Genomics 105 (2015) 168–174

Contents lists available at ScienceDirect

Genomics

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

Distributions of allele combination in single and cross loci among patients with several kinds of chronic diseases and the normal population

Li-ping Gai [a], Hui Liu [b,*], Jing-hui Cui [b], Na Ji [b], Xiao-dong Ding [a], Cui Sun [b], Lai-shui Yu [b]

[a] Department of Physics, Dalian Medical University, Dalian 116044, China
[b] College of Medical Laboratory, Dalian Medical University, Dalian 116044, China

ARTICLE INFO

Article history:
Received 18 June 2014
Accepted 27 December 2014
Available online 2 January 2015

Keywords:
Genetic polymorphism
Genome
Chronic diseases

ABSTRACT

Genetic research has progressed along with scientific and technological developments. However, it is difficult to identify frequency differences in a particular allele distribution at a single locus. Such differences can be identified by examining the allele combination distribution. We explored different mathematical methods for statistical analyses to assess the association between the genotype and phenotype. We investigated the frequency distributions of alleles, combinations of single-locus genes, and combinations of cross-loci genes at 15 loci using 447 blood samples of 200 normal subjects, 72 patients with chronic obstructive pulmonary resistance, 50 liver cancers, 75 stomach cancers and 50 hematencephalon and identified each population as having a unique gene distribution and that the distribution followed certain rules. The probability of illness followed different rules and had apparent specificity. Differences obtained using statistics of combinations of cross-loci genes are superior to single-locus gene statistics, and combinations of single-locus gene statistics are better than allelic statistics.

1. Introduction

With recent developments in clinical laboratory technology increasingly, more researchers are focused on the alleles of different genes between the normal population and patients with various diseases. Each gene has potentially different alleles. Different alleles for a gene may be present on each chromosome. Although the combination of a particular allele may be associated with a specific disease, a disease caused by a combination of alleles is not limited to the combination of alleles within a particular locus [1]. New discoveries are possible if we study the relationship between a disease and the combinations of a particular allele at different loci [2]. Generally, the allele combinations are randomly distributed for the normal population. If the allele combinations for certain disease groups aren't randomly distributed, then we can find the special allele combination associated with a particular disease using mathematical and statistical methods. Because large numbers of genes are distributed in many combinations on different chromosomes, multigenic disease combinations are very difficult to identify by human observation [3,4]. At present, cross-loci combinations of genes are very complex and difficult to assess; thus, very few researchers are studying this topic [5]. We therefore designed a computer analysis system to fill this research gap. This system can analyze frequency distributions of genes, including combinations of single-locus genes and cross-loci, from different pathogen libraries [6,7]. It is a powerful tool for pathogenic gene distribution statistics. We chose to evaluate chronic obstructive pulmonary disease (COPD), liver cancer, stomach cancer and hematencephalon to illustrate the differences between the gene distributions of diseases and those of the normal population. Patients with COPD, liver cancer, stomach cancer and hematencephalon were compared with normal subjects with respect to their gene distributions using the same allele combination technology. This analysis system can also be extended to any pathogenic allele combination.

2. Materials and methods

This study evaluated the allelic frequency distributions and allele combination frequency distributions among 200 normal subjects, 72 patients with COPD, 50 patients with liver cancer, 75 patients with stomach cancer and 50 patients with hematencephalon (in total, 447 samples at 15 loci).

To all the individuals, the genes are sampled in the patient's blood sample; detailed information was collected and provided by clinical laboratory physicians. For patients with COPD and hematencephalon, the diseases were evaluated by two to four physicians. For patients with liver cancers and stomach cancers, the patients' disease diagnoses were confirmed by operations and pathological sections.

* Corresponding author.
E-mail address: liuhui60@sina.com (H. Liu).

http://dx.doi.org/10.1016/j.ygeno.2014.12.008
0888-7543/

For example, a total of 72 Chinese patients with COPD were recruited from the Dalian 210 Hospital, Affiliated Hospital of Chifeng College, and the Affiliated Hospital of Dalian Medical University in China. We defined COPD according to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria [8]. DNA was extracted from 3 ml of peripheral blood using the Chelex 100 procedure [9,10]. The capillary electrophoresis was used on an ABI 310 automated sequencer (Applied Biosystems). The gene scan Analysis 2.1 software (Applied Biosystems) was used to determine fragment sizes. Alleles were identified by comparing the amplified fragments with the allelic ladders included in the reagent set. The alleles were labeled according to international nomenclature using the Genotyper Software package (Perkin Elmer).

