A brief overview of GWAS: discover genetic variations of diseases and phenotypes

Zhiying Peng¹

¹School of life sciences, Henan University, China

Abstract. GWAS, or Genome-wide association study, is a statistical analysis method to reveal specific genetic variations, usually single nucleotide polymorphisms, with particular phenotypes or diseases. The power to scan whole genomes from large scale samples made the method an efficient tool for information discovery. In the last decades, the application of GWAS has flourished, which benefited our understanding related to diseases, breeding and many other topics. In this review, we overviewed the history of GWAS, as well as different approaches to perform the analysis under different circumstances during different stages. Meanwhile, we also showed how different GWAS approaches benefited diverse research and application fields, and the potential limitations of the method.

1 introduction

Single nucleotide variation is the most common genetic variation in human genome. The genetic diversity caused by single nucleotide variation is called Single Nucleotide Polymorphism (SNP), which only involves the variation of a single base. This variation can be caused by the conversion or transversion of a single base, or by the insertion or deletion of a base. This polymorphism accounts for more than 90% of all known genetic polymorphisms. In terms of the influence on the characters, SNP can be divided into two types: One is non-synonymous SNP. In this case, the change of base sequence will lead to the change of protein sequence translated from it as a template, and then it may have an impact on the function of protein. Non-synonymous SNP is often the direct cause of biological character transformation and disease. The other is synonymous SNP, that is, the change of coding sequence of single base does not affect the amino acid sequence of the translated protein. However, due to the existence of linkage disequilibrium, this kind of SNP has become the most extensive and important specimen label in gene association research, and plays an important role in candidate gene association research and whole genome association research.

Gene association research is used to detect the correlation between disease status and genetic variation. Its purpose is to identify the candidate genes or genomic regions that cause specific diseases. With the completion of the Human Genome Project in 2003 and the International HapMap Project[1] in 2005, now we have a set of computer databases which contains the reference human genome sequence and human genetic variation

map. With the emergence of new technology of gene association research, these databases can quickly and accurately analyze the genetic variation in the whole genome samples that lead to the onset of disease, and help researchers find the genetic contribution of common diseases.

Genome-wide association study (GWAS) is a gene association research method which flourishes with the appearance of high-throughput genotyping, combining the Human Genome Project (HGP) and the International Haplotype Map (HapMap) Project. The publication of GWAS studies first emerged in 2002. In that year, Tanaka group and Chakravarti group conducted GWAS analysis on myocardial infarction[2] and Hirschsprung disease[3] respectively, both of which were published on Nature Genetics. These studies inspired the publication of a GWAS landmark study in 2005. In the study published on Science, Klein et al. identified 2 SNPs related to age-related macular degeneration[4]. Later, GWAS research reports of coronary heart disease[5], obesity[6], type 2 diabetes[7], triglyceride[8], schizophrenia[9] were also published. With the development of genomics and microarray technology, a large number of genetic variations associated with complex traits have been identified by GWAS. By the end of 2010, more than 1000 GWAS articles had been published (Figure 1).

Figure 1. Publication counts grouped by year on the topic of GWAS on PubMed. Keyword: (GWAS[Title/Abstract]) or (Genome-wide association study [Title/Abstract])

GWAS analysis method is to find out the SNPs that cause causal variations, the variations that lead to trait change or disease, through the overall association analysis of genome [10]. GWAS usually uses the regression analysis:

$$y = u + bx + e$$

in which y is the individual trait or phenotype, u is the average of the overall phenotypes, e is the environmental factor, x is the genotype ($x=0,1,2$), and b is the genotype coefficient. $x=0$ means that there is no meaningful SNP. $x=1$ indicates that the gene has one meaningful mutation due to SNP. And, $x=2$ means that both alleles of the gene possesses the SNP mutation. This formula shows that when the trait is not related to the SNP to be detected, the individual phenotype is the overall average phenotype; when a trait is related to SNPs, the individual phenotype and the average phenotype are significantly different, and the difference would be b or $2b$. By traversing all SNPs, we could get the SNPs related to the target traits in the whole genome. On the basis of the principles above, GWAS has derived many different methods according to the research object, process and gene typing methods. Some of these different methods are different due to different research objectives, while some are gradually formed different methods due to the development of technology. In this paper, we will briefly introduce and compare different methods and introduce their practical applications, to provide a picture showing the development path and current progression of GWAS.

2 GWAS approaches for different research objectives

Based on the internal connections between the research objectives, GWAS can be divided into two categories: the study of unrelated objectives and the study of families.

2.1 Correlation analysis for unrelated objectives

In the analysis based on unrelated objectives, we would collect a series of samples related to target traits and a series of control objectives together from a population group for analyze. Among them, case-control analysis is a straightforward method that has been widely used.

