Received 9 March 2023, accepted 20 March 2023, date of publication 29 March 2023, date of current version 5 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3262929

RESEARCH ARTICLE

Functional Annotation and Analysis of Genetic Variants

EMAD S. ABOUEL NASR[1] AND H. AL-MUBAID[2]

[1]Department of Industrial Engineering, College of Engineering, King Saud University, Riyadh 11421, Saudi Arabia
[2]Department of Computer Science, University of Houston—Clear Lake, Houston, TX 77062, USA

Corresponding author: Emad S. Abouel Nasr (eabdelghany@ksu.edu.sa)

This work was supported by the King Saud University, Riyadh, Saudi Arabia, through Researchers Supporting under Project RSP2023R164.

ABSTRACT The process of functional annotation of human genetic variants is important for some significant biomedical tasks like understanding functional genomics and gene-disease associations. Moreover, genetic variants that lead to pathogenic mutations can have harmful consequences which may lead to certain medical conditions or diseases. This paper investigates and presents two methodologies utilizing the Gene Ontology (GO) for functional annotation of genetic variants (and human gene mutations): (1) using the concept of term enrichment from the biological process taxonomy in the gene ontology, and (2) using the concept of least common subsumer (LCS) within the gene ontology tree. The second methodology induces the most significant and accurate functions for a given set of variants having one common aspect, e.g. specific disease. This method is based on the structure of the directed acyclic graph of GO to induce and identify the most significant LCS's to be used for functionally annotating the variants under investigation. We applied the methods on a large sets of genetic variants causing mutations. Our method can determine important functions and with significance level of $p<0.02$, the outcomes and results of the method are biologically significant. We found that certain ontology annotations are more related to mutations through certain genes compared with normal *bp* functions (biological process terms) in GO; for example, one of these bp functions is {GO:0031325; *positive regulation of cellular metabolic process*}. The results are basically important and suggest that a mutation can be annotated with functions from the gene ontology, e.g. biological process aspect, just like the genes. This outcome may contribute into a more complete understanding of mutation-gene-disease relationships, mutation pathogenicity, and understanding disease mechanisms.

INDEX TERMS Genetic variants, gene mutations, functional annotation.

I. INTRODUCTION

The study of genetic variants and human gene mutations is vital for several fields in bioinformatics like understanding the various genetic disorders, role of molecular processes, disease mechanisms, and gene-disease associations [1], [2], [3], [4], [5]. Essentially, a genetic variation is a change in the DNA sequence from the normal. Any change in the DNA sequence is called *genetic variation*, or simply *variation*; and that change may affect one gene, in which case it is called *mutation* [4], [6]. If the sequence change in the DNA is common in the population then we call it *polymorphism* [6], [7], [8]. The *Single Nucleotide Polymorphism*, or *SNP*, and is most common type of genetic variations, and is caused by only one change in one nucleotide in the DNA sequence [7]. Sometimes, the sequence change can be located in a coding region which may lead to a change in the expression of the gene [4]. In this paper we present a study and analysis of human genetic variations in the context of functional characterizations and annotations within the Gene Ontology (*GO*). The presented method finds the linkage between gene variants and gene functional annotations to determine and identify highly significant annotations, *i.e.,* functions, from the *biological process* taxonomy of the *GO* that characterize the target variants in the gene.

The associate editor coordinating the review of this manuscript and approving it for publication was Santosh Kumar.

Typically, when the DNA sequence is changed, such a change may impact a gene and cause some disease or medical condition [6], [7], [9]. Medical disorders and diseases caused from DNA sequence changes are collectively called *genetic disorders* [6], [7]. If the functionality of the gene is affected by a mutation in that gene, the resulting protein coded by the gene will be unable to work properly. On the other hand, a mutation is called *neutral*, or *silent*, mutation if it does not affect the gene function. For example, the mutation that does not alter the *amino acid* sequence of the coded protein is a *silent* mutation [6]. Genetic diseases, based on the size of the variation, are three major classes:–*single-gene*, –*chromosomal*, and –*multifactorial disorders* [4], [6]. Amongst the most common genetic diseases are *cystic fibrosis*, *sickle cell anemia*, and *homeochromatosis* [7].

The research work and projects on genetic variants and human gene mutations can be divided into the following groups: (1) Extracting mutation-gene-disease relationships from text and literature [5], [10]. (2) Mutation pathogenicity prediction and classification [2], [4]. (3) Study of mutations of a specific disease [4], [9], [11], [13]. (4) Mutation-disease association extraction from biomedical literature, mainly with NLP and machine learning approaches [10], [14], [15]. (5) Mutation functional analysis [2], [9], [16], [17], [18], [19].

The work in this paper focuses identifying specific *GO* annotations (functions) related to genetic variations by utilizing the functional gene annotations from *GO*. In other words, we investigate the linkages of gene mutations and gene functional annotations as shown in Figure 1 and Figure 2; as previously presented in [19]. Therefore, functional annotations have been assigned for variants from the biological process *bp* taxonomy of *GO*. For that, we presents two methodologies that utilize the gene ontology for functional annotation of genetic variants and human gene mutations as follows: (1) the first methodology uses the concept of term enrichment from the biological process, *bp*, taxonomy in the gene ontology, and (2) the second methodology uses the concept of least common subsumer *LCS* within the gene ontology tree. The second methodology induces the most significant and accurate functions for a given set of variants having one common aspect, *e.g.* specific disease. This methodology is based on the structure of the *directed acyclic graph* (DAG) of the gene ontology to identify and induce the most significant least common subsumers to be used for functionally annotating the variants under investigation. We conducted the evaluations with a number of experimental settings to (1) investigate the relationships between the mutations in a given gene g_x and the functional annotations of g_x from *bp* taxonomy, and (2) identify the significant *bp* functional annotations for gene mutations. Further, we identified several *bp* functions that if annotated (associated with) to some gene *g* then that gene (*g*) shows more mutation associations compared with other annotated genes with statistical significance ($p<0.05$). To our knowledge, there are no studies or projects in the bioinformatics literature for the annotation of mutations from the gene ontology.

II. BACKGROUND AND RELATED WORK

The demand for more extensive research to explain and identify the variations and mutations in the genetic sequences and associating (molecular or biological) functions with these mutations stems from the need for more complete understanding of human diseases and medical conditions [5], [9], [10], [11], [14], [15], [16]. Moreover, the need for studying and understanding the variations in the genetic sequences and their relationships with diseases was motivated by whole genome sequencing [5], [10], [16], [18]. Further, investigating genetic variants and mutations regarding their (biological) functions (or processes) is the least investigated dimension in mutation and genetic variants studies.

One of the public archives and freely available databases of variants and mutations is the *ClinVar* [2]. ClinVar contains human gene variations and phenotypes archive with supporting evidences for over 125,000 variants and over 200,000 submitted interpretations [2]. It stores germline and somatic variants of any type, size, or genomic location. The interpretations in ClinVar are mostly comprehensive and include structural variants that may consist of many genes (and involving over 26000 genes); for variants that affect a single gene, almost 5000 genes are represented in ClinVar [2].

In a comprehensive, and recent, genetic study, López-Urrutia et al. reported the existence of a large number of mutations that are still need to be further studied and analyzed for their functions and describes it as a 'very complex task [11]. In general, the research work in the domain of genetic variations and gene mutation date back to three decades ago [1], [2], [3], [6], [15]. We can categorize the research work and projects in this domain into the following five tasks:

– Mutation pathogenicity prediction and classification [2], [4].

– Extraction of mutation-disease associations from biomedical literature, mostly with Natural language processing, NLP, and machine learning techniques [10], [14], [15].

– Analysis of genetic variants and mutations in the context of specific diseases, e.g. Alzheimer, and Breast Cancer [4], [9], [11], [13].

– Mutation-gene-disease association extraction from biomedical text and literature [5], [10].

– Prediction and annotation of mutation functions and other (functional) analyses of mutations and genetic variations [2], [9], [16], [17], [18], [19], [30].

The proposed methods contribute to (and can be classified within) the functional analysis of mutations for the analysis and understanding of disease-mutation relationships. In a recent research, it has been found that most of the variants tend to have more than just one impact or a single biological consequence as it has commonly believed [13]. In analyzing an entire genome, within projects of Whole Genome Sequencing WGS, this fact appears the most.

In the past two decades, a large number of projects have been proposed for the extraction of mutation-disease associations or mutation-gene-disease triplets from the biomedical literature; see for example [1], [2], [3], [12], and [13].

Calabrese et al. developed a tool called SNPs&GO for predicting mutations using functional annotations from the GO [16], [17]. In [14], Doughty et al. developed and presented a system, called EMU, for mutation-disease extraction from PubMed abstracts and focused on prostate and breast cancer. The EMU tool uses a rule-based method to find all genes associated with the extracted mutations. The developers used the MutationFinder system to evaluate their EMU system; and MutationFinder is another mutations extraction tool [14]. They emphasize in their work that personalized medicine research is very important in cancer treatment, and mutation-disease relationship is crucial is this direction [14].

