Fungal Genomes: Suffering with Functional Annotation Errors

Tapan Kumar Mohanta (nostoc.tapan@gmail.com)
Yeungnam University https://orcid.org/0000-0002-3196-7746
Abeer Hashem
King Saud University
Elsayed Fathi Abd_Allah
King Saud University
Ahmed AL Harrasi
University of Nizwa

Research article

Keywords: Fungal Genome, Annotation, Calcium signaling; Calcium Dependent Protein Kinase; Selenocysteine

Posted Date: June 8th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-32751/v1

Abstract

Background

The genome sequencing data are accumulating at a rapid pace, with the current genome sequence data of more than 5780 species being publicly available at the National Center for Biotechnology Information (NCBI) database alone. However, for the researcher communities to use these data, an error-free functional annotation report is a must.

Results

Analyses of the whole proteome sequence data of 689 fungal species (7.15 million protein sequences) to find the presence of functional annotation error in several species. Hence, calcium dependent protein kinases (CDPKs) and selenoproteins were targeted for the analysis as it is absent all across the fungi kingdom. The analyses revealed the presence of protein with the functional annotation name CDPK. InterproScan analysis revealed that, none of the protein sequences tagged with name “calcium dependent protein kinase” was found to encode calcium binding EF-hands at the regulatory domain. Similarly, none of a protein sequences with annotation name associated with “selenocysteine” was found to encode Sec (U) amino acid.

Conclusion

The presence of naming of such functional annotation errors in the fungal kingdom is raised a great concern and need to address it at the earliest possible time.

Background

The term “gene annotation” is associated with detailed information of a particular gene and its translated protein products [1]. Gene annotation of an unknown gene/protein is usually performed using homology-based sequence similarity, or reference-based annotation method [2–4]. This method is the most acceptable form of gene annotation, and prediction. Sometimes a minor modification in the homology-based annotation can mislead, more particularly so in the case of fungi. The computationally predicted annotated gene name can be subjected to error with false positive results.

The most important goals of the current genome sequencing projects are to foster the advancement in the universal availability of the genome, gene, and protein sequences of hundreds and thousands of organisms. This goal will allow us investigate and correlate the genomic and molecular basis of the functions and their evolutions. These are the pre-requisite to answer any biological question. There are several genome databases present in the public domain to facilitate research in various aspects of genome, gene or proteins. A few of them are deposited at the National Center for Biotechnology

Information (NCBI), UniProt/SwissProt [5], Ensembl [6] and others. These databases often annotate the sequences by computationally predicted protein functions. However, information regarding the functional aspects of a gene/protein are available in sufficient numbers. However, there is no guarantee regarding the accuracy of the predicted annotation. In fact, the rapid accumulation of sequence data at the scale and breadth hardly contain any experimental validation. The computationally annotated sequences can be misannotated profoundly which may in turn mislead the researchers. Therefore, the propagation of errors on the day-to-day basis must be addressed at the earliest. The finding of genes and genome annotation in bacterial kingdom is comparatively easy when compared to that in eukaryotes, considering that > 90% of the bacterial genome encodes protein coding sequences. The gene finding tool reads the sequence in 6 possible reading frames (3 forwards and 3 reverse), resulting in highly accurate genes and amino acids in the correct sequence. However, in the eukaryotic system (including plants, animals, and fungi) gene finding and annotation process is difficult owing to the presence of non-coding intron sequences and the fact that genes are placed far apart from each other. In addition, the presence of large genome in the eukaryotic system makes it challenging to assemble and annotate the genome. Therefore, a less accurate automated gene annotation method is preferred for the assembly and annotation of the genes. Presently, the genome sequence data are constantly being added, resulting in rapid accumulation of data at an astounding pace. Although these data are very promising for the experimental validation and discovery of novel potential beneficial product, their use may if the sequences/genes are supplied with false positive/error prone annotation.

