NEW ERA AGRICULTURE MAGAZINE

A Research Approach to Identify Genomic Variants: GWAS

Piyusha Singh[1], Akanksha Tiwari[1], Ajeet Kumar Gupta[2], Rishab Gupta[2], Javed[2], Gyan Prakesh Pandey[2], Raja Bhaiya[2]

Introduction:

A genome-wide association study (GWAS) is a research approach used to identify genomic variants that are statistically associated with a risk for a disease or a particular trait. The method involves surveying the genomes of many people, looking for genomic variants that occur more frequently in those with a specific disease or trait compared to those without the disease or trait. Once such genomic variants are identified, they are typically used to search for nearby variants that contribute directly to the disease or trait. Genome-Wide Association Study, GWAS. The goal of genome-wide association studies is to screen the entire genome of large numbers of individuals to look for associations between millions of genetic variants within those individuals and their disease outcomes or sometimes for associations between the variants and non-disease trait such as height. The first GWAS was published in 2005 and after that, the study approach just took off exponentially. Over time, GWAS have grown significantly both in terms of sample size going from initial sample sizes of several thousand individuals to current sample sizes of tens and hundreds of thousands of individuals and in terms of the number of disease studied as well as the associated variants that have been discovered. Results from GWAS have been curated in the NHGRI-EBI GWAS catalog. The methods and results of GWAS have informed other applications in applied epidemiologic research such as gene environment studies, Mendelian randomization studies, and polygenic risk score approaches.

Current Status of GWAS

Most traits of agricultural and evolutionary importance are complex traits that are influenced by many genetic loci and environmental conditions as well as their interaction. Advances in genomic technology and methodology development and a desire to examine trait variation across diverse genetic backgrounds were the major driving forces behind the initial wave of association mapping studies in model plant and crop species. Continued progress in sequencing technologies

Piyusha Singh[1], Akanksha Tiwari[1], Ajeet Kumar Gupta[2], Rishab Gupta[2], Javed[2], Gyan Prakesh Pandey[2], Raja Bhaiya[2]

[1]Assistant Professor, [2]Research Schloar,

Acharya Narendra Deva University of Agriculture and Technology, Kumarganj, Ayodhya, U.P.

and coordinated community effort have made genome-wide association studies (GWAS) a method of choice, particularly when resequencing is conducted after the assemblage of the reference genome or when a high-density genotyping array becomes available. High-throughput phenotyping has also been expanding the trait list for GWAS.

Genome-wide association studies have investigated agriculturally important traits in many major crop species, including maize (*Zea mays* L.), wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.), soybean [*Glycine max* (L.) Merr.], sorghum [*Sorghum bicolor* (L.) Moench], barley (*Hordeum vulgare* L.), cotton (*Gossypium hirsutum* L.), and numerous other crops beyond the model plant species *Arabidopsis*. Genome-wide association studies have identified genomic regions associated with many agronomic, physiological, and fitness traits including flowering time, plant height, kernel number, stress tolerance, and grain yield

- Genome-wide association studies have also been used to study other types of phenotypes. Genome-wide association studies in rice have identified genes associated with geographical divergence and adaptation during domestication as well as with biochemical and molecular phenotypes including flavonoid, fatty acid, amino acid, and nucleic acid metabolites. Data generated by high-throughput automated phenotyping have also been analyzed by GWAS.
- Genome-wide association studies are used both to detect novel associations with valuable traits and to validate loci identified by other methods. Genome-wide association studies may be conducted as stand-alone investigations, as a component of gene cloning studies, or as the foundational step in marker-assisted selection, among other uses. In turn, exploiting this information accelerates crop breeding. For example, loci identified by GWAS on provitamin A levels in maize grain were used as the basis of marker-assisted and genomic selection for this important nutritional trait.

Opportunities of GWAS

The challenge posed by loci with low minor allele frequency was known when the GWAS approach was initially proposed. Some researchers remove variants with minor allele frequency below 5% before performing GWAS. Their argument is that because statistical power is very low for these rare alleles, preventing identification unless their effect size is extremely large, the large number of these variants only exacerbates the multiple-testing issue. However, the unequal sample size of two alleles of these variants is already considered in the test statistics the same way

as other variants and there is no a priori reason why a rare allele should not be biologically important. In fact, because of purifying selection, many deleterious alleles will be present at low frequencies. Many new statistical models have been designed to test the rare variants, often by aggregating nearby rare variants and testing their combined effects. Many tests designed for rare alleles should be implemented in software packages. Unless including many variants with low minor allele frequency inflates the genome-wide significance threshold as a result of multiple testing, we recommend the testing of variants with low minor allele frequency.