Alzheimer disease (AD) is one of the most studied diseases in the domain of genomic analysis and genetic variations analysis; and one study found that AD relates to about 30 million genetic variations (and that was one of the most comprehensive studies Butkiewicz et al. [13]). Also, they found that that a vast majority of the investigated variants (∼94%) have two or more impacts or biological consequences [13].

Wei et al. developed a tool, called GenNorm, for gene mutations extraction from biomedical literature [10], [15]. Furthermore, the tmVar tool is another text-mining software for extracting mutations from text [9]. Moreover, one of the most comprehensive resources of cancer mutations is the Catalogue Of Somatic Mutations In Cancer, COSMIC [22]. Generally, COSMIC, is considered the largest resource (or one of the largest) in the world for somatic mutations of human cancer [22]. Kordopati et al. presented a system and a tool for building links between diseases and mutations, called DES-Mutation [9]. The DES-Mutation also links mutations from 27 databases and dictionaries based on terms and phrases enriched in published mutation-disease literature [9]. They analyzed information on mutations from over 400000 Medline articles retrieved by searching for mutations [9].

DiMex, as reported in [5], is a text mining system for extracting triplets of mutations, genes and diseases {m, g, d} from abstracts. It can also extract other mutation information that can be useful for searching and displaying results more efficiently [5]. In another recent study, Butkiewicz et al. [13] proposed a methodology for functional annotation of genomic variants in the context of Alzheimer disease sequencing project [13]. Moreover, a free web-based system for functional annotation of genes and transcriptomic data is presented by Araujo et al. and is based on sequence homology search [18].

III. METHODOLOGY 1: FUNCTIONAL ANNOTATION USING TERM/FUNCTION ENRICHMENT

The main contributions of this work are: (i) analyzing and exploring the relationship between genetic variants and gene functional annotations from the *bp* taxonomy; and (ii) annotating genetic variants with functional *bp* annotations from GO.

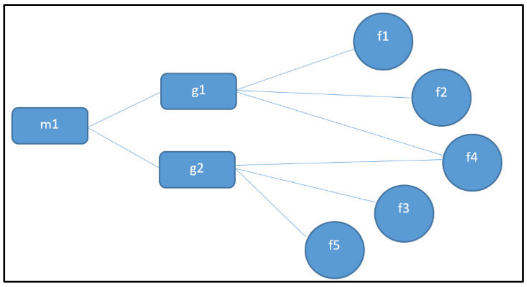


FIGURE 1. Illustration of mutation, gene, function relationships. Mutation m_1 is associated with two genes g_1, g_2, and five functions $f_1 \ldots f_5$. Function f_4 is associated with both genes g_1 and g_2.

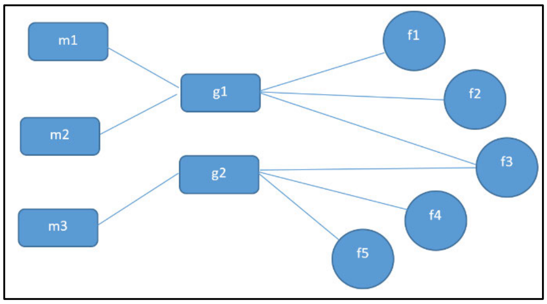


FIGURE 2. This diagram depicts the relationships among three mutations, two genes, and five functions. Both mutations m_1 and m_2 are related to gene g_1. Gene g_1 is related/annotated with three functions $f_1 \ldots f_3$. Also, both genes, g_1 and g_2, are annotated with function f_3.

A. DATA SOURCES AND PERLIMINARY ANALYSIS

We used the following two sources of mutation data: *ClinVar:* which is available within NCBI, and the *HGMD*: human gene mutation database, [1, 2, 3]. For gene-disease associations, we used the *Online Mendelian Inheritance in Man* OMIM [26]. For annotations, we utilized the *Gene Ontology* as it is the main source and the most commonly used database of functional annotations of all genetic data [17], [18], [23]. Moreover, the *GOA_human* database is used in this work for the functional annotations human genes [23]. {*note: GOA_human* includes and maintains more than 600,000 gene functional annotations).

B. GENETIC VARIANTS AND GENE ANNOTATIONS ANALYSIS

For investigating the relationships between genetic variants and functional annotations (from the *bp* taxonomy) we would like to examine the functional (*bp*) annotations of genetic variants by using *gene–bp* relationships with various types of variants as depicted in figures 1 and 2. An example of the relationship between a mutation and some (*bp*) functions is presented in Figure 1 where the mutation m_1 is associated with five functions $f_1 \ldots f_5$ through two genes g_1 and g_2. One can notice in this example that function f_4 is associated with both genes g_1 and g_2. An example with multiple mutations is depicted in Figure 2. This example in Figure 2, from our previous work [19], illustrates the relationships among 3 mutations, 2 genes and 5 functions where f_3 is in common among the three mutations.

–Analyzing the functional (bp) annotations of the genes with highest number of mutations: From *ClinVar*, we obtained the complete information and interpretations of more than 130000 variants associated with 7400 human genes. We found amongst these, around 3230 genes are associated with one variant, and more detail in the following:

Details and stats about genes and genetic variants from the *ClinVar* database.

Total number of variations examined: 134 820	
Number of genes associated with these variations: 7 448	
21 genes having 1000 or more variants	2693 genes associated with ≥5 variants
35 genes having 500 or more variants	355 " " " 4 variants
122 genes having 100 or more variants	458 " " " 3 variants
1651 genes having 10 or more variants	709 " " " 2 variants
	3233 " " " 1 variant

In an initial step for understanding the relationships between genetic variants and gene functional annotations, we investigated the genes with the most genetic variants against all human genome functional annotations *GOA* by utilizing only the biological process (*bp*) and the molecular function (*mf*) annotations. From our previous studies in [19] and [30], we extracted the genes with the highest number of variants association from the *ClinVar* database [2], as shown in Table 1. We started with each gene associated with at least 1000 mutations for a total of 21 genes [2], [19]. For example, the *Breast cancer gene 2* (BRCA2 gene Id: 675) found to be associated with 9631 variations; see Table 1. Then we obtained from *GOA* [23] both the *bp* and *mf* functional annotations of these 21 genes; Table 1. On average these (highest variation-populated) 21 genes have 42.7 *bp* functional annotations each (and 49.2 *mf* functional annotations each). Compared to the average of all annotated human genes, these findings are statistically significant with $p<0.01$ {notice that all annotated human genes have an average of 8.6 *bp* (and 9.2 *mf*) annotations as shown in Figure 3 (also reported in Table 11 in Appendix) These results, therefore, prove that when associated with more mutations, a gene tends to have more (*bp* and *mf*) functional annotations; with this difference are significant ($p<0.01$) compared with all other *bp* annotated genes. By considering only the *bp* annotation, there are in total 17717 human genes having *bp* annotations (also in Table 11).

–Analyzing the genetic variants of genes having highest number of functional annotations: We analyzed the genes with the greatest number of *bp* annotations in the *GOA* as shown in Table 2. Then we obtained all the variations of each of these genes from *ClinVar* and the results are in shown in Table 2.

From these result in Table 2, a strong correlation is observed and detected between number of genetic variations and number of functional annotations of genes. That is, these 20 genes (in Table 2) are having significantly higher than average variations per gene (144 vs 26 avg. variations per gene) of the reported variations on all genes (i.e., from all variations reported in *ClinVar* and *HGCD* we found that on average a gene is associated with ∼26 variants [2], [9]). Such genes with significantly higher number of bp annotations are associated with significantly more variations than normal genes tends to be more multifunctional genes or disease-related genes [19], [25], [30].

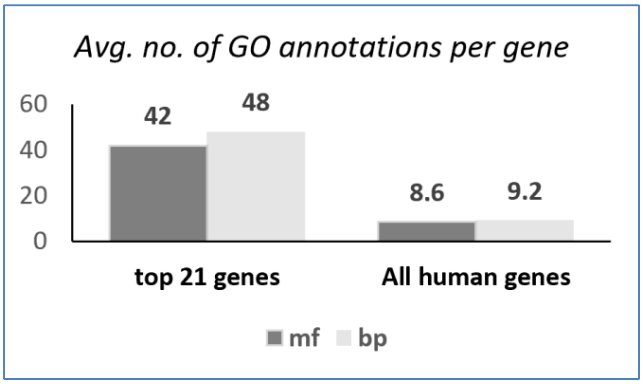


FIGURE 3. Illustrating the difference in the total number of annotations per gene for the entire human genome (all human genes) versus the top 21 genes.

–Relating functional annotations with genetic variations for eight sets of genes: we collected and compiled 8 gene sets having same number of variants. We started with only 2 variants per gene in the first set, then 10 variants per gene in the second set, and so on until the last set with ≥90 variants per gene. Each set contains 100 genes, approximately. The results are shown in Table 3 and illustrated in Figure 4. This evaluation is to prove that as the number of variations per gene increases, the number of functional annotations increases (directly proportional). This evaluation involved ∼800 genes, ∼18000 variations, ∼17000 bp annotations, and ∼15500 mf annotations. These results are statistically significant ($p<0.01$). Moreover, the *Pearson* correlation with the number of variations and bp and mf annotations are 0.977 and 0.982 respectively.