The case-control analysis method divides the samples

with and without target traits (mainly diseases) into the case group and the control group respectively. Then, the phenotype distributions and SNP differences in the whole genome of the case group and the control group are detected by GWAS methods with chi-squared test or regression analysis, and the differences in gene frequency between the two groups are compared to determine one or more genetic variations or SNPs related to the disease or phenotype. If there is a significantly difference in SNP between the two groups, it will mean that the trait or disease is related to the genetic variation. This method is often used in the study of human susceptible diseases.

Clinically, case-control analysis includes collecting a series of disease cases and control groups with sample homogeneity controlled by strict clinical standards, and a number of negative cases as the control group. For example, in 2005, Steer et al. selected 302 cases of white British with rheumatoid arthritis (RA) in London as the case group, and 374 healthy white British as the control group to study the association between 1858C/T SNP and RA in the white British population[11]. Through strict quality control and Hardy-Weinberg equilibrium test, Steer et al. concluded that the single nucleotide polymorphism had a significant impact on the risk of RA after the odds ration calculation.

Case-control analysis focuses on qualitative traits. It is difficult to find the target SNP by case-control analysis for some complex quantitative traits or traits jointly determined by environment and genes. In this situation, the random group analysis method would be preferred.

In beef cattle breeding, feed cost accounts for around 70% of the total cost. Therefore, one way to reduce the cost of beef cattle breeding is to improve feed efficiency[12]. Residual Feed Intake (RFI) is an indicator of feed efficiency. The lower the RFI is, the higher the feed efficiency is. Heras-Saldana et al. used the whole genome association to investigate the association between SNPs and RFI in Angus beef cattle in Australia by random population analysis[13]. They randomly genotyped 2190 Angus beef cattle, and identified the RFI related Quantitative Trait Loci (QTL) with GWAS to determine the genes significantly related to this trait. Heras-Saldana et al. found that the expression of 15 genes have significant impacts on RFI. The results of this study can help to improve feed efficiency through beef cattle breeding and increase the economic benefits.

2.2 Family-based correlation analysis

When the investigated individuals are genetically associated, the analysis method based on random samples, including case-control analysis, is prone to become false-positive. This is because for the GWAS regression formula $y = u + bx + e$, if there is a genetic association, there may be other unknown relationships making $bx = b_i x_i + b_j x_j$, in which x_i is the genotype due to known trait SNP (tSNP) and x_j is the genotype due to the unknown genetic variation with family association. When tSNP does not affect the traits, $b_i = 0$. However, the genetic variation of unknown family association b_j and x_j

may make the genotype coefficient bx not zero and thus cause the false-positive and population stratification. Population stratification, specifically, refers to that there are subgroups with different allele frequencies in the group. It can be caused by different factors such as artificial selection, geographical isolation, reproductive isolation, natural evolution and so on, which makes the target traits seemed correlated with unrelated loci, and finally leads to false-positive results. In order to solve this problem, one method is to correct the stratification through genome control [14] and other methods. The other way is to avoid the problem through association research based on a single family. Among them, the most widely used method is using Transmission Disequilibrium Test (TDT) [15] to analyze the relationship between SNPs and disease quality and quantitative phenotype. On the other hand, Cheikh Loucoubar et al. proposed M-TDT (Multipoint Transmission Disequilibrium Test)[16], which is a family based multiple allelic effect method extended from the original TDT to detect qualitative or quantitative traits, and can also avoid the influence of stratification.

3 GWAS of different phases

GWAS design can be divided into single-stage, two-stage and multi-stage. In single-stage analysis, researchers will collect enough samples to genotype all the target SNPs in all the subjects at one time, and then analyze the correlation between each SNP and target traits. However, practice showed that if a large number of SNPs needed to be genotyped and tested, the cost would be huge. In this contradiction, the two-stage research method came into being. The two-stage research method can significantly reduce the total number of genotyping while marginally reducing the accuracy of the results, thus greatly reducing the cost of detection. Generally, it can reduce the cost of large-scale correlation research by half, while improving the ability for disease-related markers detection[17]. The two-stage research method means that in the first stage, we analyze genome-wide SNPs in a small number of samples and screen out the most significant SNPs. Then, we genotype only the significant SNPs in a larger number of samples in the second stage. Finally, the results of the two stages are analyzed together. These two phases can be divided into more fine-grained ones, or multi-stage.

The two-stage approach was proposed by Jaya M. Satagopan and Robert C. Elston[18] in 2003. In their study, they used a two-stage approach to study the associations between genetic disease and a large number of candidate SNPs in case-control design experiments. When a large number of SNPs are tested with case-control samples, the two-stage design is more effective to identify disease-related markers.

