Analysis and Study the Correlation of Genetic Loci

Liu Yang[1], Bai Xiaojun[2] and Kang Xiaofeng[3]

[1.2] School of Computer Science and Engineering, Xi'an Technological University,

Xi'an, China, 710021;

[3] School of Science Xi'an Institute of Technology, Xi'an, Shaanxi, China 710021

Email: 365361068@qq.com

Abstract. Based on information of 1000 samples provided by single nucleotide polymorphisms, analyzing the existing SNP data, using the optimized linear mixed model for data screening of correlation analysis, the coding mode of each given locus bases (A, T, C, G) are encoded into numerical encoding, with correlation analysis of collection of evolutionary algorithms, the corresponding figure in Manhattan and risk point and the position of the gene are obtained for further operations of the transferred numerical language, and finally got the image of the corresponding computing program operation and results. By carrying out statistical analysis and inspection for results, the score genotype data are more reasonable and accurate to explain theoretically the rationality of the discovered disease locus and gene.

Keywords: SNP loci, linear mixture model, Genome-wide association analysis

1. Introduction

Single nucleotide polymorphism refers to polymorphism caused by mutation in a single nucleotide (A, T, C, G) in the genomic DNA sequence, which is the most widely distributed in the human genome and rich in genetic information polymorphism [1]-[2].Genome-wide association studies (GWAS) have been widely used in medical genetics research at home and abroad. However, the rich information about the mechanism of multi-gene complex traits contained in GWAS data has not been fully exploited. The SNP loci information was determined by analyzing the SNP loci, so we did not need to construct any hypothesis before the study [3]. A large number existence of SNP sites determine the personality differences between people and greatly relate to complex diseases. Therefore, the SNP study has a great significance and SNP as a hot issue in bioinformatics research at present has high research value both in theoretical analysis

and practical application [3].Numerous studies have shown that many phenotypic traits differences in the body' s as well as susceptibility to drugs and diseases are possibly associated with certain loci, or genes containing multiple loci. Therefore, positioning the location of loci associated with trait or disease in gene and chromosome can not only help researchers understand the genetic mechanisms of diseases and traits, but also enable people to intervene against pathogenic sites to prevent some genetic diseases occur.

2. SNP Encoding

For each sample, the base (A, T, C, G) coding was used to obtain the information of each site, which contained only the site information and the base pair corresponding to each site, A, T, G, C, in practice, in accordance with the appropriate method it is converted to numerical encoding. In the position of rs100015, the coding of different samples is a combination of T and C, there are three different coding methods TT, TC and CC. The four base pairs were coded by SNP site analysis. Below each site, 1000 lines arranged combination of A, T, G, C, and it represents genetic information in different samples on this site. Although the composition of DNA has four bases, but generally the composition of SNP only has two bases, so it is a kind of two-state markings, namely diallele. That is in the same site, there may be up to three different sample encoding. The advantage of such coding is that the measured experimental results can be quickly converted into binary machine language, which is convenient for operation and avoids a lot of complex processes. Using binary language is also conducive to decimal language conversion. Adopting SNP analysis, rapid and large-scale screening can be carried out. It is very helpful for the rapid detection of mutations in the gene.

3. Linear Mixed Model

Linear mixed model is a common method of handling large data, based on the coding information of 9445 sites on a strip of possible causative chromosome fragment in the Appendix for 1000 samples and information on a sample with genetic diseases A, uses linear mixed Model and makes a more reasonable analysis of the measured data. If the model has both the fixed effects model and random effects, it is called linear mixed model (LMM). The linear mixed model can be expressed as follows [5]-[9]:

$$EY = X\beta, Cov(Y) = \sigma_i^2 I_i$$

$$V = \sum_{i=1}^{k} U_i U_i' \sigma_i^2 \tag{1}$$

This is a common form of linear mixed model, $\sigma_1^2, \cdots, \sigma_k^2$ is unknown, it is called variance components, also the linear mixed model is known as variance component model. If the variance component is estimated, then

$$Y = X\beta + U_1\xi_1 + \cdots U_k\xi_k \tag{2}$$

$$EY = X\beta, Cov(Y) = V = \sum_{i=1}^{k} U_i U_i' \sigma_i^2 \tag{3}$$

Therefore $Y \sim N_n\left(X\beta, V\left(\sigma_i^2\right)\right)$

The problem takes use of AIC (Akaie 'Information Standards) to select the model and goodness of fit test [10].