-Functional Annotation of Genetic Variants Using Term/Function Enrichment

In this evaluation, the biological process taxonomy is employed to determine and assign functional annotations to genetic variants as follows. For a given variant v_i from *ClinVar*, we wish to determine the most significant bp annotation (function) from bp taxonomy in GO. For that, we extracted from *ClinVar* the set of all genes associated with v_i (we call it set G_i). Next, we identified the enriched bp functions for all the genes in the set G_i. We used the *GoTermFinder* tool [8] on the genes G_i to identify GO bp term enrichment with 1% significant level. For example, the variant/mutation *m.1555A>G* (ClinVar Id: 9628) is a single nucleotide variant associated with two genes MT-RNR1 and MT-CO1. It is reported in ClinVar with one of the deafness conditions (i.e., Deafness, nonsyndromic sensorineural, mitochondrial). In the literature, this medical condition was reported and confirmed by a study in [24] to be associated with purine regulation. As shown in Table 4; having our method (independently) annotated this (m.1555A>G) mutation with {purine-containing compound metabolic process; GO:0072521} indicates that the annotation of variant by our method is biologically significant [19]. In another example, the variant NM_001672.2(ASIP):c.*25A>G (*ClinVar* Id: 9308) is a single nucleotide variant in ClinVar reported as associated with the two genes AHCY and ASIP. In the *GOA* database, these two genes are associated with 14 *bp* annotations from the bp taxonomy. GOTermFinder identified one bp annotation (Behavior: GO:0007610) which is biologically significant for this mutation; see Table 5. Moreover, with this methodology, more functional bp annotations for ClinVar mutations have been identified and are shown in Table 6. These results are biologically significant for functional bp annotations of variants from ClinVar.

TABLE 1. **The top 21 genes with the most number of variations (column 2). Each associated with at least 1000 mutations; *src: ClinVar*.**

Gene	Variations	Annotations {bp+mf}	bp	mf
BRCA2	9631	81	17	64
BRCA1	6900	183	75	108
APC	4907	124	42	82
TSC2	3617	54	27	27
MSH6	3339	64	33	31
MSH2	2915	84	37	47
FBN1	2563	49	18	31
LDLR	2464	74	46	28
MLH1	2335	53	22	31
PMS2	1713	21	9	12
RYR1	1699	43	23	20
RYR2	1530	86	53	33
TSC1	1468	59	44	15
MYH7	1302	27	17	10
MYBPC3	1298	28	16	12
PTEN	1296	177	119	58
TP53	1294	416	119	297
SCN5A	1267	106	68	38
TTN	1136	78	35	43
KCNH2	1101	55	37	18
STK11	1055	68	40	28
	average	91.9	42.7	49.2

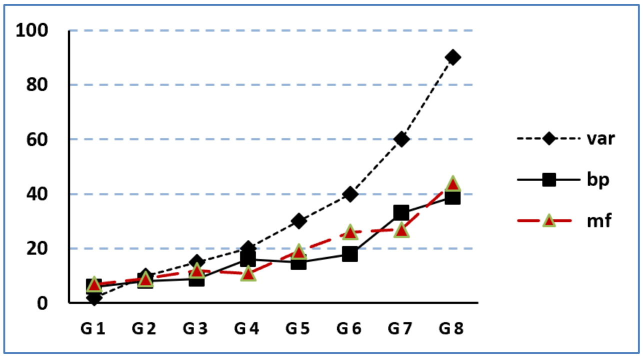


FIGURE 4. **Illustration of mean number of annotations from *bp* and *mf* for each of the eight gene sets.**

TABLE 2. **The top 20 human genes having the highest number of *bp* annotations along with the number of variants per gene. That is, for each gene from the top 20 genes having the most *bp* annotations, we extracted number of variations (from ClinVar/hgmd).**

Gene	BP Annotations	Variations
TGFB1	254	28
BMP4	202	54
TNF	194	11
AKT1	169	156
NOTCH1	166	980
PRKN	164	208
CTNNB1	161	122
WNT5A	156	107
VEGFA	150	13
SIRT1	148	14
SHH	144	154
FGFR2	136	243
BCL2	133	73
SFRP1	130	38
APOE	129	44
BMP2	129	41
SOX9	128	145
PARK7	126	61
TGFB2	124	240
APP	121	143
Average:	**153.2**	**143.8**

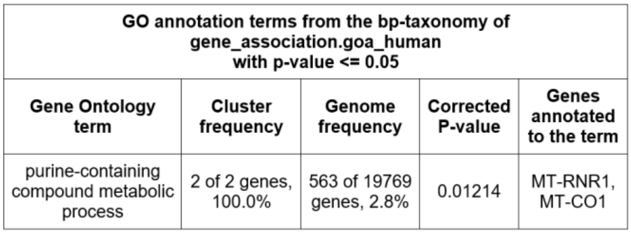
TABLE 3. **Details of the eight gene groups; Each group contains 100 genes approximately (*note:* all genes in each group have the same number of variations).**

Gene group	Var	bp	mf
G1	2	6	7
G2	10	8	9
G3	15	9	12
G4	20	16	11
G5	30	15	19
G6	40	18	26
G7	60	33	27
G8	90	39	44

IV. METHODOLOGY 2: FUNCTIONAL ANNOTATION USING TREE STRUCTURE AND LCS

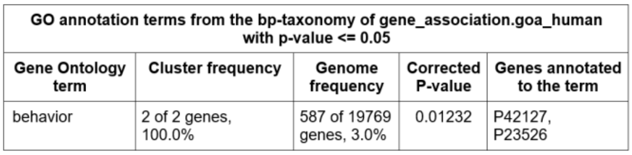
This methodology focuses on genetic variations that lead to and cause mutations which can be pathogenic and have harmful consequences. The methodology induces and assigns one or more functions to a given mutation or set of mutations having one aspect *in common* such as a disease. Such a method like this is significant for some related problems like mutation pathogenicity prediction and gene-disease association discovery [2], [7], [9], [16], [17] as explained earlier. Basically, the method relies on the *bp* taxonomy [23] to assign functions (functional annotations) to a given mutation as follows.

TABLE 4. **Identifying the GO *bp* annotation term with *p<0.02* for the mutation m.1555A>G (*ClinVar variation Id: 9628*) using the *GOTermFinder*.**



GO annotation terms from the bp-taxonomy of gene_association.goa_human with p-value <= 0.05				
Gene Ontology term	Cluster frequency	Genome frequency	Corrected P-value	Genes annotated to the term
purine-containing compound metabolic process	2 of 2 genes, 100.0%	563 of 19769 genes, 2.8%	0.01214	MT-RNR1, MT-CO1

TABLE 5. **The GOTermFinder identified the gene ontology *bp* annotation term with *p<0.02* for the mutation{*ClinVar Id 9308, NM_001672.3(ASIP):c.*25A>G*}.**



GO annotation terms from the bp-taxonomy of gene_association.goa_human with p-value <= 0.05				
Gene Ontology term	Cluster frequency	Genome frequency	Corrected P-value	Genes annotated to the term
behavior	2 of 2 genes, 100.0%	587 of 19769 genes, 3.0%	0.01232	P42127, P23526

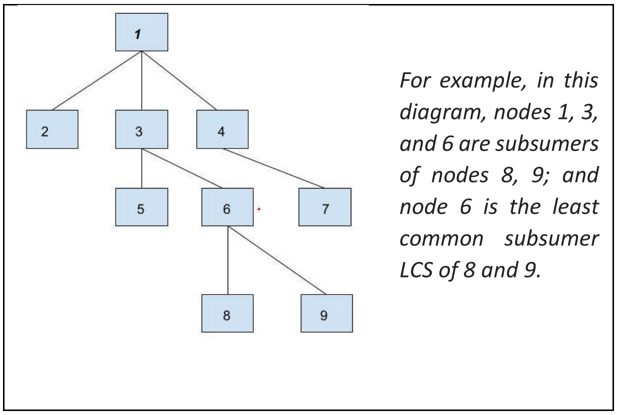
A. FUNCTIONAL ANNOTATION WITH LCS

Given a set S_m of mutations having one aspect in common, we wish to determine the most significant *bp* function (or functions) relevant to these mutations in set S_m. We used the *Gene Ontology* (GO) since it is the most widely utilized (and most comprehensive) resource for functionally annotation of genomic data, *e.g.* genes, RNA [17], [23]. *GO* is comprised of three taxonomies *Cellular Component* (cc), *Biological Process* (bp), and *Molecular Function* (mf); and typically the *bp* is the mostly used for functional annotation [17]. The mutations in the set S_m, have one aspect in common such as some disease D_i. That is, in the set S_m, each mutation $m_i \in S_m$ is found to be (and reported as) *pathogenic* for some disease D_i. Next, we extract the set G_i of all the genes of each mutation m_i: $G_i = \{g_x | g_x \text{ is a gene with mutation } m_i\}$. As mentioned earlier, we utilized the *GOA_human* data for the *bp* functions of genes (functional annotation information). That is, from the *GOA_human* database [23] we obtain all the functional annotations, from the *bp* taxonomy only, for all genes in G_i and we call it set *P*. That is

$$P = \{p_{1,\ldots,}p_n\} \tag{1}$$

such that p_j is a *bp* function annotating one gene in set G_i. Now, we would like to determine the most significant such p_j function from among all the functions in the set *P*. Then, from the *bp* taxonomy, which is a (tree-like) *Directed Acyclic Graph*, DAG, we want to identify each *least common subsumer* (*LCS*) that subsumes two or more functions $p_i \in P$.

