Abstract ID:- 3

Abstract Topic:- Genetic, genome and epigenome databases and resources

Abstract Title:- Mis-annotation profiling in DisGeNET v7.0.

Presenting author name :- Jatinder Sahota

Presenting author institute:- Guru Nanak Dev University

Co-author name:- Vasudha Sambyal

Co-author institute:-Guru Nanak Dev University

Aims:-Mis-annotations are being found frequently in many databases despite various measures to provide accurate data. These inaccuracies may increase the probability of false-positive associations from any downstream analyses involving the mis-annotated terms, requiring investigations into the mis-annotation profiles of different databases. The present study attempts to profile the frequency and categories of mis-annotations in DisGeNET v7.0, a database with a huge compilation of gene-disease association (GDA) annotations. Azoospermia, the most severe form of male infertility, has a strong genetic basis with hundreds of reported GDAs. Due to the heterogenous nature of male infertility, databases should make clear distinctions while reporting GDAs for the different male infertility sub-types. Therefore, the present study also analyzes the effects of mis-annotations on downstream analyses using the DisGeNET Azoospermia dataset.

Methods:- The abstracts of 5214 publications, with GDA annotations at DisGeNET v7.0, were manually screened from PubMed to profile the frequency and categories of mis-annotations in DisGeNET v7.0. Additionally, the GDA annotations for the DisGeNET Azoospermia dataset were assessed to check for both mis-annotated terms or GDAs. Cytoscape and Metascape were used to assess the effect of inclusion/exclusion of the mis-annotated terms on the enrichment profile of the DisGeNET Azoospermia network.

Results:- The screening of the publications referenced at DisGeNET revealed a high frequency of mis-annotations (45.84%), with a major fraction of the mis-annotations (99.96%) sourced to the BeFree dataset. Twenty-four distinct categories of mis-annotations were observed in DisGeNET, with the highest mis-annotation (21.17%) observed for genetic terminology. Furthermore, abbreviations containing separators were also mis-annotated as genes, with the highest mis-annotation (94.09%) observed for hyphenated abbreviations. The DisGeNET Azoospermia dataset consisted of 254 GDA annotations. Manual screening of the associated publications revealed the presence of both mis-annotated terms (10.24%) and mis-annotated GDA annotations (12.99%). Analysis of the azoospermia dataset using Cytoscape and Metascape showed that inclusion/exclusion of the mis-annotated genes led to the abolition of few protein-protein interaction networks, altering the enrichment profile of the Azoospermia dataset.

Conclusions:- Due to the high mis-annotation observed, the present study concludes that researchers should use the DisGeNET database with caution before arriving at any conclusion from their own datasets.

Keywords:- DisGeNET, Mis-Annotation, Gene, Data Mining, Database