© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Science and Technology

Computational Techniques for the Functional Annotation of Hypothetical ORFS in Human Chromosome 3

Sivashankari Selvarajan[1], Piramanayagam Shanmughavel[2]

[1]Assistant Professor in UGC, Innovative Programme, Department of Bioinformatics, Nirmala College for Women, Coimbatore, Tamil Nadu, India
[2]Associate Professor, Department of Bioinformatics, Bharathiar University, Coimbatore, Tamil Nadu, India

ABSTRACT

In biochemistry, a hypothetical protein encoded by a hypothetical gene is a protein whose existence has been predicted, for which there is no experimental evidence for expression *in vivo.* As a result, the function of such genes is not known. Despite several efforts, only 50-60 % of genes have been annotated in most completely sequenced genomes and their functions are known. The rest 40% of the genes in any genome is totally unknown in terms of its functions. As of September 2010, there are around 637 genes encoded as Hypothetical in NCBI. So, the present investigation focused on functional annotation of hypothetical genes in the Chromosome 3 of the Human Genome.
Keywords: Annotation, Chromosome3, Function, Hypothetical Genes

I. INTRODUCTION

The human genome project revealed the three billion base pairs encrypted within the twenty three pairs of chromosomes in the human genome. Also, the Human Genome contains 30,000 genes, constituting just 1% of the ~3 billion base pairs of the total human DNA**.** Among these, there are genes (called Hypothetical ORFs) which code for the so-called "hypothetical proteins" whose existence is either validated experimentally or predicted computationally but its function is not yet reported. Hence, after the completion of the genome sequences, the challenge ahead for all biologists is to use the data to interpret the function of the protein, the cell, and the organism. This can be achieved by a process called annotation which involves identification of genes within the chromosome, its fine structure, determination of protein products encodes by the gene and understanding the function *(*Venter *et al., 2001).* A group of these genes may be involved in many pathological disorders and hence are of pharmaceutical significance. Thus, annotation is an essential process of understanding the entire mechanism behind the cellular processes and molecular functions of a genome. However, there were inconsistencies in the accuracy of genome annotation in the initial stages which are now gone due to advancements in computational algorithms and potentiality of bioinformatics. After annotation of the Human Genome a number of genes (59%) reported by the project were hypothetical and annotated genes with unknown function *(*Venter *et al 2001)* (Table 1).

Table 1.1 The Human Genome Statistics

S.No	Topic	Statistic
1	Total size of the genome	approximately 3,200,000,000 bp
2	Percentage of DNA spanned by genes	between 25% and 38%
3	Percentage of exons	1.1 to 1.4%
4	Percentage of introns	24% to 37%
5	Occurrence rate of genes	about 12 per 1,000,000 bp
6	Percent of hypothetical genes and annotated genes with unknown function in the genome	59%

Source: Venter *et al*., 2001

As of September 2010, there are around 637 genes encoded as Hypothetical in NCBI. These hypothetical ORFs may be functionally important and play very important roles in growth, development and maintenance of *Homo sapiens*. Research is needed to unravel the function of these conserved hypothetical genes in Human to understand more about molecular mechanisms and biological significance of the entire Human Genome.

Among the 23 pairs of chromosomes in the human genome, Chromosome 3 spans almost 200 million base pairs and represents about 6.5 percent of the total DNA in cells. Identifying genes on each chromosome is an active area of genetic research. Because researchers use different approaches to predict the number of genes on each chromosome, the estimated number of genes varies. Chromosome 3 likely contains between 1,100 and 1,500 genes. Hence, the present work aims to investigate and predict the function of the Hypothetical genes in the chromosome 3 of the Human Genome using Computational approaches which will be a guide for experimental analysis. So, the present investigation focused on functional annotation of hypothetical genes in the Human Genome with the following Objectives:

1. To identify the uncharacterized hypothetical Genes in chromosome3 of the Human Genome.
2. To annotate the function of the uncharacterized hypothetical genes in the human Genome both at gene and Protein level.
3. To assign functional categories for the annotated hypothetical genes.
4. To assign transmembrane Topology and Sub-Cellular Localization for the unannotated Hypothetical Genes.