The two-stage method has been widely used since then. For example, Adriano Chio[19] et al. analyzed 545,066 SNPs in 553 patients with sporadic Amyotrophic Lateral Sclerosis (ALS) and 2,338 controls using a two-stage GWAS method. In the first stage, they

tested 555,352 SNPs in 553 patients and 2,338 controls, and got 7,600 SNPs that may be related to ALS. In the second stage, they further analyzed the 7,600 SNPs. The results showed that the correlation between these SNPs and ALS did not reach the Bonferroni threshold significance, indicating that ALS might have greater genetic and clinical heterogeneity.

4 GWAS based on different genotyping methods

No matter how the research object is selected or the experimental process is designed, GWAS needs to obtain SNP information by genotyping.

Genotyping obtains or checks the DNA sequence of an individual with biological experiments, thus providing information for genetic composition (genotype) comparison. With genotyping, we can reveal what alleles individuals inherit from their parents. The development of genotyping depends on the gene sequencing technology. Since the era of high-throughput sequencing, gene sequencing has experienced the development from the gene chip technology to Next Generation Sequencing (NGS). The genotyping methods used by GWAS can also be classified according to the same development stage.

4.1 Genotyping based on the gene chip technology

The essence of gene chip is to hybridize the target DNA sequence with the known DNA sequence. The hybridization results indicate the sequence of the target DNA. Affymetrix or Illumina's gene chip technology platform can detect and analyze a large number of samples at one time by fixing a great many DNA probes at the same time. Compared to traditional nucleotide (and protein) blotting technologies (Southern Blotting, Northern Blotting, etc.) and quantitative PCR (TaqMan probing), gene chip technology is more friendly and higher automated, and can achieve better accuracy and efficiency[20]. To be noted, the gene chip technology can be used not only for SNP genotyping, but also for expression profile analysis of high-throughput mutation detection and chromosomal aberration detection.

In 2013, in order to identify molecular markers and candidate genes related to body composition and meat quality traits in edible broilers, Liu et al. conducted a genome-wide association study on the 724 genotypes of Beijing-You Chickens by using the Illumina 60K Single Nucleotide Polymorphism (SNP) gene chip[21]. Among the ten body composition traits studied, nine were detected significantly correlated with SNPs. For example, 7 SNPs and 4 candidate genes (LCORL, LAP3, LDB2, TAPT1) in a continuous region of Chicken Chromosome 4 (GGA4) were related to carcass weight and eviscerated weight. In addition, a 0.65mb region on GGA3 was related to the pectoralis weight and percentage. By measuring the mRNA level of five different genes in this region, Liu et al. found that the change of gene GJA1 was consistent with the change of

pectoralis weight during development. Thus, Liu et al. proposed that the GJA1 gene is likely to be a functional gene affecting the development of chicken pectoralis. As the body composition and meat quality of broilers are important economic traits, the information obtained from this study is helpful to breeding, and also to the further understanding of molecular mechanism of broiler target traits.

4.2 Genotyping based on NGS

NGS sequences DNA by PCR and the reversible termination of SBS (Sequencing by Synthesis) technology using four special deoxyribonucleotides with fluorophores. Because the PCR amplification step during NGS library construction, NGS is more sensitive than gene chips on the whole. But at the same time, due to the imbalance of PCR amplification process, compared with gene-chip-based genotyping, NGS may lead to a higher error rate, and the ratio between the fragments in the sample may often be destroyed. However, compared with gene chip method, applying NGS in GWAS has many new advantages, including higher throughput sequencing, more specific applicability and higher reading quality of the sequencing data set[22], etc. Moreover, it can find a large number of rare mutations outside the common SNP and mutation sites not in the database, so as to find disease-related variation sites more effectively.

With the popularization of NGS technology, the application of NGS in GWAS is gradually increasing. In 2019, Farhad et al. conducted GWAS analysis of 12 million Single Nucleotide Polymorphisms (SNPs) from 1,252 Simmental beef cattle using NGS technology to find the SNPs that affected their muscle characteristics[23]. In the region of cattle's BTA4 chromosome related to hind leg width, Farhad et al. identified 202 SNPs significantly related to muscle properties, and found candidate genes that may have potential impact on muscle development in the region around these SNPs. Muscle growth is an important economic factor in beef industry, and the discovery about the internal mechanism of muscle growth and metabolism could help better understand the genes that regulate traits related to the potential metabolism of muscle growth.

5 Conclusion

In the past 20 years, GWAS technology has been making tremendous progress with the help of the completion of HGP and HapMap. In this paper, we summarized the technical development of GWAS from three different perspectives. These summaries covered the possible design, methods and application scenarios of GWAS research, and showed the application value of different GWAS methods through examples.