AIC=-2ln (Simulated likelihood) +2(Number of model degrees of freedom) (4)

For model (2), the AIC criterion can be derived as:

AIC=2ln (RSS $_q$) +2q (5)

4. Set Evolutionary Algorithm

(Genome Wide Association Study, GWAS) is a kind of effective method to find susceptibility genes for common diseases in which the related problems contain a lot of data and the existing methods are often difficult to solve the problem. This paper based on evolutionary algorithm set presents an effective solution to this problem. The method takes the assumed value and the probability as a new target to optimize the original problem and designs the fitness function to reflect user' s preferences. In addition, the proposed set evolution strategies by the paper show the superiority of the used methods in way of calculation results. It is characterized by the obtained best individual after the evolution of the population that is the approximate optimal solution set for the optimization problem. However, the evolution of ensemble also leads to a series of new problems, such as different sets, evolutionary individual comparison, and evolutionary strategy design.

According to the mean square error criterion:

$$E\left(\left\|\tilde{\beta}_q^* - \beta\right\|^2\right) = \left\|A\beta_1\right\|^2 + tr\left(X_q' V^{-1} X_q\right)^{-1} + \left\|\beta_1\right\|^2 \tag{6}$$

Constructed Modified Mean Square Error Criterion:

$$RMSE_q = \frac{E\left(\left\|\tilde{\beta}_q^* - \beta\right\|^2\right)}{n-q} \tag{7}$$

To carry out the variable selection, the smaller $RMSE_q$ the better, the model should choose the q independent variables when $RMSE_q$ is at the minimum. The benefit of using RMSE as the selection criterion is that, relative to the AIC criterion, using RMSE as a selection criterion unnecessarily requires the distribution type of model must be known. In practical problems, there are many information traits involved in constructing selection index, some of which are objective traits and some are non-target traits, such as the statistical distribution of the index, the desired progress of index and so on. The final calculation results show that the way of selection index can be effective to distinguish and choose different characteristics and traits of the group[10]-[11].

5. Model Selections

Follow the AIC guidelines above:

AIC=-2ln (Simulated likelihood) +2(Number of model degrees of freedom)

It can be applied to construct a variance component model AIC guideline. Likelihood function:

$$\mathrm{h}(L) = cons\tan t - \frac{1}{2}\mathrm{h}\left(\left|V\left(\sigma_i^2\right)\right|\right) - \frac{1}{2}(y - X\beta)' V\left(\sigma_i^2\right)^{-1}(y - X\beta) \tag{8}$$

If V is unknown, the estimation of $V\left(\sigma_i^2\right)$ is $\hat{V}\left(\hat{\theta}_i, \hat{\sigma}^2\right)$ can be solved by the likelihood equation, And an estimation of β is $\hat{\beta}$. Bring two estimations into there

$$\mathrm{h}(L) = cons\tan t - \frac{1}{2}\mathrm{h}\left(\left|\hat{V}\left(\hat{\theta}_i, \hat{\sigma}_i^2\right)\right|\right) - \frac{1}{2}\left(y - X\hat{\beta}\right)' \hat{V}\left(\hat{\theta}_i, \hat{\sigma}_i^2\right)^{-1}\left(y - X\hat{\beta}\right) \quad (9)$$

And the estimation is brought into the residual squared sum and taken into the AIC criterion

$$AIC = RSS\left(\hat{\theta}_i, \hat{\sigma}^2\right)_q + \mathrm{h}\left(\left|V\left(\hat{\theta}_i, \hat{\sigma}^2\right)_q\right|\right) + 2(q) \quad (10)$$

6. Analysis of Sites and Genes

6.1. Risk Point Analysis

As shown in Table 1, Statistical Table of Risk SNP Loci and in figure 1 Scatter Plots of SNP Loci: The encoded information of 9445 sites on a strip of possible causative chromosome fragment in 1000 samples was substituted therein, there came the statistical test results after calculating, as shown in Table 1, the plot of Manhattan as shown in Figure 1: The horizontal axis in the figure is 9445 site information, and the P value (-log10 (p)) in the vertical axis (i.e., Table 1). Selecting the value of P10-4 as cut-off value, when the P value is less than the cut-off value, it is the possible pathogenic sites of the disease. Suspected pathogenic sites were rs2273298, rs932372, rs12036216, corresponding to the three points above the blue line in Figure 1.