II. METHODS AND MATERIAL

The Hypothetical ORFs in chromosome 3 of the Human Genome was retrieved from NCBI [Geer *et al*., 2010]. To identify whether the ORFs can actually be genes two strategies were followed. Initially, conservation of the ORF in other organisms was determined using Homologene [Geer *et al*., 2010] and then its coding potential was calculated using Coding Potential Calculator (CPC) [Lei Kong et al., 2007]. Conservation and coding potential were determined, because the ORFs have a high probability to be functional if they are conserved and having a higher coding potential score.

Next, the annotation of the hypothetical genes which are conserved and with coding potential at the nucleotide level was done using BLAST2GO [Ana *et al*., 2005] with the following algorithm: Initially, the sequence was queried against BLAST to find homologs followed by mapping of the sequence with GO terms including running Interproscan. At the protein level, pfam [Robert *et al*., 2013] and supfam database[Pandit *et al*., 2002] were used to assign domains and superfamily to the hypothetical proteins. Finally, COG[Geer *et al*., 2010] and SCOP[Murzin et al., 2002] were used to assign functional category to the hypothetical genes. The transmembrane topology and subcellular localization were predicted for the unannotated hypothetical genes using TMHMM[Krogh et al., 2001] and PSort[Nakai and Horton, 1999]

III. RESULTS AND DISCUSSION

CHROMOSOME 3

There are 27 hypothetical ORFs in the chromosome 3 of the Human Genome. The results of Homologene and Coding Potential Calculator are presented in Table 2. It can be inferred from the Table that, all the hypothetical genes in Chromosome 3 of human are conserved; 24 genes (66%) are conserved in *Mus musculus*, 25 genes (93%) are conserved in *Pan troglodytes*, 20 genes are conserved in *Bos Taurus* (74%), 22 genes (81%) are conserved in *Rattus norvegicus*, 16 genes (59%) are conserved in *Canis lupus familiaris*, 14 genes (52%) are conserved in *Gallus gallus*, 13 genes (48%) are conserved in *Danio rerio*, 4 genes (15%) are conserved in *Drosophila melanogaster*, 3 genes (11%) are conserved in *Anopheles gambiae*, 2 genes (7%) are conserved in *Caenorhabditis elegans*, 1 gene (4%) is conserved in *Arabidopsis thaliana*, *Oryza Sativa*, *Magnaporthe grisea* and *Neurospora crassa*. Further, all the hypothetical genes in Chromosome 3 of human have strong coding potential for proteins except C3orf70 which have weak coding potential.

Table 2
Conservation and Coding Potential of Hypothetical Genes in Human - Chromosome 3

