A Recognition Method of Complex Disease Risk Gene Based on Network Integration

Ming Qiu*
Heilongjiang University, Heilongjiang, China

*E-mail: 215940191@qq.com

Abstract. In this paper, rheumatoid arthritis is studied as an example of complex diseases. As we all know, rheumatoid arthritis is a relatively complex disease of autoimmunity. It is caused by genetic and environmental factors. However, the mechanism of its pathogenesis and regulation is still unclear. This paper proposes a new method framework from the system level. We plan to use the already regulated information and functional relationships to explain the risk gene elements. For different risk gene modules, we can use the database to explore the relationship between them and the corresponding transcription factors. I believe that the results obtained will provide useful reference for the development of clinical medicine.

Keywords: Network, Complex Diseases, Risk Genes

1. Introduction

Rheumatoid arthritis, as a complex immune disease, is determined by genetic genes and living environment. There are many kinds of pathogenic factors of this kind of disease. It has genetic heterogeneity. There is no redundant correspondence between the genotype and phenotype of the disease. Many medical researchers are troubled by this problem[1].

In recent years, due to the emergence of single nucleotide polymorphism, researchers can accurately find the location of specific diseases[2]. The emergence of this method has made outstanding contributions to the prevention, diagnosis, and treatment of complex diseases. At the same time, it also creates feasibility for the development of gene drugs.

2. On the study of genetic complex diseases

In general, most complex diseases are the result of gene environment interaction. In essence, the difference between genetic factors and environmental factors in the pathogenesis of different diseases is only the difference of degree. In the research methods of epidemics, they are also different. The basic
research methods of genetic epidemics are family clustering analysis and pedigree analysis. However, compared with common genetic diseases, the research method of complex diseases is the estimation of the heritability of susceptibility and the correlation analysis of case-control study[3].

According to the researchers’ large-scale investigation of genetic complex diseases, it is found that through the simulation of in vitro experiments can effectively verify the role of certain environmental factors in the development of diseases. The relationship between environmental factors and the development of the disease can be obtained by case-control study. These studies provide the necessary basis for the follow-up gene location, cloning and prevention and control of diseases from the environmental aspects.

3. Sources and methods of gene data for complex diseases

3.1. Source of data

WTCCC data is a genetic database for studying human diseases. WTCCC contains genotype data of 1860 rheumatoid patients and 10608 normal controls.

3.2. Module identification of important risk genes

3.2.1. Location of protein network genes and extraction of key genes. In order to understand the regulatory mechanism of rheumatoid arthritis, we need to combine genetic factors with biological network information. We can use the web database to map proteins as risk genes (see Table 1). We can define a specific protein as a hub protein. By extracting hub proteins from the gene network, we can determine the main relationship between these proteins and rheumatoid arthritis[4].

3.2.2. Identification of risk gene modules of important regulatory genes. Researchers need to think of genes linked to key regulatory genes as hub gene sets. In the process of risk analysis of hub gene set, the p value in genotype experiment is used to distinguish the important sites that contribute to phenotype. Set the p value of the i gene to be represented by p_i.

\[ t_i = -2 \log p_i \]  \hspace{1cm} (1)

\[ \sum_{i=1}^{k} t_i \sim X^2(2k) \]  \hspace{1cm} (2)

If \( P < 10^{-5} \), these gene modules are considered to be important risk genes related to rheumatoid arthritis.

3.3. On the regulatory mechanism between important risk genes and transcription factors

For each module of risk gene of hub gene, we can find its target gene and corresponding transcription factors by using TRANSFAC database. We can explore the interaction of multiple transcription factors in the regulatory mechanism. The researchers found that a transcription factor can regulate multiple risk gene modules.
Table 1. Measurement of hub gene set in the experiment