It is worth mentioning that the summary in this paper only covers the classic classification methods, and there are also a great many innovative methods carried out in practice. For example, the two-stage GWAS method

proposed in the second part of this paper has experienced the replication-based analysis, a separate analysis of the samples of each stage, to a more powerful joint 2-stage analysis[24].

On the other hand, although the GWAS analysis is powerful and can help solve many problems of SNP mining, it still has some limitations in practical application. For example, linkage disequilibrium may make it difficult for GWAS to precisely locate causal loci[25]. And, the requirement of data quantity makes it difficult for GWAS to find low-frequency variation related loci[26] and predict individual phenotypes[27]. To solve these problems, some scholars are trying to improve GWAS's power by means of multiple trait association[28] or GWAS-Transcriptome Wide Association Study (TWAS)[29]. We can still expect that this genome level SNP mining technology, which has appeared for nearly 20 years, will still help us to find more disease and trait related genes.

Reference

- 1 C. M. Lewis, J. Knight, Cold Spring Harb Protoc, **2012**(3):297-306(2012)
- 2 K. Ozaki, Y. Ohnishi, A. Iida, A. Sekine, R. Yamada, et al. Nat Genet, **32**(2002)
- 3 M. M. Carrasquillo, A. S. McCallion, E. G. Puffenberger, C. S. Kashuk, N. Nouri, et al. Nat Genet, **32**(2002)
- 4 R. J. Klein, Science, **308**(2005)
- 5 J. Erdmann, T. Kessler, L. M. Venegas, H. Schunkert, Cardiovasc Res, **114**(2018)
- 6 J. R. Speakman, R. J. F. Loos, S. O'Rahilly, J. N. Hirschhorn, D. B. Allison, Int J Obes(Lond), **42**(2018)
- 7 R. M. Watanabe, Methods Mol Biol, **1706**(2018)
- 8 J. S. Dron, R. A. Hegele, Curr Atheroscler Rep, **19**(2017)
- 9 D. Rujescu, Nervenarzt, **88**(2017)
- 10 A. G. Uitterlinden, Semin Reprod Med, **34**(2016)
- 11 S. Steer, B. Lad, J. A. Grumley, G. H. Kingsley, S. A. Fisher, Arthritis Rheum, **52**(2005)
- 12 P. F. Arthur, J. A. Archer, D. J. Johnston, R. M. Herd, E. C. Richardson, et al. J Anim Sci, **79**(2001)
- 13 S. De las Heras-Saldana, S. A. Clark, N. Duijvesteijn, C. Gondro, J. H. J. van der Werf, BMC Genomics, **20**(2019)
- 14 A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, et al. Nat Genet, **38**(2006)
- 15 R. S. Spielman, R. E. McGinnis, W. J. Ewens, Am J Hum Genet, **52**(1993)
- 16 L. Cheikh, A. V. Grant, J.-F. Bureau, I. Casademont,

- N. A. Bar, et al. *Brief Bioinform*, **18**(2016)
- 17 J. M. Satagopan, E. S. Venkatraman, C. B. Begg, *Biometrics*, **60**(2004)
 - 18 J. M. Satagopan, R. C. Elston, *Genet Epidemiol*, **25**(2003)
 - 19 A. Chiò, J. C. Schymick, G. Restagno, S. W. Scholz, F. Lombardo, et al. *Hum Mol Genet*, **18**(2009)
 - 20 P. C. Hayes, C. R. Wolf, J. D. Hayes, *BMJ*, **299**(1989)
 - 21 R. Liu, Y. Sun, G. Zhao, F. Wang, D. Wu, et al. *PLoS ONE*, **8**(2013)
 - 22 J. Reis-Filho, *Breast Cancer Res*, **11**(2009)
 - 23 F. Bordbar, J. Jensen, B. Zhu, Z. Wang, L. Xu, *PloS one*, **14**(2019)
 - 24 A. D. Skol, L. J. Scott, G. R. Abecasis, M. Boehnke, *Nat Genet*, **38**(2006)
 - 25 D. Altshuler, M. J. Daly, E. S. Lander, *Science*, **322**(2008)
 - 26 J. Flannick, *Curr Diab Rep*, **19**(2019)
 - 27 U. M. Marigorta, J. A. Rodríguez, G. Gibson, A. Navarro, *Trends Genet*, **34**(2018)
 - 28 P. Turley, R. K. Walters, O. Maghzian, A. Okbay, J. J. Lee, et al. *Nat Genet*, **51** (2019)
 - 29 A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, et al. *Nat Genet*, **48**(2016)