Sl No.	Gene Name	Conservation	Coding potential
1	C3orf 23	*Pan troglodytes, Canis familiaris, Mus musculus, Rattus norvegicus, Gallus gallus, Danio rerio, Drosophila melanogaster, Anopheles gambiae, and C.elegans.*	0.523916
2	C3orf 24	*Pan troglodytes, Bos taurus, Mus musculus, Rattus norvegicus, and Gallus gallus.*	3.3725
3	C3orf 18	*Pan troglodytes, Canis familiaris, Bos taurus, Mus musculus, Rattus norvegicus, and Gallus gallus.*	4.17433
4	C3orf 45	*Pan troglodytes, Canis familiaris, Bos taurus, Mus musculus, and Rattus norvegicus.*	2.1979
5	C3orf 26	*Canis familiaris, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, and Danio rerio.*	6.02151
6	C3orf 19	*Pan troglodytes, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, Danio rerio, Drosophila melanogaster, Anopheles gambiae, S.pombe, M.grisea, and N.crassa.*	8.63648
7	C3orf 32	*Pan troglodytes, Canis familiaris, Mus musculus, Danio rerio, C.elegans, A.thaliana, and Oryza sativa.*	6.25219
8	C3orf 30	*Pan troglodytes, Canis familiaris, Bos taurus, and Mus musculus.*	6.68395
9	C3orf 70	*Pan troglodytes, Canis familiaris, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, and Danio rerio.*	0.919875 *
10	C3orf 33	*Pan troglodytes, Canis familiaris, Bos taurus, Mus musculus, Rattus norvegicus, and Danio rerio.*	1.58832
11	C3orf 38	*Pan troglodytes, Canis familiaris, Mus musculus, Rattus norvegicus, Gallus gallus, and Drosophila melanogaster*	6.8199
12	C3orf 22	*Pan troglodytes, Bos taurus, Mus musculus, and Rattus norvegicus.*	1.81438
13	C3orf 21	*Pan troglodytes, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, Danio rerio, Drosophila melanogaster, and Anopheles gambiae.*	8.37928
14	C3orf 59	*Pan troglodytes, Canis familiaris, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, and Danio rerio.*	11.1691
15	C3orf 14	*Pan troglodytes, Canis familiaris, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, and Danio rerio.*	2.54871
16	C3orf 67	*Pan troglodytes, Canis familiaris, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, and Danio rerio.*	11.3101
17	C3orf 17	*Pan troglodytes, Canis familiaris, Mus musculus, Rattus norvegicus, and Danio rerio.*	5.23593
18	C3orf 62	*Pan troglodytes, Bos taurus, Mus musculus, and Rattus norvegicus*	3.19484
19	C3orf 77	*Pan troglodytes, Mus musculus, Rattus norvegicus, and Gallus gallus.*	15.4478
20	C3orf 16	*Pan troglodytes, Bos taurus, Mus musculus, and Rattus norvegicus.*	8.51717
21	C3orf 54	*Pan troglodytes, Canis familiaris, Bos taurus, Mus musculus, Rattus norvegicus, and Danio rerio.*	3.77406
22	C3orf 43	*Pan troglodytes, Canis familiaris, Bos taurus, Mus musculus, and Rattus norvegicus.*	5.41578
23	C3orf 20	*Pan troglodytes, Bos taurus, Mus musculus, and Rattus norvegicus.*	17.5078
24	C3orf 36	*Not Conserved*	4.21101
25	C3orf 71	*Pan troglodytes*	0.484316 *
26	C3orf 34	*Pan troglodytes, Canis familiaris, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, and Danio rerio.*	2.03661
27	C3orf78	*Pan troglodytes, Bos taurus, and Gallus gallus*	1.43207

* Weak Coding Potential

Among the 27 hypothetical ORFs in Chromosome 3, 6 of them are previously characterized (Supplementary Table 1s) in Swissprot and the remaining 21 are subjected to protein and nucleotide level annotation

using Superfamily database and BLAST2GO respectively. The results are tabulated in Supplementary Table 2.

From the table, it is evident that all the 6 super families within the hypothetical genes of the Chromosome 3 were identified with reasonable confidence using Superfamily database. Further, 'C-type lectin domain' is identified with C3orf45 hypothetical gene in Chromosome 3.

The hypothetical genes C3orf33 and C3orf71 were found to contain nucleic acid binding sites with hydrolase activity and isomerase activity respectively. Similarly, C3orf43 was found to contain an ATP domain with hydrolase activity by BLAST2GO. Another type of ATP synthase was identified in C3orf77 by SUPFAM database. Among the 16 annotated hypothetical genes, majority of them are involved in Chromosome 3. It can be evident from the functional category assignment in Table 3. However, five hypothetical genes could not be assigned any functional information using this approach. Hence, transmembrane prediction and sub-cellular localization were identified for these seven hypothetical genes and presented in Table 4.