| Hub genes | Degree | Hub genes | Degree |
|-----------|--------|-----------|--------|
| PKCA      | 50     | PRKCE     | 24     |
| EGFR      | 48     | IRSI      | 24     |
| FNI       | 40     | PTPRC     | 24     |
| ADCY2     | 39     | ERBB4     | 23     |
| NGFB      | 38     | GP3A      | 22     |
| GSK3B     | 29     | GRIN2A    | 22     |
| NCOA3     | 25     | FGF1      | 23     |
| FRKCB1    | 25     | HD7       | 23     |
| CD40      | 24     | IL2RA     | 22     |
| PLCG2     | 24     | MCAF      | 24     |
| TRAF6     | 23     | VWF       | 18     |
| MHS3      | 22     | OTF3C     | 18     |
| ANK       | 23     | JNK3      | 17     |
| BIR1      | 22     | CD49B     | 17     |
| GRIN2B    | 22     | NDF       | 17     |
| DISC1     | 22     | ZO-1      | 17     |
| PDGFR     | 22     | ABLI      | 15     |
| IL6S6     | 20     | TNFAIP3   | 17     |
| CD80      | 20     | CDHH      | 17     |
| IL2RB     | 19     |           | 18     |
| MCAF      | 23     |           |        |

4. Results of numerical tests

4.1. Extraction of related genomes

The p value of the gene sequence represents the reliability of a certain genome related to rheumatoid
arthritis. In order to highlight the weight difference between unimportant genes and important genes\textsuperscript{[6]}, we take the logarithm of \( p \). In this way, the negative logarithm of \( p \) value of each gene sequence can reflect the degree of correlation with rheumatoid arthritis.

4.2. Mapping from related genomes to risk genes

We can locate risk genes through database. We need to calculate the distance of all gene sequences along the chromosome to their nearest gene splice. We found that the range is 0-5000 base pairs. The highest frequency occurs within 0-4000 base pairs.

4.3. Identification of important risk genes related to hub genes

Based on the risk analysis of key regulatory gene related gene modules, the researchers found that 37 gene modules were associated with the risk genes of rheumatoid arthritis. This finding confirms that the emergence of risk genes may lead to various complex genetic diseases.

5. Challenges and prospects of identifying sensitive genes for complex diseases

Most complex diseases are caused by genetic factors. The incidence rate of these diseases is generally over 1/1000. They have certain family tendency in clinical or epidemiological history. But this is not a typical Mendelian inheritance. The cause of complex diseases is still a difficult problem in the medical field. Researchers believe that the mode of micro effect plays an important role in the pathogenesis of complex diseases. Genes together determine the genetic susceptibility of complex diseases.

However, the identification of risk genes for complex diseases is still challenging. At present, there are still many problems to be further investigated. In recent years, the strategy of quantitative trait location has gradually matured. The research strategy of common diseases has been widely developed and applied.

6. Conclusion

The results show that most of the key regulatory genes can be identified in the important gene sequences by the analysis method. These regulatory genes are risk genes associated with rheumatoid arthritis. However, it is still a challenging task to use the existing functional relationships to explain the genetic factors of complex diseases.

References

[1] Jianghe Yao, Tiecheng Bai, Gang Wu. A Recognition Method of Red Jujube Disease Based on Portable Microscope and Pso-bp Neural Network[J]. Information Technology Journal, 2013, 12(22):6681-6685.

[2] Fu, Hao Yue, Yang, Lianping, Zhang, Xiangde. A Prioritization Method for Identifying Disease-Causative Gene Based on Hyper Graph Network[J]. Current Proteomics, 2016, 13(2):--.

[3] Chen Lin, Mukerjee Gouri, Dorfman Ruslan, etc. Disease Risk Assessment Using a Voronoi-Based Network Analysis of Genes and Variants Scores[J]. Frontiers in Genetics, 8.
[4] Li, Wan, Zhu, Lina, Huang, Hao. Identification of susceptible genes for complex chronic diseases based on disease risk functional SNPs and interaction networks[J]. Journal of Biomedical Informatics, 74:137-144.

[5] Li, Gao. Intellectual recognition of reliability analysis model based on neural network integration[C] 2012.

[6] WANG Hong, QU Xiao-Li, ZHAO Yan. Uncovering Atherosclerotic Risk Disease Gene Based on Expression and Network Topological Structure[J]. Progress in Biochemistry & Biophysics, 2010.