As shown in Fig. 5, each function (or *bp* term) is a node in the *bp* taxonomy tree; see figures 5 and 7, [23]. One of the goals of the proposed method is to identify the most significant least common subsumers (LCS's) that subsume most function nodes associated with a (target) mutation or a set of mutations. In other words, we are interested in identifying the most significant *LCS* as it will be the *most-likely* cause of the disease associated with the mutations under investigation; *i.e.* the set S_m. For each *LCS* we calculate the *significance level* as follows:



For example, in this diagram, nodes 1, 3, and 6 are subsumers of nodes 8, 9; and node 6 is the least common subsumer LCS of 8 and 9.

Algorithm 1 Functional Annotation of Mutations

```
Purpose: to annotate the mutations $m_i \in M$ with the bp function(s) in the output set $L$.
Input: A set of $n$ mutations $m_1, \ldots, m_n$ with a common aspect $A1$,
and a significance threshold $thr(default\ thr = 1.0)$.
    {*** note: the aspect A1 can be a disease or a medical condition,
e.g. Alzheimer disease AD}
Output: Set $L$ of the $k$ most significant LCSs from the
            bp taxonomy that represents the significant
            functions of the input mutations: $L = \{p_1, .., p_k\}$.
Algorithm:
1. Let $G$ be the set of all genes of the input $n$ mutations:
        $G = \{g_1, \ldots, g_n\}$.
2. Let $L$ be the set of significant functions (initially empty $L = \emptyset$)
3. For every gene $g_i \in G$ extract all the bp annotations of $g_i$ from
        GOA_Human database into the set P:   $P = \{p_1, \ldots, p_k\}$
4. For every function node $p_i \in P$ do the following:
        4.1 Calculate the significance $s(p_i)$ of node $p_i$ according to
            Equation (2).
        4.2 If $s(p_i) \geq thr$ then add the function node $p_i$ to the output
set
            $L: L = L \cup \{p_i\}$
5. Output $L$.
```

$$\text{Significance level of node } n : s(n) = \frac{\#\ of\ subsumed\ terms}{total\ \#\ of\ function\ terms} \tag{2}$$

where # denotes the number (or count); for example, *# of subsumed terms* (in eq. 2) is the count of terms submsumed by node *n*. Therefore, a node in the (*bp taxonomy*) tree gets more significance as it subsumes more functions. Thus, *significance level* of the *root* is 1 as it subsumes all the nodes in the tree. Also notice that the nominator and denominator in Equation (2) are based on the functions from the set *P* in Equation (1) above. A small section of the *bp* taxonomy tree (from the *GO*) is shown in Fig. 5. For example, Fig. 6 shows a hierarchy of four significance levels and 7 *bp* functions, $f_1 \ldots f_7$, from the *bp* taxonomy. These functions f_i are divided

TABLE 6. A sample of variations from *ClinVar* along with their annotated *bp* terms from the Gene Ontology. The functional annotations (column 4) are biologically significant.

Variation	Variation Id	Genes	GO term (*biological process*)	Corrected P-value
NC_000001.10:g.(?_236902582)_(237205889_?)dup	463531	ACTN2, MTR, RYR2, MT1HL1	response to inorganic substance cellular response to nitrogen compound	0.01236 0.01873
NM_000463.2(UGT1A1):c.145C>T (p.Gln49Ter)	12278	UGT1A, UGT1A10, UGT1A8, UGT1A7, UGT1A6, UGT1A5, UGT1A9, UGT1A4, UGT1A1, UGT1A3	flavonoid glucuronidation xenobiotic glucuronidation flavonoid metabolic process	6.46e-31 3.55e-29 3.23e-27
GRCh37/hg19 2q32.1-32.2(chr2:189322984-189910304)x3	562601	COL3A1, COL5A2, GULP1, DIRC1	collagen fibril organization cellular response to amino acid stimul	0.00202 0.00351
NC_000002.12:g.(?_165090130)_(166228992_?)del	530654	GALNT3, SCN1A, SCN2A, SCN3A, SCN9A, TTC21B, CSRNP3, LOC100506124, TTC21B-AS1, SCN1A-AS1, LOC102724058	sodium ion transmembrane transport	8.17e-05
NC_000002.12:g.(?_165874735)_(166311776_?)del	471073	SCN1A, SCN9A, TTC21B, TTC21B-AS1, SCN1A-AS1, LOC102724058	neuronal action potential membrane depolarization during action potential	0.00188 0.00211
m.1555A>G	9628	MT-CO1, MT-RNR1	purine-containing compound metabolic process	0.01214
NC_000018.9:g.76841645_78077248del1235604	242602	NFATC1, CTDP1, TXNL4A, ADNP2, KCNG2, RBFA, PQLC1, PARD6G, ATP9B, HSBP1L1	phospholipid translocation lipid translocation	0.00830 0.00971
NM_001672.2(ASIP):c.*25A>G	9308	AHCY, ASIP	behavior	0.01232

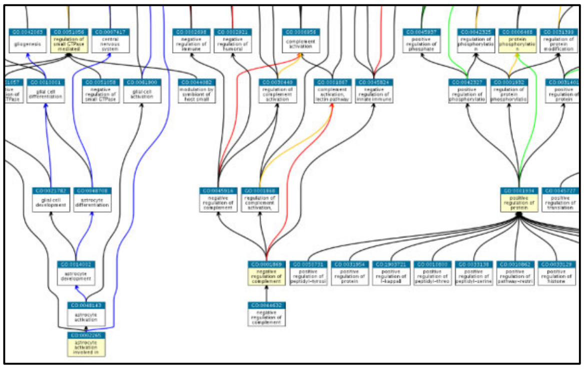


FIGURE 5. Illustration of the *bp* function terms from (a portion of) the Biological Process aspect of the Gene Ontology (GO); source: *EMBL QuickGO.*

into four significance levels {0.4, 0.6, 0.8, 1.0}. The two functions f_1 and f_2 are the most significant, Fig. 6. Moreover, Fig. 7 presents an illustrations with eight functions (8 nodes). In part (a) the annotated functions (nodes) are 5, 6, and 8 and are highlighted in yellow; whereas part (b) shows the results (*significance levels*) in three significance levels {1.0, 0.33, and 0.66}; for example, The *metabolic process* $GO : 0008152$ (node 3) subsumes two nodes (nodes 5 and 8) out of three and so its significance is 0.66 (= 2/3 as shown in the figure. Algorithm 1 presents the details of the process as follows

This algorithm summarizes the technique for identifying the most accurate and relevant functions for annotating a mutation or a set of mutations associated with a given disease or medical condition. Notice that the algorithm can identify more than one *LCS* if *thr* (the threshold) is less than 1.0; and for $thr = 1.0$ there will be only one LCS that subsumes all the functions in the set P (line 3 of the algorithm).

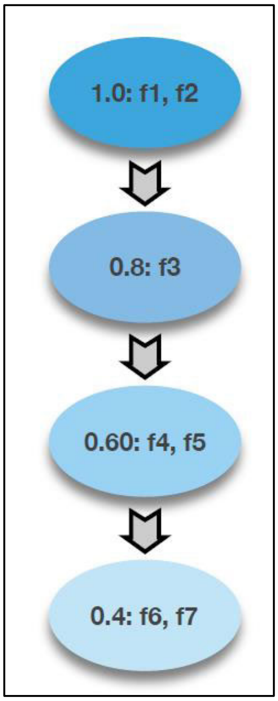
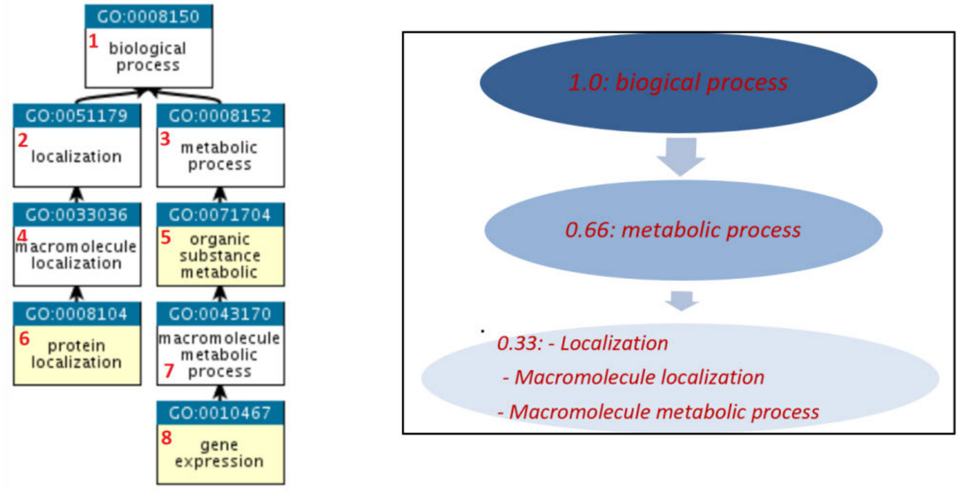


FIGURE 6. This figure illustrates four levels of significance with sets of functions from *lcs's*.