Because of the quantity of samples, loci can be increased without limit. For input and statistical convenience, we used the number to represent gene number. We get the biological mechanisms by searching the GenBank database if we know the loci, chromosomal location and allele number. See "Appendix Table 1 — Parameters of loci amplified" in the additional file for reference.

A sample of the library (screenshot of the data of 50 patients with liver cancer) is shown in Fig. 1. The data input rules for the 72 patients with COPD, 75 patients with stomach cancer and 50 patients with hematencephalon and 200 normal subjects were identical to those shown in Fig. 1. In the additional file the "source sample" sheet of "Excel-1 (50 patients with liver cancer)" provides further details. Each library comprises 15 loci: D8S1179, D21S11, D7S820, CSF1PO, D3S1358, D5S818, D13S317, D16S539, D2S1338, D19S433, vWA, D12S391, D18S51, D6S1043, and FGA. They are simplified as D8, D21, D7, CS, D3, D5, D13, D16, D2, D19, vWA, D12, D18, D6, and FGA. The Arabic number of each locus indicates the allele number. Each locus has two different chromosomes, one each from the father and mother. These two chromosomes are designated "a" and "b." The first column contains the patient's identification number, the second column contains the doctor's coding (patient's real name omitted), the third column contains the diagnosis, and all remaining columns contain the loci. "NULL" is the terminator of the samples.

2.1. Probability statistics of alleles at specific loci

Fig. 2 is a screenshot of the data of the 72 patients with COPD and illustrates the probabilities of allele distributions. The actual data are shown in the "gene number statistics sheet" of "2 Excel-2 (72 patients with COPD)" in the additional file. The same statistical algorithms were used for the 200 normal subjects, the 50 patients with liver cancer, 75 patients with stomach cancer and 50 patients with hematencephalon. We can get the allele numbers arranged in ascending order and their distributed quantities. The decimals represent the distribution of allelic probabilities at specific loci. For example, the probability distribution of allele 7 at locus D7 is 0.007. Blank cells in columns C to Q indicate that the probability distribution of the allele is zero at the corresponding locus.

The probability distribution of alleles at different loci was calculated using the Microsoft macro programming technology [11–13], and statistical analysis was performed using specific mathematical methods [14–16] involving permutations, combinations, etc. In this probability calculation, P represents the probability distribution of allele *i* at certain loci. $P = m/2/n$, where m represents the number of allele *i* at certain loci and n represents the quantity of the sample. Each locus has two different chromosomes, one each from the father and mother. Therefore, the maximum quantity of alleles at the same locus is 2n. The probability is considered to be statistically meaningful if the difference is $\geq 5\%$.

2.2. Differences in probability of alleles

The statistical analysis revealed that the numbers of alleles at different loci had specific effects on the patients. Some differences in the probability distribution were close to 10% and met certain rules. The allele probability distributions were divided into three types among the patients with liver cancer and COPD and the normal population: the first affects only liver cancer, the second affects COPD, and the third affects both liver cancer and COPD. The data are shown in

