Similarity study of single nucleotide polymorphism (SNPs) data

Wenjun Yu, Bin Hu and Yu Huang*
College of informatics, Huazhong Agricultural University, Wuhan, Hubei, China

*Corresponding author e-mail: yhuang@mail.hzau.edu.cn

Abstract. Based on the classification of patients, the analysis of genetic data has important complementary significance for predicting the progress of patients' diseases and subsequent treatment. Massive sequencing data provides the basis for genetic analysis. We used GAMETES to simulate single-nucleotide polymorphisms (SNPS) data, and proposed correlation clustering analysis algorithms to provide a scientific basis for understanding the consistency of clinical data and genetic data.

1. Introduction
With the rapid development of medical and information technology, it is necessary to accurately predict the development of patients' future diseases. In order to prevent diseases and early diagnosis, and to provide accurate post-treatment to patients, research on clinical data and sequencing data has important social value and scientific significance [1]. The data in various medical databases currently include a wealth of molecular biology and genomics information, and the classification of the information for clinical disease prediction can make the treatment plan more targeted and rational use of relevant medical resources. This is inseparable from the processing, mining, integration and classification of data. The classification (similarity) research on clinical data and genotype data is particularly important. This paper mainly studies the classification of genetic data, which are SNPs data.

The problem of similarity metrics for specific types of genetic data (so-called SNP data) comes from the GENICA (Interdisciplinary Study Group on Gene Environment Interaction and Breast Cancer in Germany) breast cancer Case-control study. The GENICA study aims to investigate the effects and interactions of SNP loci and exogenous risk factors. Based on this, we can study whether the similarity of SNPs data and the similarity of clinical data are consistent. If there is consistency, this will greatly broaden our understanding of the underlying biological mechanisms of complex diseases and the lack of genetic information. Moreover, other relevant research results indicate that genetic interactions emphasized by convergent biological functions caused by environmental and lifestyle factors are important determinants of complexity and comorbidity. Therefore, we can also discover new genetic variants through their consistency studies, as well as their genetic interactions and cooperative regulatory mechanisms, which will facilitate the development of many applications, such as accurate biological validation, drug repositioning, and prediction of disease progression.

The main principle of genetic data similarity analysis is to establish similar groups by measuring the distance between pairs of SNPs data. In recent years, research on the similarity of SNPs data has emerged in an endless stream, involving various infectious diseases, metabolic diseases, heart diseases, skin diseases, and nervous system diseases. For example, in predicting heart failure, the similarity classification of genetic data has achieved good predictive effects [2-5]. In predicting the survival rate
of kidney transplantation [6-9], the similarity algorithm of SNPs data provided by related authors is not only improved existing prediction model, and the prediction efficiency is also improved. The technical measures adopted are basically similar. The main problem is that the genetic data used is different, which is expressed: 1. SNPs sites; 2. The total amount of data. Sufficient data can greatly improve the accuracy of SNPs data similarity studies.

Based on the simulated clinical data of 5000 patients, effective SNPs on 22 pairs of chromosomes were simulated by GAMETES. The similarity algorithm we presented was used to calculate the similarity between pairs of patients, and the classification was performed accurately, which laid the foundation for the similarity comparison between genetic data and clinical data.

2. Method

2.1. SNPs data simulation

GAMETES [10] contains algorithms that can generate complex SNP disease models for simulation studies. Furthermore, GAMETES can quickly and accurately generate random n-locus models with specific genetic constraints, including heritability, secondary to SNPs, allele frequency and population prevalence. In addition, GAMETES includes a simple dataset simulation strategy that can be used to quickly generate an archive of simulated datasets for a given genetic model. Based on this, we generated the SNPs data by GAMETES, which contain 214 valid SNPs on 22 pairs of chromosomes, as shown below:

14 155151086 rs5983826 T C . . PR GT 0/0 0/0 1/1…

The data file is in VCF (Variant Call Format) format, from left to right:

(1) Chromosome code/name, which mutation site is found from which chromosome segment of the reference sequence;
(2) POS: Base-pair coordinate, base pair coordinates, ectopic point relative to the leftmost position of the reference genome (belonging to the 1-coordinate system: counting from 1);
(3) ID: Variant identifier, the name of the variant site (corresponding to the ID in the dbSNP database; if not, the default is to indicate that he is a novel variant);
(4) REF: Allele 2 code (missing='N');
(5) ALT: Allele 1 code (missing='.'), the base type and number of the position variation, separated by commas; for the SNP is a single base change, for InDel is the number of base changes;
(6) FILTER: Left blank ('.'), whether the next site is to be filtered out. If PASS is displayed, indicating that the next site is consistent with the reference sequence, then this site has a greater probability of being a variant site;
(7) INFO: Normally 'PR'; '.' when --real-ref-alleles specified, with the description to understand additional information about the site (including the most information, in the form of Tag=Value, semicolon separated);
(8) FORMAT: 'GT' (signaling the presence of genotype calls), variant site format, GT: sample genotype (genotype), separated by two lines (here 1/1), indicating diploid samples genotype. 0 represents the allele of the ref in the sample, 1 represents the allele of the sample variant, and 2 represents the allele of the second variant. 0/0 means that the site is homozygous in the sample, consistent with ref; 0/1 means that the site is heterozygous in the sample, there are two genotypes of ref and variant; 1/1 means homozygous for the site, and variant Consistent. A total of 214 pairs.

2.2. SNPs data filtering

Based on the computational requirements, we need to filter the above data files, leaving only the patient serial number and SNPs value pairs. For a diploid organism, the GT value represents the two alleles carried by this sample at this site. 0 means the same as REF; 1 means the same as ALT; 2 means the second ALT. When there is only one ALT allele, 0/0 means pure sum and is consistent with REF; 0/1 means heterozygous, two allele one is ALT and one is REF; 1/1 means pure sum and both are ALT.

For the above reasons, the transformation rule is as follows: add the minor allele of the two points (allele) to a number, so 0/0=0, no minor allele, 1/1=2, there are two minor alleles, 0/1=1/0=1, does not
distinguish whether this allele comes from father or mother. Any missing or incorrect sites are always expressed as 0. The final data file is a CVS (Comma-separated values) structure, as shown in the following figure:

Fig1. SNPs file, the leftmost column represents the patient number.

2.3. Similarity algorithm and analysis results

We perform a search for patterns in the current data set through various clustering and classification methods. Here we only consider the problem of proper measurement of the proximity of two variables or subjects as an essential basis for further cluster analysis [11-15].

In general, clustering methods are useful tools for detecting structures and generating assumptions about potential relationships in complex data environments. There are two possible goals for searching patterns in data: identifying groups of similar objects or topics or groups that identify similar variables throughout or within a subpopulation. Comparing the individual gene profiles and comparing the genetic information of subpopulations, we discuss possible choices for similarity measures, especially similarity measures based on matching and mismatch counts. Based on this, we introduce a new matching coefficient using a more flexible weighting scheme to solve the general problem of SNP data comparison: most homozygous reference sequences relative to homologous and heterozygous SNPs mask the level and difference of interest. The calculation results are shown in the following table:

| Pathway                  | Number of SNPs |
|--------------------------|----------------|
| DNA repair               | 12, 14, 17, 18, 19, 23 | 35 |
| Signal transduction      | 33, 64, 75, 76,   | 90 |
| Growth factors           | 70             | 41 |
| Oncogene                 | 31             | 101 |
| Metabolism of steroid hormones | 101, 102, 105, CYP1A1, CYP1B1 | 203 |

3. Discussion

The SNPs similarity calculation method we proposed is a promising tool for detecting the general structure of SNPs data and finding potential differences between cases and controls, especially the assumptions between cases and controls. The traditional matching coefficient [16] is proposed by Müller et al., and the comparison of other similarity coefficients and specific flexible matching coefficients is more detailed. Adding new variables seems to have little effect on the general structure of the tree, so the applied measurements seem to produce a conservative structure.
In general, cluster analysis helps to understand the data in depth, but in a complex data set, it can be reasonably combined with other methods. In order to detect interactions between loci and exogenous factors, there are many other classification methods, such as classification trees, collection methods, SVM, multi-dimensional reduction and logistic regression, designed to identify the "best" combination of features. Due to the heterogeneity of the case group, the low penetrance of the relevant genetic variants, and the number of competing models, these SNPs data methods often produce high false classification rates [17-20].

Based on the above, we proposed combined clustering and classification methods to obtain more insights and validate relevant biological hypotheses by pre-selecting variables or by combining several methods to suggest potential impact factors.