B. EXPERIMENTS AND RESULTS WITH LCS

Firstly, the mutation data in our experiments were retrieved from *ClinVar* hosted within *NCBI* [1], [2], [3]. As in the first methodology, we used *GO* as the main functional annotations source in this approach [17], [18], [23]. For gene-disease association information we relied on *OMIM* [26]; and we used the *GOA_human* [23] database for all



(a) (b)

FIGURE 7. A simple illustration of methodology 2. (a) A section of the bp taxonomy with eight nodes. The functions/nodes of interest are highlighted with the yellow nodes (i.e., *nodes* 5, 6, 8). Node 1 (GO:0008150 *biological process*) is the lcs for all three of them with significance level = 1.0; Node 3 subsumes two nodes (5 and 8) so, its sig. level = 2/3 = 0.66. Node 2 has sig. level = 1/3 = 0.33 because it subsumes only one node (6). Both nodes 4 and 7 are also like node 2 with sig. level = 0.33. This way we have three levels of significance {0.33, 0.66, 1.0}. (b) Node 1 (GO:0008150; *biological process*) is the most significant *lcs* at the default threshold thr=1.0. At *thr* ≥ 0.66, we get two nodes: node 5 (GO:0008152; *metabolic process*) and node 1.

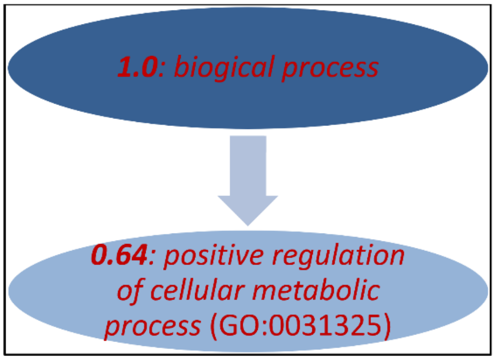

1.0: biogical process
0.64: positive regulation of cellular metabolic process (GO:0031325)

FIGURE 8. Illustration of the most significant *LCS* of all 96 *bp* functional annotations of the mutations of Alzheimer disease.

TABLE 7. The mutations we studied from *ClinVar*; more than 155,000 mutations involving more than 18,000 genes.

Type	# of Mutations	# of Genes	# of Associated Conditions
pathogenic	82321	9270	7473
benign	72070	9044	2572

TABLE 8. The top genes associated with Alzheimer disease; source *OMIM Morbidmap*.

Gene symbol	Gene Id	MIM Id
HFE	3077	613609
NOS3	4846	163729
PLAU	5328	191840
A2M	4353	103950
MPO	2	606989
APP	351	104760

functional annotations of human genes [23]. Table 7 contains the mutations, genes, and the associated conditions as were retrieved from the *ClinVar* [1], [2], [3]. The attention in our experiments was focused on the *pathogenic* mutations (more than 82000 pathogenic mutations involving more than 9200 genes), Table 7.

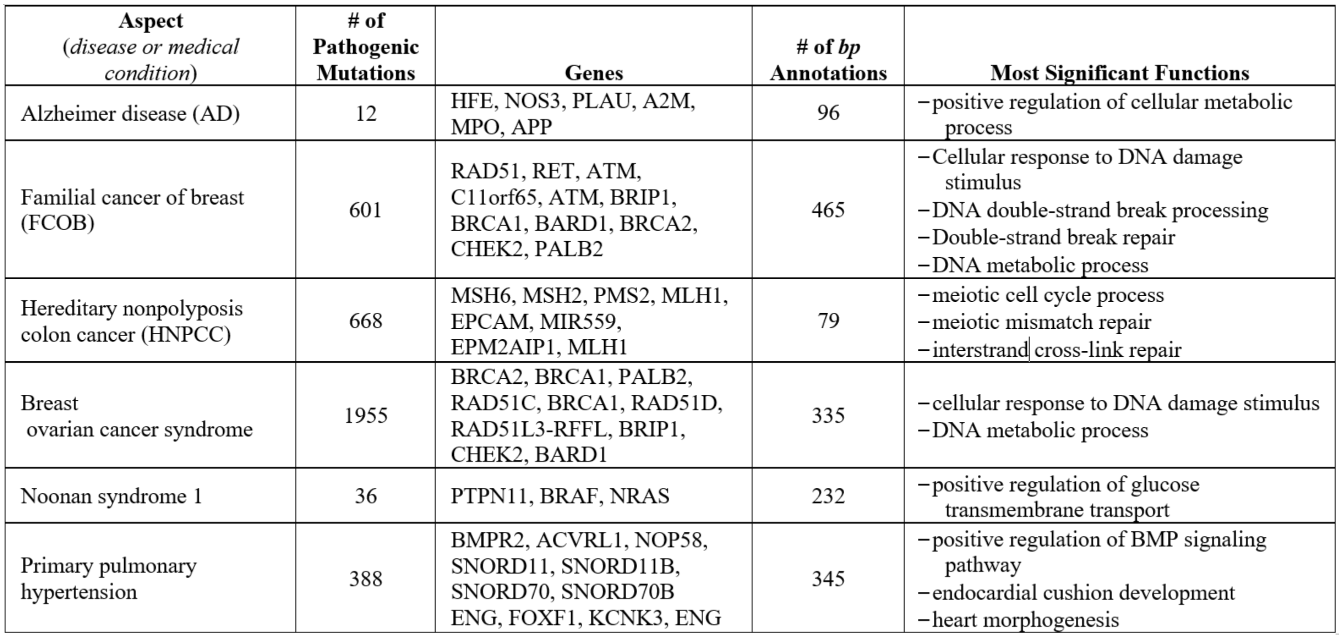
1) ALZHEIMER DISEASE (AD)

In examining the Alzheimer Disease, we found six genes associated with it in *OMIM* (Table 8). We determined a total of 96 *bp* functions annotating the six *AD* genes as per the databases of *OMIM* and the *GOA_human*. Considering the subtree (from the *bp* taxonomy) the includes all these (96) *bp* functions, we found that the *bp* function {GO:0031325; *positive regulation of cellular metabolic process*} to be the most significant *LCS* for all these 96 *bp* annotations/functions with *significance level* = 0.64 and *thr* = 0.60 as shown in Fig. 8. For *thr* = 1.0 we seek a *bp* function that subsumes all these 96 which found to be only the *root* {GO:0008150; *biological process*}.

2) HEREDITARY NONPOLYPOSIS COLON CANCER(HNPCC)

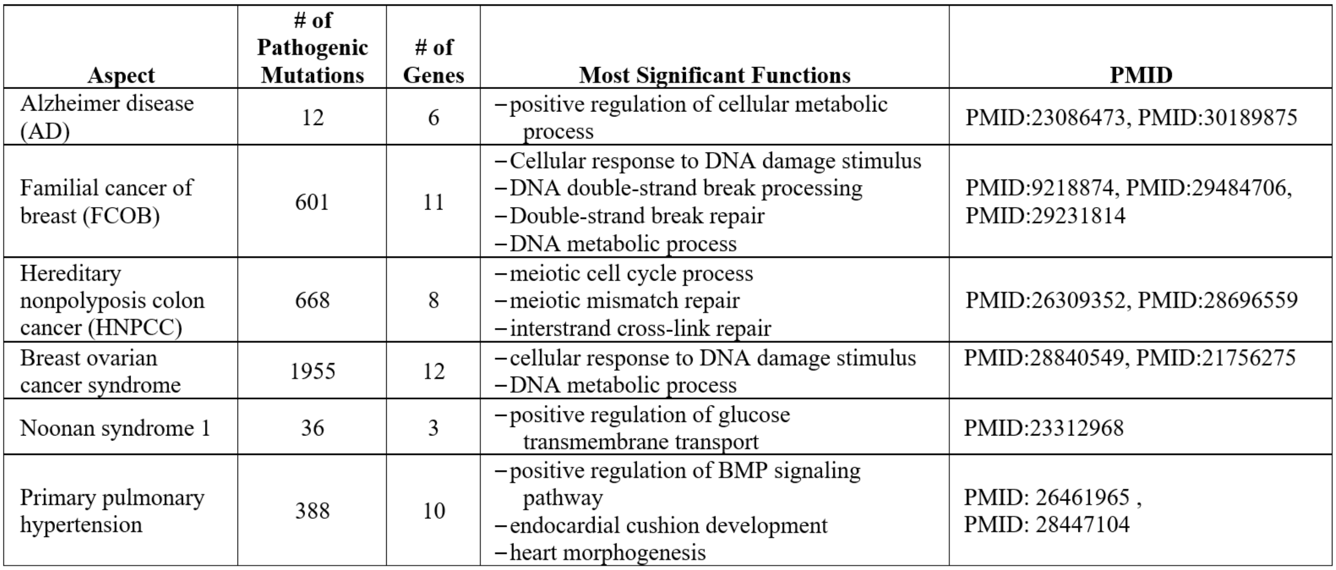
The most significant functions that characterize of mutations reported as pathogenic for *HNPCC* disease found by our method are three functions {*meiotic cell cycle process*, *meiotic mismatch repair*, and *interstrand cross-link repair*}. Table 9 contains these results along with the eight genes reported for these mutations; also in Table 9.