Synthetic associations are misleading associations that occur when GWAS identifies noncausal SNPs as more significant than truly causal variants. The most significant peak may actually be located in a different LD block from the true gene, making the gene very difficult to identify. This may happen in the case of allelic heterogeneity, when multiple independent alleles of a given gene are present in a population. If each allele affects the phenotype similarly, none of the alleles responsible will be perfectly correlated with the trait of interest and so their tagging SNPs may not be detected by GWAS. However, there may be SNPs in a different location that are associated with the presence or absence of all alleles responsible. These SNPs may then be detected as synthetic associations. For example, in the case of the *Hd1* gene that controls days to heading in rice, allelic heterogeneity prevented SNPs in the true gene from being significant. However, the adjacent linkage block included SNPs that were well correlated with functional vs. nonfunctional versions of the gene and therefore could be detected. If the true causal alleles are rare and therefore already difficult to detect, this problem can be exacerbated. Rare alleles can also produce synthetic associations even in the absence of allelic heterogeneity. For example, sickle cell anemia is controlled by a single rare allele, but Dickson et al. showed that common variants in other locations in the genome are significantly associated with this allele by chance. Genome-wide associations studies may then detect these common SNPs as synthetic associations. Genome-wide association study approaches based on genes or regions rather than SNPs have been helpful in addressing this problem. However, more work remains to be done in developing these methods, particularly because several of the methods developed have yet to be implemented in freely available software.

Because of these challenges, other methods of setting significance thresholds for GWAS have been developed. In one example, Simple M addresses the dependency among markers by calculating the number of effective

markers and then using as the genome-wide significant threshold. The number of effective markers is obtained as the number of principal components that cumulatively capture a high percentage (e.g., 99.5%) of variance in the pairwise correlation matrix for all SNPs, which is derived from the composite LD among SNPs. Another method, a sliding-window approach for locally inter correlated markers with asymptotic distribution errors corrected (SLIDE), is designed to account for the correlation among SNPs within a sliding window and corrects for the departure of the true null distribution of the statistic from the asymptotic distribution.

Conclusion

Genome-wide association studies have successfully identified thousands of loci associated with agronomic and other traits in crop species, and several methods have been developed to improve power and computational speed. As the development of GWAS in crops continues, it may emulate the recent progress of GWAS in human diseases. Human disease GWAS have been coupled with tools including in silico models, tissue-specific resources, and wet-lab experiments to understand the biological functions of significant loci and to determine the causal and protective alleles. This deeper understanding has led to tangible results including new drug targets and therapeutic optimization for individual patients depending on genotype. Today, especially in the major species of crops, similar validation tools are being developed or are already available. It is worthwhile to point out the synergistic relationship between GWAS and genome editing. While GWAS is a major tool to identify genes underlying complex traits, providing targets for genome editing to generate engineered alleles, improved genome editing enables the validation of gene function under different genetic backgrounds, which can inform the method research of GWAS. Ultimately, the biological understanding gained from complementing GWAS with many genomic, phenomic, biotechnological, and data analytical tools in crop species may in turn be used to produce genetically modified plants containing specific alleles known to influence desirable traits including drought resistance, increased yield, and improved nutritional quality.

- Genome-wide association studies techniques were developed beginning in the 1990s, and the first GWAS were published in the early 2000s.
- Genome-wide association studies dissect complex traits by associating genotypic variants with phenotypic variation in a large panel. Although unrelated individuals are preferred, this is often not the case when we assemble

existing samples to form the study panel.

- Because of the existence of population structure and kinship among samples within the panel, naïve GWAS analysis with simple linear regression results in many false positives.
- Genome-wide association studies methods have been developed to improve computational speed by improving the efficiency of solving the MLM equations. Other methods improve power by alternative kinship calculation methods and by including multiple markers as covariates; these methods often improve efficiency as well.
- Ongoing challenges include analyzing variants with low minor allele frequency, avoiding synthetic associations, and understanding differences among results generated by various GWAS methods.
- Candidate gene prioritization methods help in moving from GWAS results to biological understanding.
- Continued methodology development in GWAS is needed and funding support for methodology development and software implementation benefits a wide range of research disciplines.