Table 3: Functional Categories for the Hypothetical Genes in Chromosome 3

S.No	Group Code	Description	*Number* [a]
COG category			
1	J, A, K, L, B	Information storage and processing	*1*
2	D, Y, V, T, M, N, Z, W, U, O	Cellular processes and signaling	-
3	C, G, E, F, H, I, P, Q	Metabolism	*2*
4	R, S	Poorly characterized	-
SCOP Category			
5	RF, RE, P, MA, RG, SB, D, IA, N, NA, O, OA	PROCESSES	*3*
7	CA, C, CB, E, EA,F, G, GA, GB, H, RA, RB, RC, I, M, Q	METABOLISM	*6*
8	HA, HB, HC, HE, R, RD, ST	GENERAL	*6*
9	B, J, K, L,LB, Y	INFORMATION	*1*
10	A, LA, OB, T, TA, HD	REGULATION	-
10	S, SA	OTHER	-
11	*NONA*	*NOT ANNOTATED*	-

[a] '–'indicates there is no newly annotated gene in this COG or SCOP functional category.

On perusal of Table 3, it is evident that many of the hypothetical genes in Chromosome 3 are involved in metabolism and only meager number of genes performs process, information and regulation activities. Similarly, COG functional assignment reveals nearly equal number of hypothetical genes is involved in metabolism and information storage. This validates that Chromosome 3 is a store house of many metabolically important proteins and enzymes.

Table 4: Topology and Localization of Un-annotated Hypothetical Genes in Chromosome 3

S.No.	Gene	Subcellular Localization	TM Topology
1	C3orf 30	Nucleus	No
2	C3orf 22	Extracellular	No
3	C3orf 67	Extracellular	No
4	C3orf 36	Nucleus	No
5	C3orf 23	mitochondrion	No

It is evident from Table 4 that, majority of the un-annotated hypothetical genes in the human chromosome 3 are localized in nucleus, Extracellular region and none of them has trans membrane topology.

IV. CONCLUSION

To sum up, out of 21 uncharacterized hypothetical genes in Chromosome 3, five of them could not be

annotated functionally. But their membrane topology and sub cellular localization are predicted to aid in experimental expression studies. Two of the un-annotated hypothetical genes in Chromosome 3 are localized in Nucleus and another two in extracellular region, and only one is localized in Mitochondria. There is no membrane topology seen among the un-annotated hypothetical genes in Chromosome 3 of the Human Genome. The present work has resulted in the identification of function for majority of the hypothetical proteins in the Human Chromosome 3 which has to be validated experimentally.

V. REFERENCES

[1] Ana Conesa,Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón1and Montserrat Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics, 2005, Volume 21, Issue 18 Pp. 3674-3676.

[2] Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH. The NCBI BioSystems database. Nucleic Acids Res. 2010 Jan; 38(Database issue):D492-6.

[3] Krogh A, Larsson B, von Heijne G, Sonnhammer EL Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001 Jan 19;305(3):567-80.

[4] Lei Kong, Yong Zhang, Zhi-Qiang Ye, Xiao-Qiao Liu, Shu-Qi Zhao, Liping Wei and Ge Gao, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine, Nucleic Acids Research, 2007, Volume 35, Issue suppl 2 Pp. W345-W349

[5] Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.

[6] Nakai K and Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, Trends Biochem Sci. 1999 Jan;24(1):34-6.

[7] Pandit SB, Gosar D, Abhiman S, Sujatha S, Dixit SS, Mhatre NS, Sowdhamini R, Srinivasan N, SUPFAM--a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. Nucleic Acids Res. 2002 Jan 1;30(1):289-93.

[8] Robert D. Finn, Alex Bateman, Jody Clements, Penelope Coggill ,Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L. L. Sonnhammer, John Tate and Marco Punta, Pfam: the protein families database, Nucleic Acids Research, 2013, Volume 42, Issue D1, Pp. D222-D230.

[9] Venter,J.C., Adams,M.D., Myers,E.W. et al. (2001) The sequence of the human genome. Science, 291, 1304–1351.