TABLE 9. The most significant *bp* functions for six sets of mutations with certain diseases or medical conditions.



Aspect (*disease or medical condition*)	# of Pathogenic Mutations	Genes	# of *bp* Annotations	Most Significant Functions
Alzheimer disease (AD)	12	HFE, NOS3, PLAU, A2M, MPO, APP	96	–positive regulation of cellular metabolic process
Familial cancer of breast (FCOB)	601	RAD51, RET, ATM, C11orf65, ATM, BRIP1, BRCA1, BARD1, BRCA2, CHEK2, PALB2	465	–Cellular response to DNA damage stimulus –DNA double-strand break processing –Double-strand break repair –DNA metabolic process
Hereditary nonpolyposis colon cancer (HNPCC)	668	MSH6, MSH2, PMS2, MLH1, EPCAM, MIR559, EPM2AIP1, MLH1	79	–meiotic cell cycle process –meiotic mismatch repair –interstrand cross-link repair
Breast ovarian cancer syndrome	1955	BRCA2, BRCA1, PALB2, RAD51C, BRCA1, RAD51D, RAD51L3-RFFL, BRIP1, CHEK2, BARD1	335	–cellular response to DNA damage stimulus –DNA metabolic process
Noonan syndrome 1	36	PTPN11, BRAF, NRAS	232	–positive regulation of glucose transmembrane transport
Primary pulmonary hypertension	388	BMPR2, ACVRL1, NOP58, SNORD11, SNORD11B, SNORD70, SNORD70B ENG, FOXF1, KCNK3, ENG	345	–positive regulation of BMP signaling pathway –endocardial cushion development –heart morphogenesis

TABLE 10. The results of the method with the outcomes verified from the biomedical literature.



Aspect	# of Pathogenic Mutations	# of Genes	Most Significant Functions	PMID
Alzheimer disease (AD)	12	6	–positive regulation of cellular metabolic process	PMID:23086473, PMID:30189875
Familial cancer of breast (FCOB)	601	11	–Cellular response to DNA damage stimulus –DNA double-strand break processing –Double-strand break repair –DNA metabolic process	PMID:9218874, PMID:29484706, PMID:29231814
Hereditary nonpolyposis colon cancer (HNPCC)	668	8	–meiotic cell cycle process –meiotic mismatch repair –interstrand cross-link repair	PMID:26309352, PMID:28696559
Breast ovarian cancer syndrome	1955	12	–cellular response to DNA damage stimulus –DNA metabolic process	PMID:28840549, PMID:21756275
Noonan syndrome 1	36	3	–positive regulation of glucose transmembrane transport	PMID:23312968
Primary pulmonary hypertension	388	10	–positive regulation of BMP signaling pathway –endocardial cushion development –heart morphogenesis	PMID: 26461965 , PMID: 28447104

3) FAMILIAL CANCER OF BREAST (FCOB)

There are *4059* variants total reported with the 35 genes of the *Familial cancer of breast* disease (FCOB). Of the 35 disease genes, 11 genes are reported with variants pathogenic to FMOC, and these 11 genes have 601 pathogenic variants reported. The first 3 genes are: PALB2, CHEK2, PALB2, and are related to (and reported with) pathogenic variants; namely with 232, 126, and 72 pathogenic variants respectively. The functions that are essentially the most significant *LCS's* are reported in Table 9. These functions are induced from the *bp* subtree subsuming more than 400 function nodes to achieve majority threshold with $thr \geq 0.66$.

4) PRIMARY PULMONARY HYPERTENSION

The 388 mutations of this disease were examined by our method with one aspect: Primary pulmonary hypertension. As shown in Table 9, there are 11 genes reported with these pathogenic mutation. Our method was able to induce three functions characterizing these mutations: *positive regulation of BMP signaling pathway*, *heart morphogenesis*, and *endocardial cushion development*.

5) NOONAN SYNDROME 1

We investigated with our proposed method 36 pathogenic mutations reported with one shared aspect, *Noonan* *syndrome 1*. The method was able to annotate these mutations with one function: *positive regulation of glucose transmembrane transport*; and this fact is reported and validated in the (biomedical) literature in PubMed (PMID:23010278).

6) BREAST OVARIAN CANCER SYNDROME

Two functions characterize the mutations that are reported pathogenic to the breast and ovarian cancer and are shown in Table 9. The table also contains the 10 genes with which the mutations are reported, Table 9.

C. DISCUSSION AND ANALYSIS

By assigning one *bp* function term (or more) for gene mutations from the *bp* taxonomy, we essentially characterize their functional consequences. This will basically lead to an increased and a more complete understanding of the mutation pathogenicity and their relationship with genes and diseases. We identified published reports and research articles in *PubMed* to verify the results as shown in Table 10. Also, we were able to verify the results in Table 9 by examining the biomedical literature seeking for experimental evidences and functional information of these mutations. These results indicate that we can further our knowledge about pathogenicity of mutations by investigating the specific functional consequences of these mutations.

V. DISCUSSION AND CONCLUSIONS

The annotation and functional characterization of all human genes has been an active area of research in the biomedical domain for the past two decades. Then recently, more interest is being shifted towards functional annotation of gene mutations and genetic variations [13], [16], [18], [20]. As biological sequences, or part of sequences, genetic variants and mutations have roles in the total molecular functions of the sequences. Further, the functional characterizations of genetic variants are important for research projects related to genetic diseases and functional genomics [18]. In general, a human gene with multiple functions tends to have more mutations associated with it than normal. This is intuitive since a *multifunctional* gene is most likely a disease-gene, and thus a highly studied gene, and therefore, more mutations reported for it than the normal [25]. However, this relationship has never been studied within the context of genetic variations. Thus, the results and outcomes of this work can connect mutations with gene functions computationally using the publicly available genes, the gene ontology, mutations, and gene variation databases [1], [2], [3], [7], [17], [23], [26]. This research produced a number of relationships connecting mutations with functional pb annotations with statistical significance ($p < 0.05$); see for example experiment 3. For example, it has been reported in Section IV above that the *FCOB* disease is annotated with more than 450 bp annotation (Table 9) and one of the statistically significant functional annotations is the *Cellular Response to DNA damage stimulus*. Searching the literature shows that it has been already reported that FCOB (aka familial breast cancer) is intimately related to DNA damage response [27]. Another example, it has been found that *Noonan syndrome 1* makes patients more sensitive to glucose which is shown by our method as in Table 10 [28].

Trembath and Harrison reported that in Pulmonary arterial hypertension (PAH), raised pulmonary artery pressures lead to progressive right heart hypertrophy and eventual failure [29]. This validates the finding of the proposed method (last row in Table 10) that the *bp* annotation term *Heart mophongenesis* (GO:0003007) is significant for disease *Primary Pulmonary Hypertension* (PPH) [29].