Table. 1 Statistical Table of Risk SNP Loci

SNP	A1	F_A	F_U	A2	CHISQ	P	OR
RS2273298	G	0.338	0.229	A	29.25	6.38E-08*	1.719
RS932372	G	0.147	0.09	A	15.55	8.03E-05*	1.742
RS12036216	A	0.093	0.15	G	15.22	9.57E-05*	0.581
RS2807345	T	0.259	0.189	C	14.09	1.74E-04	1.5
RS4391636	C	0.273	0.349	T	13.48	2.41E-04	0.7005

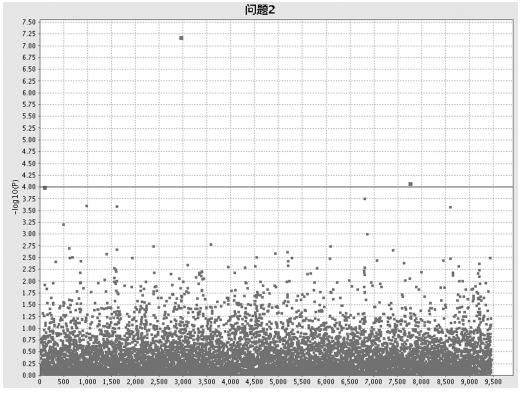

Fig.1 Scatter Plots of SNP Loci

6.2 Risk Gene Analysis

Table 2 is Risk Gene Statistics Table and Figure 2 shows Gene Scatter Plot.The gene locus information of the 300 data files in the gene info folder is imported into the mixed linear model established above, and

the result of the Manhattan map is obtained, as shown in Fig.2. Wherein, the horizontal axis is 300 genes, the vertical axis is EMP1 in the table, is the negative logarithm of probability. According to the calculation result, selecting EMP1 = 10-2 as the cut-off value, that is, the position shown in blue line. Table 2 shows the EMP1 values and site information of SNP for the five abnormal genomes. Combined with the critical value of EMP1, the genes segment 245,102,274 among these 300 gene fragments possibly associated with the disease can be seen from Fig.2, corresponding to the 3 points above the blue line in Fig.2.

Table. 2 Risk Gene Statistics Table

GENEID	NSNP	NSIG	ISIG	EMP1	SNPS
245	14	2	2	*4.00E-04**	rs932372|rs6699113
102	10	3	3	*0.0012**	rs2273298|rs6696978|rs6541080
274	44	2	2	*0.005199**	rs9426306|rs2788891
298	23	1	1	0.012	rs12145450
235	8	1	1	0.0122	rs12028945

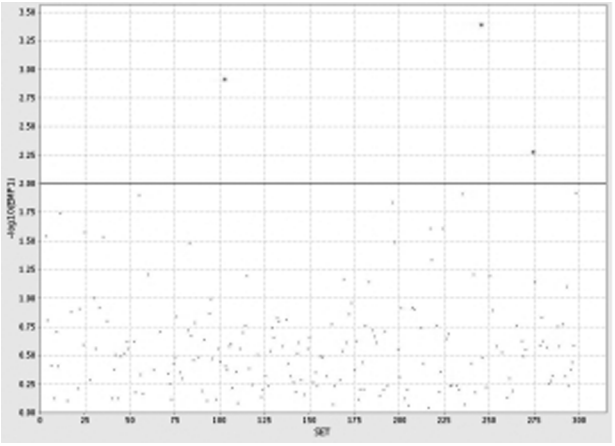

Fig. 2 Gene Scatter Plot

6.3 Analysis of Trait Correlation Loci

As shown in Table 3, Table of Traits and Loci Correlation and in Figure 3, the Scatter Plots of Related Sites: in Table 3, as the information of 10 associated traits and the coding information of 9445 loci of 1000 samples are given, the information data are substituted into the optimized model to output the Manhattan diagram shown in Figure 3. Among them, the horizontal axis represents 9445 loci, and the vertical axis also represents the negative logarithm of probabilities, namely P value. As we can seen from Figure 3, the closer to P = 1, the more dense the distribution of points is, and then with the P value decreases, the point distribution becomes more sparse. According to the actual situation, at this time the critical value of P is taken for P = 10-4. Table 3 shows the five suspected sites ranging case, we can see that less than the critical point of the P value site is only one, is the site rs35615194, that is, and at this point there are 10 traits associated with sites. Corresponding to the point above the blue line in Fig 3.