It is well known that fungi do not encode the “calcium dependent protein kinase” (CDPK) gene family in its genome. In addition, fungi also do not encode the selenocysteine/selenoprotein [7] in its proteome. Therefore, target was made to find the CDPK proteins and selenocysteine/selenoprotein in the proteome of the fungi.

Results

To understand the possibilities of the presence CDPK genes/proteins in the fungi, the annotated proteome files of 689 fungal species were downloaded from the NCBI. The data file constituted approximately 7.15 million protein sequences of the species across the Fungal Kingdom, from Ascomycota, Basidiomycota, Blastocladiomycota, Chytridiomycota, Glomeromycota, Microsporidia, Mucoromycota, Neocallimastigomycota, Opisthokonta, and Zoopagomycota. The species were found to encode 17 to 32854 proteins per species. From the 689 studied fungal species, most belonged to the Ascomycota (67.63%) phylum followed by the Basidiomycota (21.62%) and Microsporidia (3.77%). We searched for the presence of CDPK proteins in the proteome files of these 689 species. The search resulted in the finding of 521 protein sequences associated with the annotation term “calcium/calmodulin-dependent protein kinase” in 197 species (Supplementary File 1). Later, all of the 521 protein sequences with annotation term “calcium/calmodulin-dependent protein kinase” were scanned with the Scanprosite [8] and MEME suite [9] to identify the presence of kinase and regulatory domain (EF-hand domain). Analyses revealed the presence of N-terminal domain and kinase domain in

all the protein sequences, whereas none of the single protein sequence was found to contain the auto-inhibitory domain and the regulatory domain with 4 calcium binding EF-hands in it.

A similar contradictory result was noted regarding the presence of selenocysteine amino acid encoding machinery in the fungi. When search was made, at least 134 protein sequences were found with the functional annotation name associated with “selenocysteine” in 112 species from the 689 studied species (Supplementary File 2). When the search was made to find the presence of Sec (U) amino acid in the protein sequences, none of the sequence was found to contain U amino acid (Fig. 1).

Discussion

The CDPKs were characterized by the presence of N-terminal domain, kinase domain, an auto-inhibitory domain, and a regulatory domain [10, 11]. The regulatory domain is characterized by the presence of 4 calcium-binding EF-hands [12, 13]. The EF-hands present in the regulatory domain of the CDPKs are conserved and contain D-x-D conserved amino acid at the 14th and 16th position, which are responsible for binding of Ca^{2+} ions [14]. In addition, the CDPKs contain the N-terminal palmytoylation and myristoylation sites [14]. Hence, it was important to find the presence of all 4 domains in the CDPKs and N-terminal signal sequences of palmitoylation and myristoylation sites. However, N-terminal signal sequences and regulatory domain were not found in these sequences, which raised the question of, how the genes/proteins were annotated as calcium/calmodulin-dependent protein kinase in the complete absence of calcium-binding EF-hand domain in the proteins. Even, the homology-based annotation does not result in such a misleading report, when there is a complete lack of EF-hand containing regulatory domain. It is well known that fungi do not encode for the CDPK in its genome, whereas it is present in the plant and animal kingdom [10, 14, 15]. The calcium dependent protein kinases play diverse roles in plants and animals [10, 14, 15]. In plants, CDPK regulates growth, development, and biotic and abiotic stress tolerance [16–19]. In fungi, the calcium-signaling events are regulated by calmodulins, calcineurin B-like proteins, calmodulin-like proteins, calcineurin-responsive zinc finger transcription factor, Ca^{2+} ATPase, Ca^{2+}/H^{+} exchangers, high-affinity calcium system, low-affinity calcium system, transient receptor potential (TRP)-like calcium channels, and mitochondrial calcium channel [20, 21]. However, fungi do not encode for the *CDPK* gene family for the calcium signaling event. The basal cytoplasmic calcium level in fungi ranges from 50 to 200 nM and fungi store the maximum of their Ca^{2+} ion in the vacuole (approximately ~ 95%) and calmodulin, calcineurin B-like proteins and other bring the Ca^{2+} homeostasis irrespective of the presence of CDPKs [20]. The cellular Ca^{2+}-channels, cation/proton exchange regulate the filamentous growth in the fungi associated with cell division, hyphal tip growth, and hyphal branching [22]. The vacuolar Ca^{2+}-ATPase in the fungi is closely related to the plasma membrane Ca^{2+}-ATPase (PMCA) -type pump. The PMCA contain a cytosolic auto-inhibitory domain at the C-terminal end, which is relieved by the binding of calmodulin into it [20, 23]. The presence of auto-inhibitory domain in PMCA in fungi is functionally similar to the auto-inhibitory domain of the CDPK in the plant and animals. This may be a possible similar structural and functional unit of the fungi with regard to the CDPK in order to conduct calcium-signaling events; hence, the fungi do not encode the CDPKs.

