Infection, Genetics and Evolution 46 (2016) 118–123

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Novel tuberculosis susceptibility candidate genes revealed by the reconstruction and analysis of associative networks

Elena Yu. Bragina [a,*], Evgeny S. Tiys [b,c], Alexey A. Rudko [a], Vladimir A. Ivanisenko [b], Maxim B. Freidin [a]

[a] Laboratory of Population Genetics, Research Institute of Medical Genetics, Tomsk NRMC, Nabereznaya Ushaiki Str. 10, Tomsk 634050, Russia
[b] Laboratory of Computer-Assisted Proteomics, The Federal Research Centre Institute of Cytology and Genetics of The Siberian Branch of the Russian Academy of Sciences, Lavrentiev Ave. 10, Novosibirsk 630090, Russia
[c] Laboratory of Computer Genomics, Novosibirsk State University, Pirogova Str. 2, Novosibirsk 630090, Russia

ARTICLE INFO

Article history:
Received 2 July 2016
Received in revised form 25 October 2016
Accepted 30 October 2016
Available online 31 October 2016

Keywords:
Tuberculosis
Susceptibility
Genes
Associative network

ABSTRACT

Tuberculosis (TB) is a common infectious disease caused by *M. tuberculosis*. The risk of the disease is dependent on complex interactions between host genetics and environmental factors. Accumulated genomic data, along with novel methodological approaches such as associative networks, facilitate studies into the inherited basis of TB. In the current study, we carried out the reconstruction and analysis of an associative network representing molecular interactions between proteins and genes associated with TB. The network predominantly comprises of well-studied key proteins and genes which are able to govern the immune response against *M. tuberculosis*. However, this approach also allowed us to reveal 12 proteins encoded by genes, the polymorphisms of which have never been studied in relation to *M. tuberculosis* infection. These proteins include surface antigens (*CD4*, *CD69*, *CD79*, *CD80*, *MUC16*) and other important components of the immune response, inflammation, pathogen recognition, cell migration and activation (*HCST*, *ADA*, *CP*, *SPP1*, *CXCR4*, *AGER*, *PACRG*). Thus, the associative network approach enables the discovery of new candidate genes for TB susceptibility.

1. Introduction

The study of the mechanisms of human susceptibility to infections caused by pathogenic microorganisms such as *M. tuberculosis* is an important and productive field of human genetics. It aims to reveal genetic variants predisposing to tuberculosis (TB) in order to facilitate the development of new effective measures to combat the disease. For this reason, both candidate genes and genome-wide association studies are usually carried out. The results of these studies confirm the importance of host genetic components in the risk of contracting TB; however, their reproducibility is poor and the convincing examples of causal genes are solitary.

Contemporary technologies and tools for the analysis of the human genome, allow for detailed estimations of the significance of genetic variability in the development of infectious diseases (Meyer and Thye, 2014). However, the understanding of gene function is impossible without systems biology, which, if applied to human pathology, offers a unique approach for studying multifactorial diseases. Systems biology allows revealing molecular links between diseases and identifies shared genes and pathways to achieve better understanding of the molecular origin of different disease groups. For instance, such the integrative analysis helped discover that leprosy susceptibility genes are also crucially important for the development of autoimmune diseases (Zhang et al., 2016). In systems biology methodology, a network is a central object that represents links between multiple objects (Koonin, 2011). In gene networks, the nodes represent genes and the connections are their interactions (Barabási and Oltvai, 2004). Thus, a gene network is a group of genes functioning in a coordinated fashion, providing control for vital functions in an organism (Kolchanov et al., 2000). Studying the patterns of functioning of such networks may lead to the identification of genes responsible for the development of diseases. For instance, a number of studies using gene network methodologies have revealed novel genes that are important for the development of breast cancer (Pujana et al., 2007), inherited ataxia (Lim et al., 2006), asthma (Hwang et al., 2008), diabetes mellitus (Kussmann et al., 2013; Liu et al., 2007), neurodegenerative diseases (Parikshak et al., 2015), and cardiovascular diseases (Shangguan et al., 2014). Moreover, it is apparent that a breakthrough in the analysis of pathological processes and the development of effective therapeutics for complex diseases (including infectious disorders) is impossible without using the knowledge about functioning of molecular networks.

The wealth of experimental data in molecular biology provides a strong basis for the development of methods for the reconstruction of gene networks based on the automatic analysis of scientific texts and databases (Rzhetsky et al., 2004). One of the most popular softwares

* Corresponding author.
E-mail address: elena.bragina72@gmail.com (E.Y. Bragina).

http://dx.doi.org/10.1016/j.meegid.2016.10.030
1567-1348/

for such analysis is STRING (Szklarczyk et al., 2015) which uses "text-mining" technology for the reconstruction of networks. Another software, ANDSystem, was developed for automatic extraction of knowledge about molecular genetic interactions between proteins, genes, mRNA, metabolites, biological processes and diseases, from the texts of scientific publications, and their presentation in the form of associative semantic networks. Unlike STRING, ANDSystem provides more detailed descriptions of interactions between objects (Ivanisenko et al., 2015). The ANDSystem has been used to reconstruct associative networks for myopia and glaucoma (Podkolodnaya et al., 2011) as well as to describe molecular genetic interactions for co-morbid (Glotov et al., 2015) and inverse co-morbid pathological conditions (Bragina et al., 2014).

In this study, we set out to construct an associative network for TB using the ANDSystem software to reveal key molecules driving the disease pathogenesis.

2. Material and methods

We used the ANDSystem software to reconstruct an associative network for pulmonary TB. Extraction of the data by ANDSystem is carried out by applying a text-mining algorithm to the ANDCell knowledge base, followed by the visualisation of the results by the ANDVisio software (Ivanisenko et al., 2015).