Supplementary Table 1: Characterized Hypothetical ORFs in Chromosome 3 of Human

S. No	Gene Name	Component, Process, Function	Domain/Superfamily	Function	COG	SCOP
1	C3orf 26	F: ATP Binding F: ATP-Dependent helicase activity F: Nucleic acid binding	Helicases superfamily	Information storage and processing: DNA replication, recombination and repair	L	--
2	C3orf 32	-	DnaJ-class molecular chaperone with C-terminal	Cellular processes: Posttranslational modification,	O	--

				protein turnover, chaperones		
3	C3orf 21	C: Integral to membrane C: Membrane F: Transferase activity, Transferring Glycosyl Groups	Glycosyltransfe rase family A Nucleotide-diphospho-sugar transferases	Cell envelop biogenesis	M	-
4	C3orf 59	-	Mab-21 protein	Unknown function	S	-
5	C3orf 16	-	Ankyrin repeats	General	-	*R*
6	C3orf 34	-	Phosphoenolpyr uvate carboxylase	Metabolism: Energy production and conversion	C	--

Supplementary Table 2: Functional annotation of Hypothetical Genes in Chromosome 3

Sl No.	GENE NAME	COMPONENT, PROCESS, FUNCTION	DOMAIN / SUPERFAMILY	FUNCTION	COG	SCOP
1	C3ORF18	C: Integral to Membrane	-	General	-	R
2	C3ORF45	C: Integral to Membrane	C-Type Lectin	Processes_EC: Cell Adhesion	-	MA
3	C3ORF70	P: Biological Process; C: Cellular Component; F: Molecular Function	-	General	-	R
4	C3ORF33	C: Integral to Membrane F: Hydrolase Activity, Acting on Ester Bonds F: Nucleic Acid Binding	Staphylococcal Nuclease	Metabolism: Nucleotide Transport and Metabolism	-	F
5	C3ORF38	P: Apoptosis	NTF2-LIKE	Processes_IC: Transport	-	RF
6	C3ORF14	-	Tetrahydrobiopterin Biosynthesis Enzymes-Like	Metabolism: Nucleotide Transport and Metabolism	-	F
7	C3ORF17	C: Integral to Membrane	GFP-Like	Metabolism: Energy Production and Conversion	-	C
8	C3ORF77	-	V-Type ATP Synthase Subunit C; Fuma C-Terminal Domain-Like	Metabolism: Electron Transfer/Transport	-	CA;E
9	C3ORF54	-	Acyl-CoA Dehydrogenase NM Domain-Like	Metabolism: Coenzyme Metabolism and transport	-	H

10	C3ORF43	C: Integral to Membrane C: Proton-Transporting Two-Sector ATPase Complex, Catalytic Domain; C: Integral to Membrane; F: Hydrolase Activity, Acting on Acid Anhydrides, Catalyzing Transmembrane Movement of Substances	-	Cellular Processes: Inorganic Ion Transport And Metabolism	P	-
11	C3ORF20	C: Cytoplasm C: Integral to Membrane	-	General	-	R
12	C3ORF71	C: Integral to Membrane F: Cholestenol Delta-Isomerase Activity; F: Nucleic Acid Binding; F: Endonuclease Activity; C:Integral To Membrane; F: Metal Ion Binding; P: RNA Catabolic Process; P: Sterol Metabolic Process; C: Extracellular Region; C: Endoplasmic Reticulum Membrane; P: Transcription; F: Ribonuclease Activity	-	Metabolism : Lipid Metabolism; Information Storage and Processing: Transcription	I;K	-
13	C3ORF78	C: Integral to Membrane, Mitochondria	-	General	-	R
14	C3ORF62	C: Centrosome; C: Microtubule; F: Protein Binding	-	General	-	R
15	C3ORF19	-	Tropomyosin; XRCC4, C-Terminal Oligomerization Domain	Processes_IC: Cell Motility, Cytoskeleton; Information: DNA Replication, Recombination, Repair	-	N;L
16	C3ORF24	-	Bet v1-Like	General	-	R