Selenoproteins contain Sec amino acid, which is encoded by UGA codon. The proteins associated with the reactive oxygen species signaling machinery (glutathione peroxidase) contain Sec amino acid [24]. The presence of Sec amino acid has been reported in plants [25], animals, [26] and bacteria [27]. However, the presence of Sec amino acid has not been reported in the fungi [7]. Previously Mariotti et al., (2015) also reported that fungi do not contain Sec amino acid [7]. Therefore, the presence of ambiguous gene/protein annotation name with “selenocysteine” in the fungal genome/proteome is quite a concern for the researcher community. Therefore, it is important to provide an insight to the annotation strategy of the fungal genome, making it important for the researchers across the globe to consider it as a serious problem, requiring an immediate address. Previous report also reported the genome annotation error in bacteria [28, 29]. Several reasons may be accounted for the misannotation of the gene/protein sequences. However, the parameters for the placement of lower limit for the coding sequences can be one of the reasons. Another most possible reason may be the lack of “gold standard” of reference sequences. However, when the misannotation occur at the super-family level, that is CDPK, it is quite a concerning matter. A recent comparative study of the protein-coding and lncRNA transcript in the RefSeq and Gencode human gene database led to the finding that only 27.5% of the Genecode transcript had the exact match with the introns at the same position corresponding to the RefSeq genes [30]. Even after 19 years of continuous effort, the exon-intron boundary of the human genome is not yet settled. The problem in yeast and *Arabidopsis* is even worse than in human [30]. The advancement of RNA-sequencing could slightly rectify the problem, as the full-length transcript can directly align with the genome to reveal the exon-intron structure. The Mammalian Gene Collection that includes the gene of humans and a few other species could reduce the error rate through the RNA-seq approach. The modern annotation pipeline MAKER uses RNA-seq data and aligns with the database of other proteins and provides the correct annotation names. Although the RNA-seq has its own limitation, it remains a viable alternative to remove the error-prone annotation. The error in assembly can also lead to the errors in the annotation. Although the automated genome annotation process is good enough to cope with the pace for the sequencing of big and large number of genomes, any minor error in the existing annotation can directly propagate the error to the other species with immediate effect.

Conclusion

To overcome the annotation error in fungi, it is advisable to conduct the annotation on the basis of homology/orthology-based sequence similarity with a “gold standard”. In addition, co-localization of functionally-linked genes, experimental proteomics approach and similarity of motifs and sequence profiles search can add extra benefits towards error-free functional annotation of genes. The mitigation of false-positive annotation error that is propagating on the regular basis can create detrimental effect, more specifically in the case of pathway analysis. Therefore, we should be extra cautious toward the report of genome sequencing, assembly, and annotation.