name	sample from	Diagnosis	D8a	D8b	D21a	D21b	D7a	D7b	CSa	CSb	D3a	D3b	D5a	D5b	
1	x1	liver cancer	11.00	14.00	30.00	32.20	8.00	10.00	11.00	12.00	16.00	16.00	10.00	10.00	NULL
2	x2	liver cancer	11.00	14.00	29.00	30.00	8.00	11.00	12.00	12.00	15.00	16.00	11.00	12.00	
3	x3	liver cancer	11.00	13.00	30.00	31.20	11.00	12.00	11.00	12.00	18.00	18.00	10.00	11.00	
4	x4	liver cancer	11.00	12.00	30.20	31.00	11.00	13.00	11.00	13.00	16.00	16.00	10.00	11.00	
5	x5	liver cancer	10.00	15.00	30.00	30.00	9.00	11.00	10.00	12.00	15.00	16.00	9.00	9.00	
6	x6	liver cancer	16.00	16.00	29.00	30.20	10.00	11.00	10.00	10.00	16.00	16.00	8.00	9.00	
7	x7	liver cancer	13.00	15.00	30.00	30.00	11.00	12.00	12.00	12.00	16.00	17.00	11.00	12.00	
8	x8	liver cancer	14.00	15.00	30.00	30.20	10.00	11.00	11.00	12.00	16.00	16.00	10.00	11.00	
9	x9	liver cancer	13.00	13.00	30.00	31.00	8.00	8.00	12.00	13.00	15.00	16.00	10.00	13.00	
10	x10	liver cancer	13.00	14.00	32.20	32.20	11.00	11.00	10.00	11.00	15.00	15.00	11.00	13.00	
11	x11	liver cancer	10.00	15.00	30.00	31.20	9.00	11.00	10.00	12.00	15.00	16.00	9.00	13.00	
12	x12	liver cancer	11.00	14.00	29.00	32.20	11.00	11.00	10.00	12.00	16.00	16.00	11.00	12.00	
13	x13	liver cancer	11.00	11.00	30.00	30.00	11.00	12.00	9.00	13.00	15.00	17.00	12.00	12.00	
14	x14	liver cancer	14.00	15.00	28.00	30.00	12.00	12.00	10.00	11.00	16.00	17.00	11.00	11.00	
15	x15	liver cancer	12.00	13.00	31.20	32.20	8.00	11.00	11.00	12.00	15.00	15.00	10.00	12.00	
16	x16	liver cancer	10.00	13.00	29.00	29.00	11.00	11.00	11.00	12.00	15.00	17.00	10.00	11.00	
17	x17	liver cancer	14.00	15.00	29.00	30.00	11.00	12.00	10.00	12.00	15.00	17.00	8.00	11.00	
18	x18	liver cancer	12.00	14.00	29.00	31.00	11.00	12.00	9.00	12.00	14.00	15.00	11.00	12.00	
19	x19	liver cancer	14.00	14.00	29.00	32.00	11.00	11.00	10.00	11.00	15.00	16.00	7.00	11.00	
20	x20	liver cancer	10.00	15.00	30.00	33.20	11.00	12.00	12.00	18.00	15.00	18.00	12.00	13.00	
21	x21	liver cancer	13.00	16.00	30.00	33.00	10.00	10.00	10.00	12.00	15.00	17.00	9.00	13.00	
22	x22	liver cancer	12.00	14.00	29.00	29.00	9.00	10.00	10.00	14.00	16.00	18.00	9.00	12.00	
23	x23	liver cancer	11.00	13.00	28.00	29.00	11.00	13.00	11.00	12.00	15.00	17.00	13.00	13.00	
24	x24	liver cancer	12.00	14.00	29.00	30.00	11.00	11.00	11.00	12.00	15.00	17.00	11.00	12.00	
25	x25	liver cancer	10.00	13.00	29.00	29.00	8.00	11.00	12.00	12.00	15.00	17.00	10.00	11.00	
26	x26	liver cancer	11.00	12.00	29.00	30.00	10.00	13.00	10.00	12.00	15.00	17.00	10.00	12.00	
27	x27	liver cancer	8.00	14.00	30.00	31.00	8.00	12.00	10.00	10.00	15.00	16.00	9.00	10.00	
28	x28	liver cancer	14.00	15.00	28.00	30.30	10.00	11.00	10.00	12.00	15.00	16.00	11.00	11.00	
29	x29	liver cancer	10.00	14.00	30.00	32.00	10.00	11.00	10.00	11.00	16.00	17.00	12.00	12.00	
		NULL													

Fig. 1. A sample of the library (screenshot of the data of 50 patients with liver cancer) is shown. Each library comprises 15 loci: D8, D21, D7, CS, D3, D5, D13, D16, D2, D19, vWA, D12, D18, D6, and FGA. The Arabic number of each locus indicates the gene number. Each locus has two different chromosomes, one each from the father and mother. These two chromosomes are designated "a" and "b." The first column contains the patient's identification number, the second column contains the doctor's coding (patient's real name omitted), the third column contains the diagnosis, and all remaining columns contain the loci. "NULL" is the terminator of the samples.