Genomics 109 (2017) 158–164

Contents lists available at ScienceDirect

Genomics

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

Candidate novel long noncoding RNAs, MicroRNAs and putative drugs for Parkinson's disease using a robust and efficient genome-wide association study

Zahra Mortezaei, Hossein Lanjanian, Ali Masoudi-Nejad *

Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

ARTICLE INFO

Article history:
Received 10 December 2016
Received in revised form 2 February 2017
Accepted 15 February 2017
Available online 21 February 2017

Keywords:
Single-nucleotide polymorphism
Parkinson's disease
Linkage disequilibrium
Genome-wide association studies
Long noncoding RNA
Transcription factor binding site

ABSTRACT

Parkinson's disease (PD) is a neurodegenerative disorder with serious symptoms of which, are not clearly demonstrated at the beginning stages of the disease, making treatment challenging. Understanding the genetic causes of PD can be useful for determining its mechanisms and proposing treatments and preventive methods. For different populations with different genetic backgrounds and lifestyles, genome-wide association studies (GWASs) represent a crucial approach for genetic analysis. In this study, a robust and efficient GWAS without dimensionality reduction applied to evaluate heritability and genetic causes of PD in the German and US populations. The results show higher rate of PD heritability in the German population. Moreover, 25 significant SNPs have been determined, as well as five newly identified candidate genes associated with PD and some potential drug candidates. Analysis also reveals various long noncoding RNAs (lncRNAs), microRNAs and transcription-factor binding sites (TFBSs) with potential in the prevention and treatment of PD.

1. Introduction

Neurodegenerative diseases, such as Alzheimer's and Parkinson's, are disorders in which neurons have been selectively lost. They are characterized by the continuous degeneration of function and structure of the peripheral or central nervous system. After Alzheimer's, Parkinson's disease (PD) is the second most common neurodegenerative disease in aged populations. PD worldwide afflicts approximately 6 million people. Clinically, PD is manifested by postural instability, bradykinesia, rigidity and resting tremors. The most important risk factor for developing PD is aging. PD was initially described by James Parkinson in 1817, but its cause and successful treatment strategies remain unknown [1].

PD cannot be proven by testing, and its aetiology remains unknown. Difficulties in movement are typical symptoms associated with PD. When PD is in its early stages, it is difficult to distinguish, but it is easier to discern as the symptoms gradually worsen. There are no precise treatments and cures for PD or the prevention of its progression; therefore, for general physicians and neurologists, it is challenging to manage PD [2]. In sporadic PD cases, the mechanism of pathogenesis can be more clearly defined by changes in specific genes, long noncoding RNAs (lncRNAs), transcription factor binding sites (TFBSs) and microRNAs.

The human genome has been successfully sequenced, and the specific genes that play roles in PD are being elucidated; approximately 13 genetic mutations have been discovered that play important roles in PD [3]. Some of these genes are important for PD heritability and code for special proteins that cause the disorder, and some treatment strategies can be considered for this disease base on these discoveries. One method includes drugs that interact with specific genes; another method uses specific genes to treat or prevent PD experimentally, i.e. gene therapy [4]. In addition, previous studies have shown that TFBSs have potential to affect the expression of genes controlling PD [5].

In addition to genes, non-coding RNAs (ncRNAs) have important impacts on the progression and pathogenesis of PD. For example, the impact of microRNAs and lncRNAs on PD has been previously documented. In many diseases, the importance of misexpression and mutations in ncRNAs has been implicated, and ncRNAs have been used therapeutically [6]. Because so few genes, lncRNAs, TFBSs and microRNAs have been found to be associated with PD, genetic variations in the susceptibility for PD are unknown; thus, there is no cure for PD. The detection of the genetic causes of PD hold the greatest promise for treatment and prevention.

In this study, a genome-wide association study (GWAS) was applied to examine this hypothesis. In medical genetics, GWAS contributes to a better understanding of disease pathogenies and aetiologies. Association analyses of this type are usually performed with a linkage

* Corresponding author.
E-mail address: amasoudin@ut.ac.ir (A. Masoudi-Nejad).
URL: http://www.LBB.ut.ac.ir (A. Masoudi-Nejad).

http://dx.doi.org/10.1016/j.ygeno.2017.02.004
0888-7543/

disequilibrium (LD) analysis [7]. An LD analysis, which is an important and effective analysis for detecting single-nucleotide polymorphisms (SNPs) that associate with quantitative traits, requires randomly collected samples, although this fact has usually been ignored. This oversight occurs because samples are not usually randomly selected. Most studies that have estimated the LD for random samples have ignored the non-randomness of the samples. One such example is the 'chromosome counting' method applied by Hill to estimate the LD parameter for random samples when the genotypes of the two loci were given [8]. Methods originally developed for random samples display severe biases when non-random samples are used. To decrease the level of biases in LD analyses when factors exist, such as non-random samples or population structures, some studies have applied likelihood-based statistical methods that are flexible for genetic association analysis [9,10]. It has been shown that likelihood-based methods perform well, with high statistical power and fewer false-positive results [11,12].