Methods

All the annotated proteome files of the fungal proteomes from 689 species were downloaded from the publicly available NCBI (National Center for Biotechnology Information) database, which provided approximately 7.15 million protein sequences. All the annotated proteome files of fungi were downloaded available till 10th of March 2020. A simple search was made to find the protein sequences associated with the annotation name “selenocysteine/selenoprotein”. The resulted sequences were analyzed for the amino acid composition to find the selenocysteine amino acids and the results were noted. A full proteome file of *Chlamydomonas reinhardtii* was used as a reference to compare the presence and absence of Sec (U) amino acid. A similar study was conducted for the presence of *CDPK* gene family in the fungi. The protein sequences resulted with the annotation name “calcium dependent protein kinase” were subsequently analyzed in the Scanprosite [8] and MEME suite [9] in order to find the presence of calcium binding EF-hand domains using the default parameters. All the analysis was conducted on the Linux-based platform.

Abbreviations

CDPK: calcium dependent protein kinase, Sec: Selenocysteine, NCBI: National Center For Biotechnology Information, MEME: multiple Em of motif elicitation, N-terminal: amino-terminal, EF-hand: elongation factor hand, TRP: transient receptor potential, PMCA: plasma membrane Ca^{2+}-ATPase

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not Applicable

Availability of data material

All the studied data were taken from publicly available databases and data associated with the manuscript is provided in supplementary file.

Competing of interest

There is no competing of interest to declare

Funding

Not applicable

Author contribution

TKM: conceived the idea, collected and annotated the genome sequences, analysed and interpreted the data and drafted the manuscript, AA: revised the manuscript. All authors have read and approved the manuscript

Acknowledgement

Not available

References

1. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, et al. The Ensembl automatic gene annotation system. Genome Res. 2004;14:942–50.
2. Cai Y, Bork P. Homology-Based Gene Prediction Using Neural Nets. Anal Biochem. 1998;265:269–74.
3. Taher L, Rinner O, Garg S, Sczyrba A, Brudno M, Batzoglou S, et al. AGenDA: homology-based gene prediction. Bioinformatics. 2003;19:1575–7.
4. Meyer IM, Durbin R. Gene structure conservation aids similarity based gene prediction. Nucleic Acids Res. 2004;32:776–83.
5. Consortium TU. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2018;47:D506–15.
6. Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, et al. Ensembl variation resources. Database. 2018;2018.
7. Mariotti M, Guigó R. Evolution of selenophosphate synthetases: emergence and relocation of function through independent duplications and recurrent subfunctionalization Running Title : Phylogeny of selenophosphate synthetases Keywords : selenocysteine, gene duplication, sub. Genome Res. 2015;25:1256–67.
8. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucleic Acids Res. 2012;41:D344–7.
9. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202–8.
10. Mohanta TK, Mohanta N, Mohanta YK, Bae H. Genome-Wide Identification of Calcium Dependent Protein Kinase Gene Family in Plant Lineage Shows Presence of Novel D-x-D and D-E-L Motifs in EF-Hand Domain. Front Plant Sci. 2015;6:1146.
11. Mohanta TK, Mohanta N, Mohanta YK, Parida P, Bae H. Genome-wide identification of Calcineurin B-Like (CBL) gene family of plants reveals novel conserved motifs and evolutionary aspects in calcium signaling events. BMC Plant Biol. 2015;15:189.
12. Chandran V, Stollar EJ, Lindorff-Larsen K, Harper JF, Chazin WJ, Dobson CM, et al. Structure of the Regulatory Apparatus of a Calcium-dependent Protein Kinase (CDPK): A Novel Mode of Calmodulin-target Recognition. J Mol Biol. 2006;357:400–10.

