EBioMedicine xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

A Systems Approach to Refine Disease Taxonomy by Integrating Phenotypic and Molecular Networks

Xuezhong Zhou [a,1], Lei Lei [b,1], Jun Liu [c,1], Arda Halu [d,1], Yingying Zhang [c,f], Bing Li [b,c], Zhili Guo [c,g], Guangming Liu [a], Changkai Sun [h,i,j,k], Joseph Loscalzo [e], Amitabh Sharma [d,e,*], Zhong Wang [c,**]

[a] School of Computer and Information Technology and Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing 100044, China
[b] Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, No.16, Nanxiaojie, Dongzhimennei, Dongcheng District, Beijing 100700, China
[c] Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, No.16, Nanxiaojie, Dongzhimennei, Dongcheng District, Beijing 100700, China
[d] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115, USA
[e] Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA
[f] Dongzhimen Hospital, Beijing University of Chinese Medicine, No.5 Haiyuncang, Dongcheng District, Beijing 100700,China
[g] Jiaxing Traditional Chinese Medicine Affiliated Hospital of Zhejiang Chinese Medical University, No. 1501, Zhongshan East Road, Jiaxing, Zhejiang 314000, China
[h] School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, 116024, China
[i] Research Center for the Control Engineering of Translational Precision Medicine, Dalian University of Technology, Dalian 116024, China
[j] State Key Laboratory of Fine Chemicals, Dalian R&D Center for Stem Cell and Tissue Engineering, Dalian University of Technology, Dalian 116024, China
[k] Liaoning Provincial Key Laboratory of Cerebral Diseases, Institute for Brain Disorders, Dalian Medical University, Dalian 116044, China

ARTICLE INFO

Article history:
Received 27 November 2017
Received in revised form 14 March 2018
Accepted 3 April 2018
Available online xxxx

Keywords:
Disease taxonomy
Network medicine
Disease phenotypes
Molecular profiles
Precision medicine

ABSTRACT

The International Classification of Diseases (ICD) relies on clinical features and lags behind the current understanding of the molecular specificity of disease pathobiology, necessitating approaches that incorporate growing biomedical data for classifying diseases to meet the needs of precision medicine. Our analysis revealed that the heterogeneous molecular diversity of disease chapters and the blurred boundary between disease categories in ICD should be further investigated. Here, we propose a new classification of diseases (NCD) by developing an algorithm that predicts the additional categories of a disease by integrating multiple networks consisting of disease phenotypes and their molecular profiles. With statistical validations from phenotype-genotype associations and interactome networks, we demonstrate that NCD improves disease specificity owing to its overlapping categories and polyhierarchical structure. Furthermore, NCD captures the molecular diversity of diseases and defines clearer boundaries in terms of both phenotypic similarity and molecular associations, establishing a rational strategy to reform disease taxonomy.

1. Introduction

Disease taxonomy plays an important role in defining the diagnosis, treatment, and mechanisms of human diseases. The principle of the current clinical disease taxonomies, in particular the International Classification of Diseases (ICD), goes back to the work of William Farr in the nineteenth century and is primarily derived from the differentiation of clinical features (e.g. symptoms and micro-examination of diseased tissues and cells) (Council et al., 2011). Despite its extensive clinical use, this classification system lacks the depth required for precision medicine with the limitations of its rigid hierarchical structure and, moreover, it does not exploit the rapidly expanding molecular insights of disease phenotypes. For example, many diseases (e.g. cancer, chronic inflammatory diseases) in the current disease taxonomies have either high genetic heterogeneity (Bianchini et al., 2016; McClellan and King, 2010) or manifestation diversity (Arostegui et al., 2014; Jeste and Geschwind, 2014; Mannino, 2002), which give little basis for tailoring treatment to a patient's pathophysiology. Furthermore, disease comorbidities (Hu et al., 2016; Lee et al., 2008; Hidalgo et al., 2009), temporal disease trajectories (Jensen et al., 2014) in clinical populations, various molecular relationships between disease-associated cellular components and their connections in the interactome (Blair et al., 2013; Goh et al., 2007; Barabasi et al., 2011; Rzhetsky et al., 2007; Zhou et al., 2014), and many successful drug repurposing cases (Li and Jones, 2012; Chong and Sullivan Jr., 2007; Ashburn and Thor, 2004; Wu et al.,