In this study, case-control samples were used based on selection; they were not random. We developed a new method to perform GWAS with reduced biases. The study of PD heritability and biological concepts related to this disease are very important, and dimensionality reduction has not been performed in this analysis to consider all biologically important SNP data. However, the results of the GWAS for all the SNPs were further analysed using a comprehensive system biological analysis to distinguish the genes, lncRNAs, microRNAs and TFBSs associated with PD. Afterward, the genes that might contributed to PD were subjected to further biological analysis to identify specific potential drugs that could be useful for treating PD.

2. Materials and methods

The likelihood-based machine learning approach applied in this GWAS study was an expectation maximization (EM) algorithm. For this algorithm, the conditional probability distributions for the markers and disease genotypes are latent variables that can be written with respect to population genetic parameters. The population genetic parameters in this study were genotypic distributions at markers, disease loci and coefficients of LD. The aim of the EM algorithm was to estimate the disease allele frequency and coefficients of LD. Therefore, the likelihood function of those parameters constructed, and the results of estimating the parameters in the EM algorithm converged to the maximum likelihood estimation. In addition, the conditional probability distributions provided functions of the penetrance parameters that characterized PD heritability [10]. Details of that EM algorithm are available in Supplementary Note.

The input data applied in the GWAS in this study were similar to those used in a previous study [10]. That dataset was initially published by Simon-Sanchez et al. [11]. This case-control dataset contained 4005 individuals (971 cases and 3034 controls) from the United States and 1686 individuals (742 cases and 944 controls) from Germany. Those individuals were genotyped using Infinium BeadChip (Illumina, Inc.) for 507,861 SNPs. Details of the genotyping and the preparation of the input dataset for the GWAS are explained in Fig. S1 [13].

After applying the EM algorithm to these data without dimensionality reduction, the results of that algorithm were further analysed. First, a special threshold value of 2.4E-07 was selected. The significant SNPs

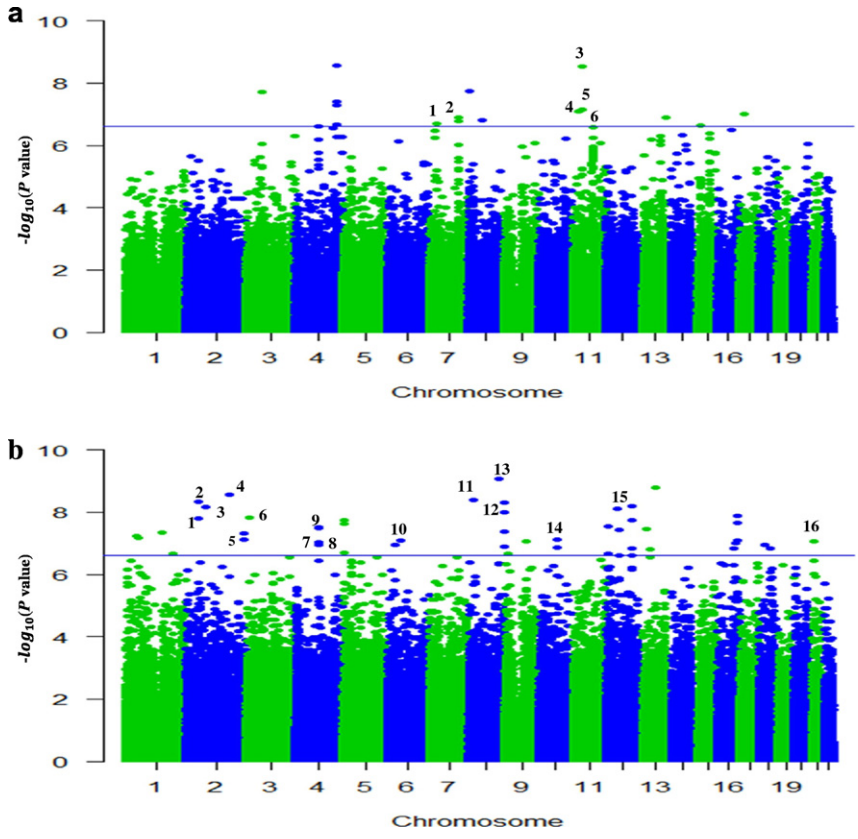

Fig. 1. Manhattan plots showing the results of the genetic association analysis for additive heritability factors. The graphs represent the results estimating the linkage disequilibrium for the additive effect of disease genes in the a) US and b) German populations. The threshold value of 2.4E-07 is represented by a blue horizontal line in each graph. The numbers refer to important significant SNPs. In US population, number 1 is rs3779331, 2 is rs859522, 3 is rs10500796, 4 is rs1726764, 5 is rs1941817, 6 is rs11605276. In German population, number 1 is rs1474406, 2 is rs3100218, 3 is rs1447108, 4 is rs7564397, 5 is 17190254, 6 is rs1605527, 7 is rs2736990, 8 is rs11931074, 9 is rs1398908, 10 is rs2072638, 11 is rs2736050, 12 is rs2009817, 13 is rs4875773, 14 is rs2492448, 15 is rs10849446, 16 is rs2070535.