13. Asai S, Ichikawa T, Nomura H, Kobayashi M, Kamiyoshihara Y, Mori H, et al. The variable domain of a plant calcium-dependent protein kinase (CDPK) confers subcellular localization and substrate recognition for NADPH oxidase. J Biol Chem. 2013/04/08. American Society for Biochemistry and Molecular Biology; 2013;288:14332–40.
14. Mohanta KT, Yadav D, Khan LA, Hashem A, Abd_Allah FE, Al-Harrasi A. Molecular Players of EF-hand Containing Calcium Signaling Event in Plants. Int J Mol Sci. 2019;20(6):1476.
15. Braun AP, Schulman H. The Multifunctional Calcium/Calmodulin-Dependent Protein Kinase: From Form to Function. Annu Rev Physiol Annual Reviews. 1995;57:417–45.
16. Mohanta TK, Sinha AK. Role of Calcium-Dependent Protein Kinases during Abiotic Stress Tolerance. Abiotic Stress Response Plants. 2016. p. 185–206.
17. Shi S, Li S, Asim M, Mao J, Xu D, Ullah Z, et al. The Arabidopsis Calcium-Dependent Protein Kinases (CDPKs) and Their Roles in Plant Growth Regulation and Abiotic Stress Responses. Int J Mol Sci MDPI. 2018;19:1900.
18. Mohanta TK, Occhipinti A, Atsbaha Zebelo S, Foti M, Fliegmann J, Bossi S, et al. Ginkgo biloba responds to herbivory by activating early signaling and direct defenses. PLoS One. 2012;7:e32822.
19. Gao X, Cox KL Jr, He P. Functions of Calcium-Dependent Protein Kinases in Plant Innate Immunity. Plants (Basel, Switzerland). MDPI; 2014;3:160–76.
20. Tisi Marco;Groppi,Silvia;Belotti,Fiorella. R. Calcium homeostasis and signaling in fungi and their relevance for pathogenicity of yeasts and filamentous fungi. AIMS Mol Sci. 3:505–49.
21. Liu S, Hou Y, Liu W, Lu C, Wang W, Sun S. Components of the Calcium-Calcineurin Signaling Pathway in Fungal Cells and Their Potential as Antifungal Targets. Eukaryot Cell. American Society for Microbiology Journals; 2015;14:324–34.
22. Benčina M, Bagar T, Lah L, Kraševec N. A comparative genomic analysis of calcium and proton signaling/homeostasis in Aspergillus species. Fungal Genet Biol. 2009;46:93–104. A.
23. Brini M, Calì T, Ottolini D, Carafoli E. The plasma membrane calcium pump in health and disease. FEBS J. John Wiley & Sons, Ltd; 2013;280:5385–97.
24. Borchert A, Kalms J, Roth SR, Rademacher M, Schmidt A, Holzhutter H-G, et al. Crystal structure and functional characterization of selenocysteine-containing glutathione peroxidase 4 suggests an alternative mechanism of peroxide reduction. Biochim Biophys Acta - Mol Cell Biol Lipids. 2018;1863:1095–107.
25. Mohanta TK, Khan AL, Hashem A, Abd_Allah EF, Al-Harrasi A. The Molecular Mass and Isoelectric Point of Plant Proteomes. BMC Genom. 2019;20:631.
26. Labunskyy VM, Hatfield DL, Gladyshev VN. Selenoproteins: molecular pathways and physiological roles. Physiol Rev American Physiological Society. 2014;94:739–77.
27. Zhang Y, Romero H, Salinas G, Gladyshev VN. Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and evolution of selenocysteine from redox active cysteine residues. Genome Biol. 2006/10/20. BioMed Central; 2006;7:R94–R94.

28. Brenner SE. Errors in genome annotation. Trends Genet. 1999;15:132–3.
29. Devos D, Valencia A. Intrinsic Errors in Genome Annotation. Trends Genet. 2001;17:429–31.
30. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. Genome Biology; 2019;19–21.