* Correspondence to: A. Sharma, Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115, USA.
** Correspondence to: Z. Wang, Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, No.16, Nanxiaojie, Dongzhimennei, Dongcheng District, Beijing 100700, China.
E-mail addresses: amitabh.sharma@channing.harvard.edu (A. Sharma), wangzh@mail.cintcm.ac.cn (Z. Wang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ebiom.2018.04.002

2016; Evans et al., 2005) altogether demonstrate the vague boundary between different diseases in current disease taxonomies. Moreover, the deep understanding of diseases based on the advances in disease biology, bioinformatics, and multi-omics data necessitates the reclassification of disease taxonomy (Mirnezami et al., 2012).

In the past decade, efforts to reclassify diseases based on molecular insights have increased with studies related to molecular-based disease subtyping in different disease conditions, such as acute leukemias (Golub et al., 1999; Alizadeh et al., 2000), colorectal cancer (Dienstmann et al., 2017), oesophageal carcinoma (Cancer Genome Atlas Research et al., 2017), pancreatic cancer (Bailey et al., 2016), cancer metastasis (Chuang et al., 2007), neurodegenerative disorders (Mann et al., 2000), autoimmunity disorders (Ahmad et al., 2003), multiple cancer types across tissues of origin (Hoadley et al., 2014), and a network-based stratification method for cancer subtyping (Hofree et al., 2013). Further insights will arise from integrating all types of biomedical data with a single framework to exploit disease-disease relationships. Data integration methods that utilize multiple types of data, including ontological and omics data, have been used to classify and refine disease relationships (Gligorijevic and Przulj, 2015; Menche et al., 2015; Gligorijevic et al., 2016). Despite these efforts, the development of a molecular-based disease taxonomy that links molecular networks and pathophenotypes still remains challenging (Menche et al., 2015; Hofmann-Apitius et al., 2015; Jameson and Longo, 2015).

Here, we aim to refine a widely used clinical disease classification scheme, the ICD. To achieve this, we first quantify the category similarity among the ICD chapters using ontology-based similarity measures and investigate the molecular connections of disease pairs in the same ICD chapters. Furthermore, we seek the correlation between category and molecular similarity, and check for the heterogeneity of molecular specificity and correlated boundary between categories in ICD taxonomy. Finally, we construct a new classification of diseases (NCD) with overlapping structures. The aim is to provide clear boundaries between distinct diseases belonging to different categories using a new disease classification scheme (Fig. 1 & Fig. S3).

