2002, Nature, pp. 387-391.

13. *Selection at Linked Sites Shapes Heritable Phenotypic Variation in C. elegans.* **Matthew V. Rockman, Sonja S. Skrovanek, Leonid Kruglyak.** 6002, 2010, Science, Vol. 330, pp. 372-376.

14. *SGD: Saccharomyces Genome Database.* **Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D.** 1, January 1998, Nucleic Acids Research, Vol. 26.

15. **Loader, William S. Cleveland and Clive Loader.** *Smoothing by Local Regression: Principles and Methods.* USA : AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ.

16. *The evolutionary dynamics of eukaryotic gene order.* **Hurst, L.D., C. Pal, and M.J. Lercher.** March 2003, Nature Genetics, Vol. 33.

17. *Probability.* **Jim, Pitman.** s.l. : United States of America: Springer Science Business Media, LLC., 2006.

18. *The Simes method for multiple hypothesis testing with positively dependent test statistics.* **K., Sarkar S. and K, Chang C.** 1997, Journal of the American Statistical Association, Vol. 92, pp. 1601-1608.

19. *GO-function: deriving biologically relevant functions fromstatistically significant functions.* **JingWang, Xianxiao Zhou, Jing Zhu, Yunyan Gu,Wenyuan Zhao, Jinfeng Zou and Zheng Guo.** 2, Briefings in Bioinformatics, Vol. 13, pp. 216-227.

20. *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* **Ashburner M, Ball CA, Blake JA,.** 2000, Nat Genet, Vol. 25, pp. 25-19.

21. *Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.* **Camon E, Magrane M, Barrell D.** 2004, Nucleic Acids Res, p. 32.

22. *GO-2D: Identifying 2-dimensional cellular-localized functional modules in Gene Ontology.* **J., Zhu, J., Wang and Z., Guo.** 30, 2007, BMC Genomics, Vol. 8.

23. *GO-function: deriving biologically relevant functions from statistically significant functions.* **J, Wang, et al., et al.** 2, 2012, Brief Bioinformatics, Vol. 13, pp. 216-227.

24. *Technique of Molecutar Analysis: A step by step guide to phylogeny reconstruction.* **Langdale, C. Jill Harrison and Jane A.** 2006, The Plant Journal, Vol. 45, pp. 561-572.

25. **DM, Mount.** *Bioinformatics: Sequence and Genome Analysis (2nd ed.).* NY : Cold Spring Harbor Laboratory Press: Cold Spring Harbor, 2004.

26. *On the complexity of multiple sequence alignment.* **L, Wang and T, Jiang.** 4, 1994, Journal of couputational Biology, Vol. 1, pp. 337-348.

27. *Settling the intractability of multiple alignment.* **Elias and Isaac.** 7, 2006, Journal of Computational Biology, Vol. 13, pp. 1323–1339.

28. *Genetic pleiotropy in Saccharomyces cerevisiae quantified by high-resolution phenotypic profiling.* **Ericson, E., et al.** 2006, Molecular Genetics and Genomics, Vol. 275(6), pp. 605-14.

29. *Sequencing and comparison of yeast species to identify.* **Kellis, M., et al.** 2003, Nature, Vol. 423(6937), pp. 241-54.

30. Variation in chromosome number. [Online] http://www.ndsu.edu/pubweb/~mcclean/plsc431/chromnumber/number1.htm.

31. *Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization.* **Spellman, Paul T., et al., et al.** 12, December 1, 1998, Molecular Biology of the Cell, Vol. 9, pp. 3273-3297.

32. Yeast Cell Cycle Analysis Project website. [Online] http://genome-www.stanford.edu/cellcycle/.

33. *Everything You Ever Wanted to Know About Saccharomyces cerevisiae Telomeres: Beginning to End.* **Wellinger, Raymund J. and Zakian, Virginia A.** 2012, Genetics, Vol. 191.

34. *TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes. Nat. Biotechnol.* **Judson N. and Mekalanos, J.J. (2000).**

35. **Ross, S.** *A First Course in Probability*. s.l. : Macmillan, 1976.

36. *Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language.* **P., Resnik.** 1999, Journal Of Artificial Intelligence Research, Vol. 11, pp. 95-130.

37. *Semantic similarity based on corpus statistics and lexical taxonomy* . **J., Jiang and D., Conrath.** 1997, Proceedings of Intertional Conference Research on Computational Linguistics , pp. 19-33.

38. *AN information theoretic definition of similarity.* **D, Lin.** 1998, ICML, pp. 296-204.

39. *A new method to measure the semantic similarity of GO terms.* **J., Wang, et al., et al.** 2007, Bioinformatics, Vol. 23, pp. 1274-1281.

40. http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Clustering_Overview.htm. [Online]

41. Clustering Overview. [Online] http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Clustering_Overview.htm.

42. Hierarchical Clustering Overview. [Online] http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Agglomerative_Hierarchical_Clustering_Overview.htm.

43. **Wikipedia.** http://en.wikipedia.org/wiki/Gene. [Online]

44. *Genetics: what is a gene?* **H, Pearson.** 7092, 2006, Nature, Vol. 441, pp. 398–401.

45. *DNA Study Forces Rethink of What It Means to Be a Gene.* **Pennisi, Elizabeth.** 5831, Science, Vol. 316, pp. 1556–1557.