Table 3 Table of Traits and Loci Correlation

SNP	A1	F_A	F_U	A2	CHISQ	P	OR
RS35615194	C	0.3943	0.2677	T	18.12	*2.08E-05**	1.781
RS2274334	T	0.3008	0.4154	G	14.25	1.60E-04	0.6056
RS84353	A	0.2053	0.1201	G	13.37	2.56E-04	1.893
RS9439605	A	0.3028	0.4094	G	12.37	4.36E-04	0.6265
RS10754914	A	0.5061	0.3957	G	12.31	4.50E-04	1.565

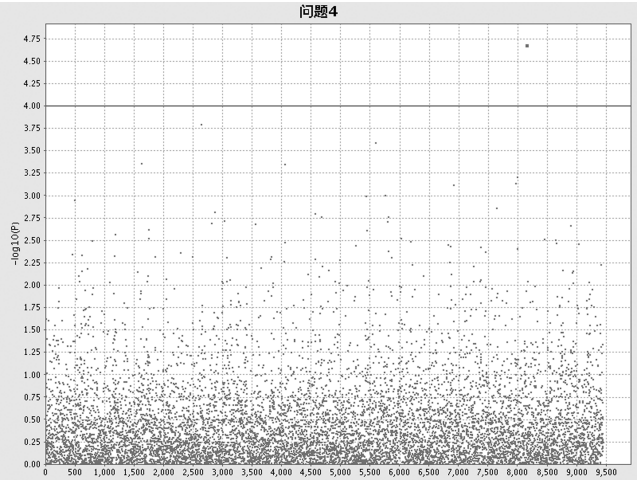

Fig.3 Scatter plots of related sites

7. Conclusion

This paper uses information of the supplied 1000 samples to analyze the loci of the gene to get SNP locus information, and uses linear mixed model to analyze the association set evolutionary algorithm to obtain the location of risk points and genes and the corresponding Manhattan diagram, which was converted into numerical computation language for calculating, and finally got the corresponding image and results. Meanwhile, a new genome-wide association analysis procedure is given and the locus associated with the traits is visualized in the form of image, and algorithm based on the original one is improved, which makes the genotype data more accurate, and theoretically shows the reasonableness of the found pathogen and gene. This method has a very important position in the future theoretical research and practical application.

References

[1]Luo Xu-hong. Genetic Linkage Analysis of Diseases at Gene Level [D]. Ningbo University, 2014.p33-35.
[2]Wu Xiaoping. Genome-wide Association Analysis and Genome Selection of Dairy Cows Based on SNP Chip and Full Sequencing Data [D]. China Agricultural University, 2014.p23-45.
[3]He Wei-Ming.Population SNP Locus Detection and Genotyping Based on Re - sequencing Data [D].

South China University of Technology, 2013.p45-50.
[4]Wang Yu. Model Selection of Linear Mixed Models [D]. Beijing University of Technology, 2003.p12-14.
[5]Ye Hui-liang. Further Study of Linear Mixed Model [J]. Chongqing University, 2013.p33-34.
[6]Zhao Yan-yan. Statistical Analysis Method of Large Scale Data and Its Application Based on Linear Mixed Model [D]. Southwest Jiao tong University, 2013 (4). p23-25.
[7]YE Hui-liang. Further Study of Linear Mixed Models [J]. Chongqing University, 2013. p33-34.
[8]Zhao Jin-fang. Application of Repeated Linear Mixed Model in Medical Research [D]. Shanxi Medical University, 2002 (2). p13-14.
[9]Ye Dao-jun.Study on the Properties of SNP Disease Model and the Comparison of Detection [D] Xi'an University of Electronic Science and Technology, 2011.p34-45.
[10]Wang DG,Fan JB,Siao CJ,et al.Large-scale Identification, Mapping, and Genotyping of single-nucleotide Poly Morphisms in the Human Genome.Science,1998,280:p1077-1082.
[11]Li W. Sadler LA.Low Nucleotide Diversity in Man. Genetics, 1991, 129:p513-523.

Author Brief and Sponsors or Supporters.

Received date: 2016-10-18
About the Author: Liu Yang (1990-),male,Shanxi people,Current PhD student,research direction:Computer Software Theory and Research.E-mail: 365361068@qq.com