We used a "tuberculosis pulmonary" search query to build the network, followed by a re-assessment through the expert review to remove the network redundancy caused by the incomplete formalization of the texts of research articles. The following parameters were taken into account for the expert review: 1) participation of a protein in TB pathogenesis; 2) correct protein/gene name recognition; 3) correct context (protein/gene mentioned in the abstract is really associated with the disease); 4) nominal statistical significance ($p < 0.05$) for the association between the protein and TB. After the review, the associative network was shrunk to contain only the most essential objects and their interactions.

We used the public databases Ensembl (http://www.ensembl.org/index.html), NCBI (http://www.ncbi.nlm.nih.gov/), and HuGE Navigator (https://phgkb.cdc.gov/HuGENavigator) to find out whether there are known polymorphisms of the genes in the network associated with TB or immune-mediated diseases.

Gene ontology-based overrepresentation analysis was conducted using the BiNGO program with the following parameters: hypergeometric test; Benjamini-Hochberg FDR-correction; significance level = 0.05; ontology = go-basic.obo (data-version: releases/2016-03-23); annotation = UniProt-GOA (Submission Date: 3/16/2016).

3. Results and discussion

Using the ANDSystem software we compiled a list of 131 different proteins and genes associated with TB according to data published in scientific literature. After the expert review, the list was reduced to the 40 genes/proteins most essential in TB pathogenesis (Supplementary Table 1; Fig. 1).

In the associative network, genes and proteins are separate components (Fig. 1). The network comprises 40 genes and their 45 protein products. The links between genes and the proteins they encode are defined by the type "expression". This link type is assessed in case of the presence of a Gene Id link in SwissProt card for the protein. The total number of proteins exceeds the total number of genes in the network because for *HLA-A* and *HLA-DRB1* genes there were 7 proteins; MHC

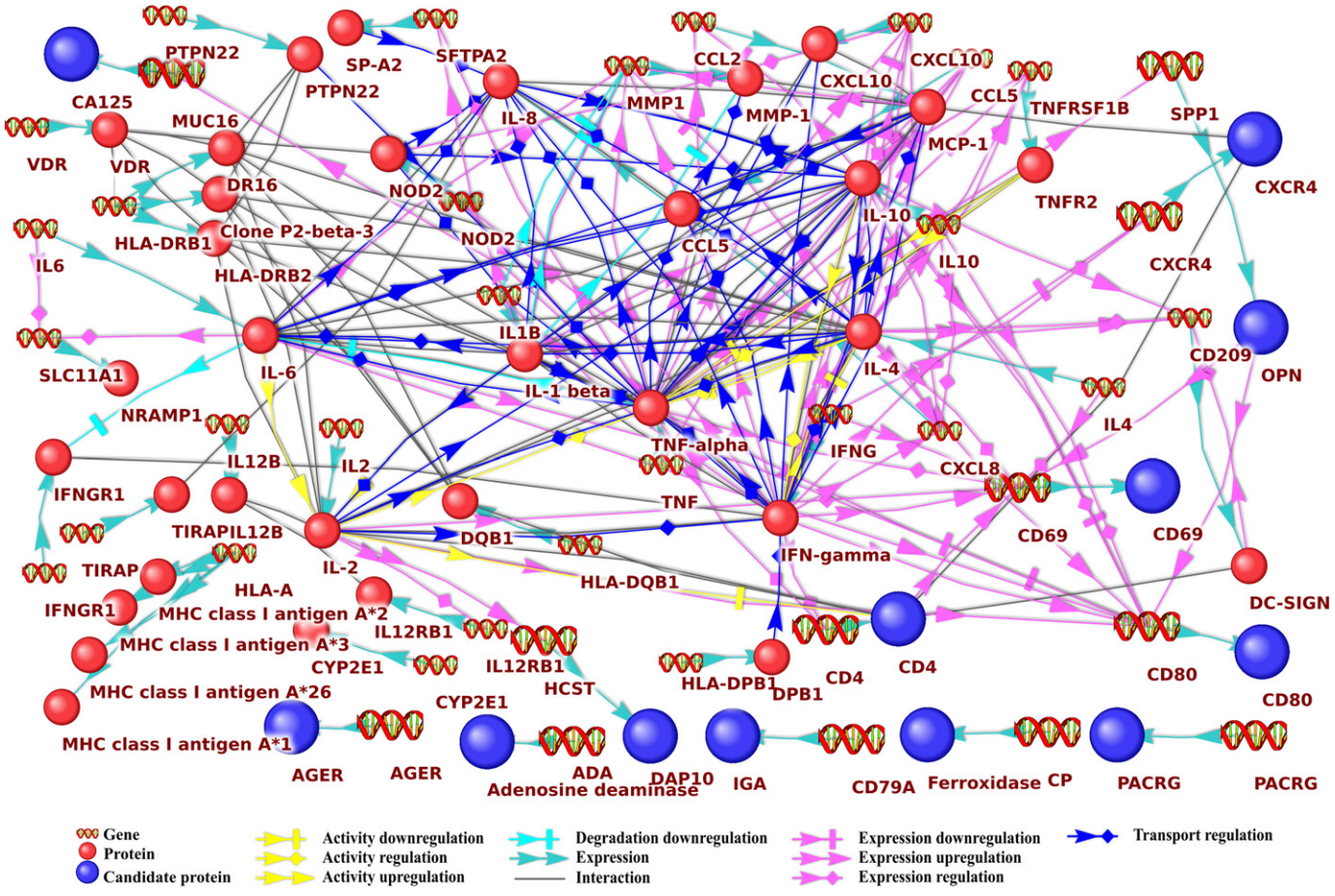

Fig. 1. Molecular genetic network of functional interactions between human genes/proteins associated with tuberculosis (associative network).