VI. CONCLUSION

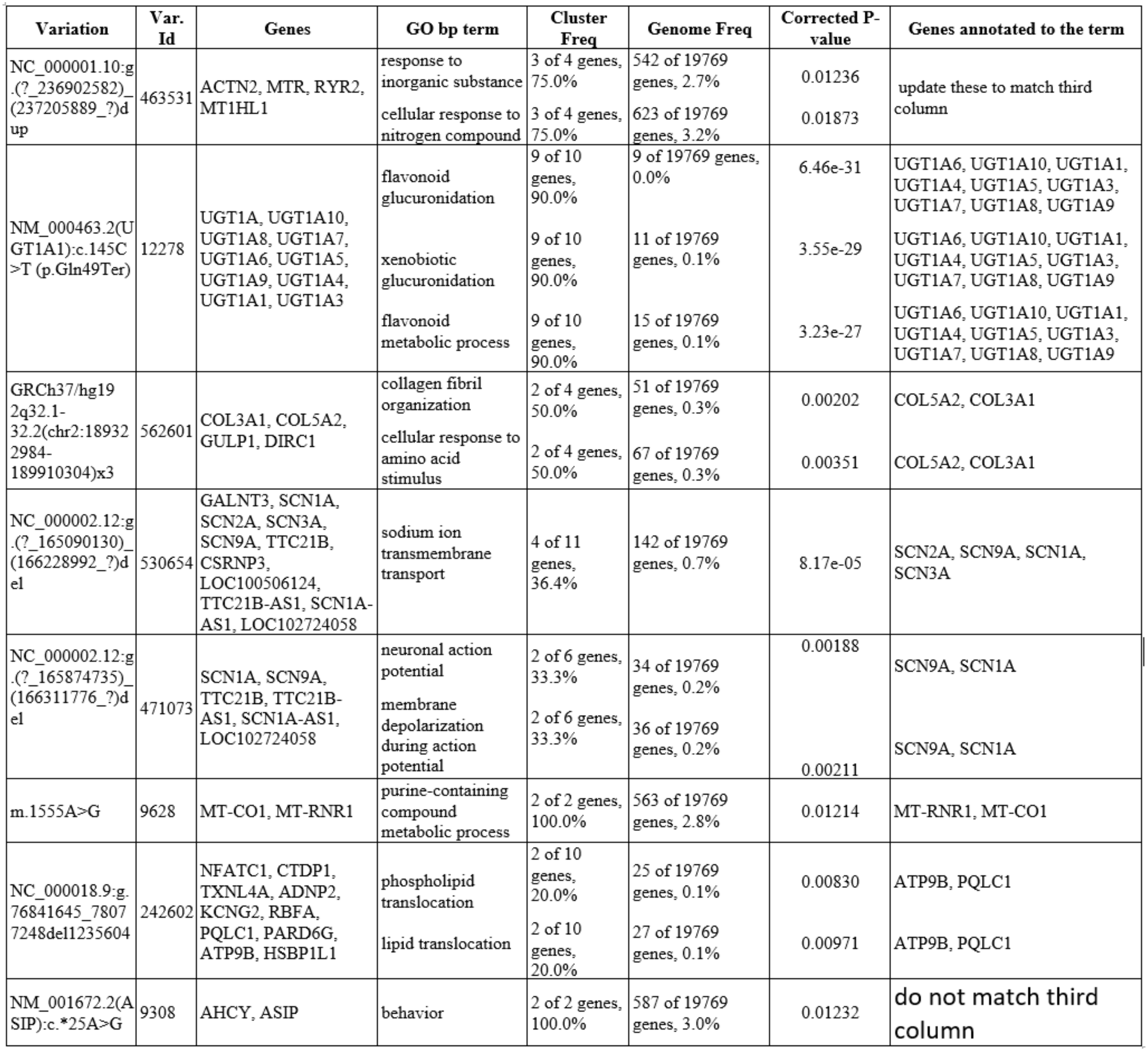
In this paper, we presented an analysis of genetic variants in relationship with biological process annotations from the *Gene Ontology*; also explored the relationships between genetic variations and bp annotations for some diseases and genes. We also reported results of the analysis which are considered interesting and important. Moreover, a method is presented for functional annotations of mutations and genetic variants from the *bp* taxonomy of the gene ontology. To the best of our knowledge there are no such work or projects in the literature that analyzes mutations for functional annotations from the gene ontology (we only found the work of Capriotti et al [17] which uses gene annotations as features in the prediction of disease association of a given variation [17]).

In the overall results and experimental analysis of this work, we found a clear connection in relating mutations with functional annotations, and the relationship is statistically significant especially in experiment 3 (where we used *Man-Whitey* statistical significance test with $p < 0.0001$). Hence,

TABLE 11. Details of the GOA human database [17].

<table>
<tr><th colspan="2">GOA Human</th></tr>
<tr><td colspan="2">Total: 313,274 total <i>bp</i> and <i>mf</i> annotations (bp: biological process, mf: molecular function)</td></tr>
<tr><td>BP:
152,396 total <i>bp</i> annotations
17,717 unique genes (having bp annotations)
Avg bp annotations per gene: 152396/17717 = 8.6 bp/gene
<i>In Bp:</i>
- 2529 genes with only 1 bp annotations
- 2335 genes with only 2 bp annotations</td><td>MF:
160,879 total <i>mf</i> annotations
17,639 unique genes (having mf annotations)
Avg mf annotations per gene: 160879/17639 = 9.2 mf/gene
<i>In Mf:</i>
-2520 genes with only 1 mf annotations
-2474 genes with only 2 mf annotations</td></tr>
</table>

TABLE 12. The full details of the variations reported in Table 6 with their annotated bp terms from the Gene Ontology. ****note: this table is only for reviewing and may not be included in the final version of this paper.*



Variation	Var. Id	Genes	GO bp term	Cluster Freq	Genome Freq	Corrected P-value	Genes annotated to the term
NC_000001.10:g.(?_236902582)_(237205889_?)dup	463531	ACTN2, MTR, RYR2, MT1HL1	response to inorganic substance	3 of 4 genes, 75.0%	542 of 19769 genes, 2.7%	0.01236	update these to match third column
			cellular response to nitrogen compound	3 of 4 genes, 75.0%	623 of 19769 genes, 3.2%	0.01873	
NM_000463.2(UGT1A1):c.145C>T (p.Gln49Ter)	12278	UGT1A, UGT1A10, UGT1A8, UGT1A7, UGT1A6, UGT1A5, UGT1A9, UGT1A4, UGT1A1, UGT1A3	flavonoid glucuronidation	9 of 10 genes, 90.0%	9 of 19769 genes, 0.0%	6.46e-31	UGT1A6, UGT1A10, UGT1A1, UGT1A4, UGT1A5, UGT1A3, UGT1A7, UGT1A8, UGT1A9
			xenobiotic glucuronidation	9 of 10 genes, 90.0%	11 of 19769 genes, 0.1%	3.55e-29	UGT1A6, UGT1A10, UGT1A1, UGT1A4, UGT1A5, UGT1A3, UGT1A7, UGT1A8, UGT1A9
			flavonoid metabolic process	9 of 10 genes, 90.0%	15 of 19769 genes, 0.1%	3.23e-27	UGT1A6, UGT1A10, UGT1A1, UGT1A4, UGT1A5, UGT1A3, UGT1A7, UGT1A8, UGT1A9
GRCh37/hg19 2q32.1-32.2(chr2:189322984-189910304)x3	562601	COL3A1, COL5A2, GULP1, DIRC1	collagen fibril organization	2 of 4 genes, 50.0%	51 of 19769 genes, 0.3%	0.00202	COL5A2, COL3A1
			cellular response to amino acid stimulus	2 of 4 genes, 50.0%	67 of 19769 genes, 0.3%	0.00351	COL5A2, COL3A1
NC_000002.12:g.(?_165090130)_(166228992_?)del	530654	GALNT3, SCN1A, SCN2A, SCN3A, SCN9A, TTC21B, CSRNP3, LOC100506124, TTC21B-AS1, SCN1A-AS1, LOC102724058	sodium ion transmembrane transport	4 of 11 genes, 36.4%	142 of 19769 genes, 0.7%	8.17e-05	SCN2A, SCN9A, SCN1A, SCN3A
NC_000002.12:g.(?_165874735)_(166311776_?)del	471073	SCN1A, SCN9A, TTC21B, TTC21B-AS1, SCN1A-AS1, LOC102724058	neuronal action potential	2 of 6 genes, 33.3%	34 of 19769 genes, 0.2%	0.00188	SCN9A, SCN1A
			membrane depolarization during action potential	2 of 6 genes, 33.3%	36 of 19769 genes, 0.2%	0.00211	SCN9A, SCN1A
m.1555A>G	9628	MT-CO1, MT-RNR1	purine-containing compound metabolic process	2 of 2 genes, 100.0%	563 of 19769 genes, 2.8%	0.01214	MT-RNR1, MT-CO1
NC_000018.9:g.76841645_78077248del1235604	242602	NFATC1, CTDP1, TXNL4A, ADNP2, KCNG2, RBFA, PQLC1, PARD6G, ATP9B, HSBP1L1	phospholipid translocation	2 of 10 genes, 20.0%	25 of 19769 genes, 0.1%	0.00830	ATP9B, PQLC1
			lipid translocation	2 of 10 genes, 20.0%	27 of 19769 genes, 0.1%	0.00971	ATP9B, PQLC1
NM_001672.2(ASIP):c.*25A>G	9308	AHCY, ASIP	behavior	2 of 2 genes, 100.0%	587 of 19769 genes, 3.0%	0.01232	do not match third column

this work contributes to annotation of gene variants using the gene ontology as the source of functional information [23]. We also showed that genes with the most mutations are also among the highest in number of functional *bp* annotations associated with them. Further, genes with the most *bp* functions are also among the highest in having variations associated with them. We also showed that for a given variation, the assigned *bp* functions and annotations, by the proposed method, are biologically significant. These outcomes and results collectively shall augment our knowledge and understanding of gene variants and gene-mutation- disease relationships and mechanisms.

APPENDIX

See Tables 11 and 12.