4. Acknowledgments

This work was financially supported by the Fundamental Research Funds for the Central Universities (Program No. 2662015PY057) and the State Scholarship Fund (No. 201706765027).

References

[1] Krapohl E, Plomin R. Genetic link between family socioeconomic status and children’s educational achievement estimated from genome-wide SNPs[J]. Molecular psychiatry, 2016, 21(3): 437.

[2] Li L, Cheng W Y, Glicksberg B S, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity[J]. Science translational medicine, 2015, 7(311): 311ra174-311ra174.

[3] Dozmorov M G, Cara L R, Giles C B, et al. GenomeRunner web server: regulatory similarity and differences define the functional impact of SNP sets[J]. Bioinformatics, 2016, 32(15): 2256-2263.

[4] Knowles L L, Massatti R, He Q, et al. Quantifying the similarity between genes and geography across Alaska’s alpine small mammals[J]. Journal of Biogeography, 2016, 43(7): 1464-1476.

[5] Elbsayoni I S, Lorenz A J, Guttieri M, et al. A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat[J]. Plant science, 2018, 270: 123-130.

[6] Wang L, Oehlers S H, Espenschied S T, et al. CPAG: software for leveraging pleiotropy in GWAS to reveal similarity between human traits links plasma fatty acids and intestinal inflammation[J]. Genome biology, 2015, 16(1): 190.

[7] Roshyara N R, Scholz M. Impact of genetic similarity on imputation accuracy[J]. BMC genetics, 2015, 16(1): 90.

[8] Sahl J W, Lemmer D, Travis J, et al. NASP: an accurate, rapid method for the identification of SNPs in WGS datasets that supports flexible input and output formats[J]. Microbial genomics, 2016, 2(8).

[9] Cao S, Loladze A, Yuan Y, et al. Genome-wide analysis of tar spot complex resistance in maize using genotyping-by-sequencing SNPs and whole-genome prediction[J]. The plant genome, 2017.

[10] Urbanowicz R J, Kiralis J, Sinnott-Armstrong N A, et al. GAMEetes: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures[J]. BioData mining, 2012, 5(1): 16.

[11] Coelho L P, Kultima J R, Costea P I, et al. Similarity of the dog and human gut microbiomes in gene content and response to diet[J]. Microbiome, 2018, 6(1): 72.

[12] Capobianchi M R, Uleri E, Caglioti C, et al. Type I IFN family members: similarity, differences and interaction[J]. Cytokine & growth factor reviews, 2015, 26(2): 103-111.

[13] Sahl J W, Lemmer D, Travis J, et al. The Northern Arizona SNP Pipeline (NASP): accurate, flexible, and rapid identification of SNPs in WGS datasets[J]. BioRxiv, 2016: 037267.

[14] Cremers R G, Galesloot T E, Aben K K, et al. Known susceptibility SNPs for sporadic prostate cancer show a similar association with “hereditary” prostate cancer[J]. The Prostate, 2015, 75(5): 474-483.

[15] Okawa K, Ohno M, Kashimura A, et al. Loss and gain of human acidic mammalian chitinase activity by nonsynonymous SNPs[J]. Molecular biology and evolution, 2016, 33(12): 3183-3193.
[16] Laucou V, Launay A, Bacilieri R, et al. Extended diversity analysis of cultivated grapevine Vitis vinifera with 10K genome-wide SNPs[J]. PloS one, 2018, 13(2): e0192540.
[17] Zhao G, Marceau R, Zhang D, et al. Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression[J]. Genetics, 2015, 199(3): 695-710.
[18] Chen X, Sun Y Z, Guan N N, et al. Computational models for lncRNA function prediction and functional similarity calculation[J]. Briefings in functional genomics, 2018, 18(1): 58-82.
[19] Jin Z, Chung J W, Mei W, et al. Regulation of nuclear–cytoplasmic shuttling and function of Family with sequence similarity 13, member A (Fam13a), by B56-containing PP2As and Akt[J]. Molecular biology of the cell, 2015, 26(6): 1160-1173.
[20] Dussex N, Chuah A, Waters J M. Genome-wide SNPs reveal fine-scale differentiation among wingless alpine stonefly populations and introgression between winged and wingless forms[J]. Evolution, 2016, 70(1): 38-47.