Figures

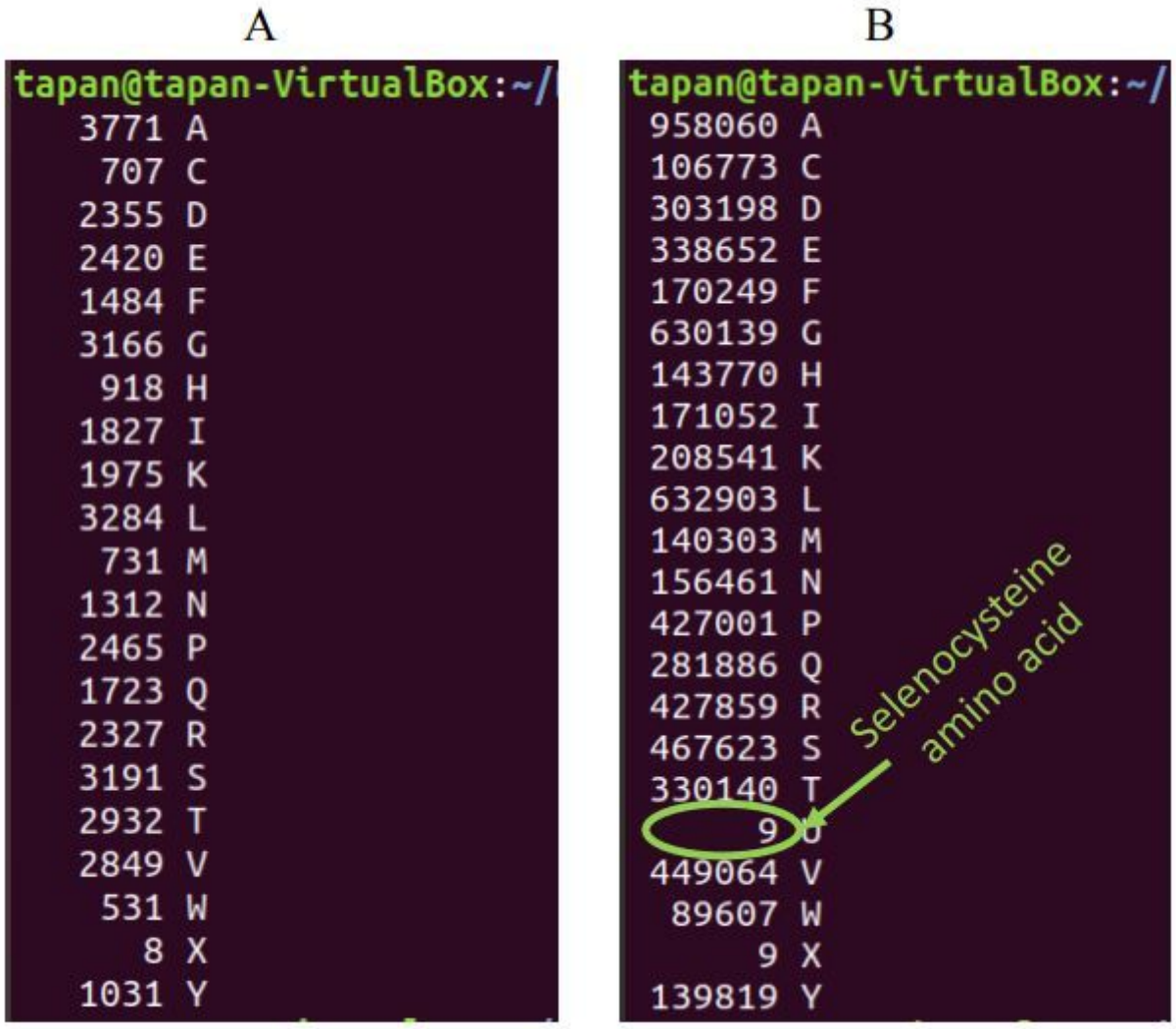

Figure 1

(A) Amino acid composition of the fungal protein sequences with annotation name associated with selenoprotein/selenocysteine. No selenocysteine amino acid was detected in the protein's sequences annotated with selenocysteine/selenoprotein name. (B) Amino acid composition of the Chlamydomonas reinhardtii proteome showing the presence of Sec (U) amino acid. The whole proteome of C. reinhardtii was taken as a reference to show the presence of U amino acid in it.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile2.txt
- SupplementaryFile1.xls