ACKNOWLEDGMENT

The authors extend their appreciation to King Saud University for funding this work through Researchers Supporting Project number (RSP2023R164), King Saud University, Riyadh, Saudi Arabia.

REFERENCES

[1] M. Krawczak, "The human gene mutation database," *Trends Genet.*, vol. 13, no. 3, pp. 121–122, Mar. 1997.

[2] M. J. Landrum, J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, and D. R. Maglott, "ClinVar: Public archive of interpretations of clinically relevant variants," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D862–D868, Jan. 2016.

[3] P. D. Stenson, E. V. Ball, M. Mort, A. D. Phillips, J. A. Shiel, N. S. T. Thomas, S. Abeysinghe, M. Krawczak, and D. N. Cooper, "Human gene mutation database (HGMD): 2003 update," *Hum. Mutation*, vol. 21, no. 6, pp. 577–581, Jun. 2003.

[4] N. Mahdieh and B. Rabbani, ''An overview of mutation detection methods in genetic disorders,'' *Iranian J. Pediatrics*, vol. 23, no. 4, pp. 375–388, 2013.

[5] A. S. M. A. Mahmood, T.-J. Wu, R. Mazumder, and K. Vijay-Shanker, ''DiMeX: A text mining system for mutation-disease association extraction,'' *PLoS ONE*, vol. 11, no. 4, Apr. 2016, Art. no. e0152725.

[6] The New York-Mid-Atlantic Consortium for Genetic and Newborn Screening Services, *Understanding Genetics: A New York, Mid-Atlantic Guide for Patients and Health Professionals*. Washington, DC, USA: Genetic Alliance, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK115568/

[7] U.S. National Library of Medicine. (Sep. 2018). *Genetic Home Reference GHR*. [Online]. Available: https://ghr.nlm.nih.gov/

[8] (2018). *GOTermFinder*. [Online]. Available: https://go.princeton.edu/cgi-bin/GOTermFinder

[9] V. Kordopati, A. Salhi, R. Razali, A. Radovanovic, F. Tifratene, M. Uludag, Y. Li, A. Bokhari, A. AlSaieedi, A. Bin Raies, C. Van Neste, M. Essack, and V. B. Bajic, ''DES-mutation: System for exploring links of mutations and diseases,'' *Sci. Rep.*, vol. 8, no. 1, Sep. 2018, Art. no. 13359.

[10] K. Opap and N. Mulder, ''Recent advances in predicting gene–disease associations,'' *F1000Res. J.*, vol. 6, p. 578, Apr. 2017.

[11] E. López-Urrutia, V. Salazar-Rojas, L. Brito-Elías, M. Coca-González, J. Silva-García, D. Sánchez-Marín, A. D. Campos-Parra, and C. Pérez-Plasencia, ''BRCA mutations: Is everything said?'' *Breast Cancer Res. Treatment*, vol. 173, no. 1, pp. 49–54, Jan. 2019.

[12] J. D. Burger et al., ''Hybrid curation of gene–mutation relations combining automated extraction and crowdsourcing. Database,'' *J. Biol. Databases Curation*, vol. 2014, no. 1, pp. 1–13, 2014.

[13] M. Butkiewicz, E. E. Blue, Y. Y. Leung, X. Jian, E. Marcora, A. E. Renton, A. Kuzma, L.-S. Wang, D. C. Koboldt, J. L. Haines, and W. S. Bush, ''Functional annotation of genomic variants in studies of late-onset Alzheimer's disease,'' *Bioinformatics*, vol. 34, no. 16, pp. 2724–2731, Aug. 2018.

[14] E. Doughty, A. Kertesz-Farkas, O. Bodenreider, G. Thompson, A. Adadey, T. Peterson, and M. G. Kann, ''Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature,'' *Bioinformatics*, vol. 27, no. 3, pp. 408–415, Feb. 2011.

[15] C.-H. Wei and H.-Y. Kao, ''Cross-species gene normalization by species inference,'' *BMC Bioinf.*, vol. 12, no. S8, pp. 1–11, Dec. 2011.

[16] R. Calabrese, E. Capriotti, P. Fariselli, P. L. Martelli, and R. Casadio, ''Functional annotations improve the predictive score of human disease-related mutations in proteins,'' *Hum. Mutation*, vol. 30, no. 8, pp. 1237–1244, 2009.

[17] E. Capriotti, P. L. Martelli, P. Fariselli, and R. Casadio, ''Blind prediction of deleterious amino acid variations with SNPs&GO,'' *Hum. Mutation*, vol. 38, no. 9, pp. 1064–1071, Sep. 2017.

[18] F. A. Araujo, D. Barh, A. Silva, L. Guimarães, and R. T. J. Ramos, ''GO FEAT: A rapid web-based functional annotation tool for genomic and transcriptomic data,'' *Sci. Rep.*, vol. 8, no. 1, p. 1794, Jan. 2018.

[19] H. Al-Mubaid, ''Analysis of gene variants for functional annotations,'' in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Siena, Italy, Jul. 2019, pp. 1–7, doi: 10.1109/CIBCB.2019.8791476.

[20] X. Liu, C. Wu, C. Li, and E. Boerwinkle, ''dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs,'' *Hum. Mutation*, vol. 37, no. 3, pp. 235–241, Mar. 2016.

[21] A. Martín-Navarro, A. Gaudioso-Simón, J. Álvarez-Jarreta, J. Montoya, E. Mayordomo, and E. Ruiz-Pesini, ''Machine learning classifier for identification of damaging missense mutations exclusive to human mitochondrial DNA-encoded polypeptides,'' *BMC Bioinf.*, vol. 18, no. 1, pp. 1–11, Dec. 2017.

[22] COSMIC. (2018). *The Catalogue of Somatic Mutations in Cancer*. [Online]. Available: https://cancer.sanger.ac.uk/cosmic/download

[23] (2019). *Gene Ontology Annotation (GOA) Database*. [Online]. Available: https://www.ebi.ac.uk/GOA

[24] L. X. Zhong, S. Kun, Q. Jing, C. Jing, and Y. Denise, ''Non-syndromic hearing loss and high-throughput strategies to decipher its genetic heterogeneity,'' *J. Otol.*, vol. 8, no. 1, pp. 6–24, Jun. 2013.

[25] H. Al-Mubaid, ''Gene multifunctionality scoring using gene ontology,'' *J. Bioinf. Comput. Biol.*, vol. 16, no. 5, Oct. 2018, Art. no. 1840018.

[26] *Online Mendelian Inheritance in Man (OMIM)*, McKusick-Nathans Inst. Genet. Med., Johns Hopkins Univ., Baltimore, MD, USA, 2018. [Online]. Available: https://www.omim.org/

[27] X. Zhu, T. Tian, M. Ruan, J. Rao, W. Yang, X. Cai, M. Sun, G. Qin, Z. Zhao, J. Wu, Z. Shao, R. Shui, and Z. Hu, ''Expression of DNA damage response proteins and associations with clinicopathologic characteristics in Chinese familial breast cancer patients with BRCA1/2 mutations,'' *J. Breast Cancer*, vol. 21, no. 3, p. 297, 2018, doi: 10.4048/jbc.2018.21.e38.

[28] T. Pakladok, Z. Hosseinzadeh, I. Alesutan, and F. Lang, ''Stimulation of the Na$^+$-coupled glucose transporter SGLT1 by B-RAF,'' *Biochem. Biophys. Res. Commun.*, vol. 427, no. 4, pp. 689–693, Nov. 2012, doi: 10.1016/j.bbrc.2012.09.062.

[29] R. C. Trembath and R. Harrison, ''Insights into the genetic and molecular basis of primary pulmonary hypertension,'' *Pediatric Res.*, vol. 53, no. 6, pp. 883–888, Jun. 2003, doi: 10.1203/01.PDR.0000061565.22500.E7.

[30] H. Al-Mubaid, ''Gene mutation analysis for functional annotations using graph heuristics,'' in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Siena, Italy, Jul. 2019, pp. 1–6, doi: 10.1109/CIBCB.2019.8791494.

EMAD S. ABOUEL NASR received the Ph.D. degree in industrial engineering from the University of Houston, TX, USA, in 2005. He is currently a Professor with the Industrial Engineering Department, College of Engineering, King Saud University, Saudi Arabia, and a Professor with the Mechanical Engineering Department, Faculty of Engineering, Helwan University, Egypt. His current research interests include CAD, CAM, rapid prototyping, advanced manufacturing systems, and collaborative engineering.

H. AL-MUBAID received the Ph.D. degree in computer science from The University of Texas at Dallas, in 2000. He is currently a Professor in computer science and computer information systems and the Program Chair of Computer Information Systems with the University of Houston—Clear Lake, TX, USA. His research interests include natural language processing and bioinformatics and include data mining, machine learning, and biomedical text mining. He is working to prove the virtue of the context-based approach to natural language processing (NLP). He developed several algorithms and systems to prove the point, including detecting and correcting context-based errors, word prediction, word classification, text categorization, biomedical term disambiguation, and document clustering. He received the Distinguished Dissertation Award for his Ph.D. degree, in 2000.