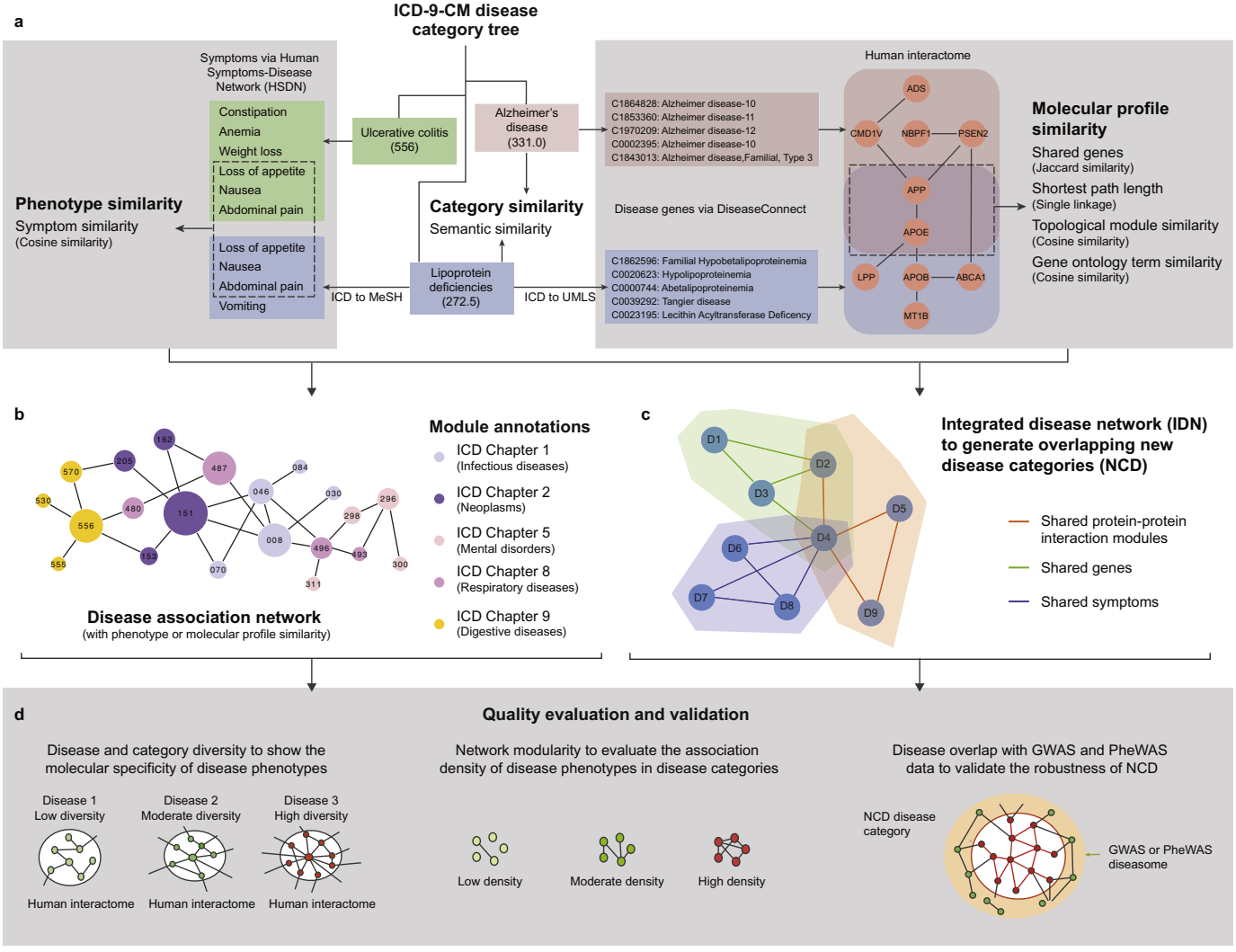

Fig. 1. Overview of the new disease taxonomy construction and validation. a. Similarity calculation between the disease pairs in ICD taxonomy, including the calculation of 1) category similarity; 2) Phenotype similarity (based on ICD-MeSH term mapping) and 3) Molecular profile similarities (based on ICD-UMLS term mapping) of disease pairs in ICD; b. Module or community annotations of disease association network by chapters in ICD or NCD. We generate disease association network, in which nodes represent diseases and the link weights represent their corresponding phenotype or molecule profile similarities. The module annotations of the disease network correspond to ICD chapters or NCD categories; c. Construction of integrated disease network (IDN) and generation of NCD. The links of IDN are fused from the multiple similarities (e.g. phenotype similarity and shared gene similarity). Based on IDN, NCD is generated by community detection algorithms with overlapping disease members; d. Quality evaluation and validation of ICD and NCD. The molecular specificity (or inverse molecular diversity) and network modularity are used for evaluation and comparison of the quality of two disease taxonomies. Furthermore, we validate the robustness of NCD with two independent phenotype-genotype association datasets, namely GWAS and PheWAS.

Please cite this article as: Zhou, X., et al., A Systems Approach to Refine Disease Taxonomy by Integrating Phenotypic and Molecular Networks, EBioMedicine (2018), https://doi.org/10.1016/j.ebiom.2018.04.002