Principles of GWAS

Table of Contents

1.Resource groups

In our research, scientists had three times GWAS with different control groups. Their study used individuals from:

1) Three tropical breeds, Senepol (SE), Carora (CR), and Romosinuano (RS)

2) Three tropical cross-bred lines consisting of Senepol ×Angus (SNG, SNGSE), Senepol ×Holstein (SHO), and Romosinuano ×Angus (RAN, ANRAN) (demonstrate the slick phenotype) The SHO animals were back-crossed to pure Holsteins over four generations or line-bred between crosses and selected for the SLICK. Overall they ranged from 40 to 97% HO according to pedigree analysis. The SNGSE and ANRAN were also the progeny of back-crossing between SNG and pure Senepol and RAN crossed back to pure Angus.

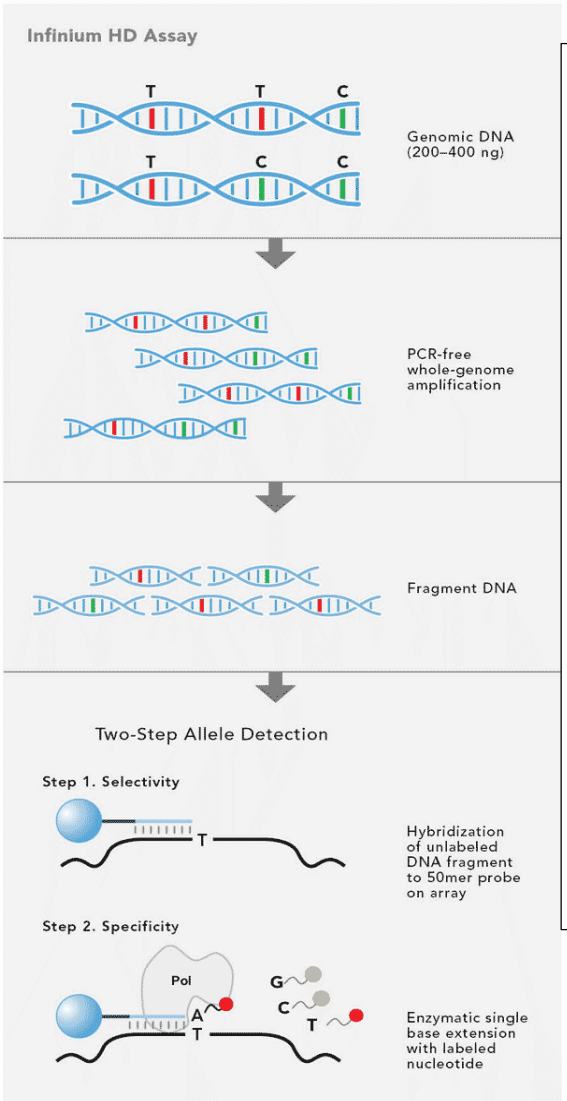
3) Additionally, individuals of the breeds Angus (AN), Red Poll (RP), N'Dama (ND), Holstein (HO), Brown Swiss (BS), and East African Zebu (ZB), all having non-SLICKs and an ancestral relationship to the slick breeds, were used for comparison in various analyses described throughout the Materials and Methods section. All animals were characterized as having a slick or wild-type hair coat based on visual observations of hair length. The SHO were additionally validated as SLICK using the weight of hair clipped. Senepol, Red Poll, and all three cross-bred lines were validated using SLICK microsatellite markers

The researchers explained “Due to the nature of breeds such as the Senepol and Carora being predominantly slick coated, ancestral breeds such as the Red Poll, N'Dama, and Brown Swiss were used to serve as non-slick control animals. We also included East African Zebu cattle as a control breed based on recent population modeling of Senepol cattle. Non-slick coated Angus and Holstein were used as controls to counter the slick coated cross-breed lineages. We hoped to minimize population stratification and reduce false SNP association to breed specific alleles by taking this approach.”[14] (Population stratification and false association will be explained more later)

In a second GWAS, they removed the Carora cases and Brown Swiss controls and in a third GWAS, they removed the Romosinuano, RAN, ANRAN, and Angus. The systematic removal of slick breeds and their ancestral groups was to identify any shifting of the GWAS peak location and the subsequent effect on the *p*-value of associated SNPs.[14]

2.Genotyping:

Technologies for genotyping developed rapidly, which dramatically increased the efficiency of GWAS. Here I want to clarify two most famous companies which all focus on chips for SNPs. These two companies all concentrate on producing chips for biological application. Basically, Illumine produce “Beadchips” whereas Affymetrix focus on “GeneChips.” Chips of Affymetrix includes whole genome chips, Axiom chips, SNP 6.0 chips, Mouse Diversity chips, and 3K, 5K, 10K, 70K specialized chips. For Illumina, their chips include Genotyping--SNP chips, Iselect/Goldengate chips, Genomeexpression chips, DALS chips and microRNA chips. Here, in our example, researchers make use of Illumina Bovine HD Beadchips for an in-depth genome-wide investigation of the slick phenotype. Then Genome Studio software was used to calculate and cluster over 777 thousand SNP marker genotypes from BovineHD.14



The Infinium HD Assay protocol features single-tube sample preparation and whole-genome amplification without PCR or ligation steps, significantly reducing labor and sample-handling errors.

Process:

1). Fix amount for DNA

2). Incubation of products of DNA amplification

3). Fragmentation of products of DNA amplification

4). Precipitation and suspension of DNA

5). Preparing BeadChips

6). Hybridization of DNA fragment to BeadChips

7). Extension and Pigmentation of BeadChips

8). Surrounding BeadChips

9). Scanning BeadChips

3. Quality Control

Data for genome-wide association studies (GWAS) demand a fair amount of preprocessing and quality control (QC), especially SNP genotypes.[49] So after genotyping, what we should consider next is data cleaning and quality control.[13]

i) Quality control for sampling. Normally, researchers will make use of "Kolmogorov-Smirnov" method to do the test of normal distribution of data. And then use Minitab 15 to do the Johnson transformation for those data which don't follow the normal distribution. Make use of Cervus procedure of maximum likelihood method to revise the information of pedigree and matching situation of SNPs.

ii) Genotyping quality control is basically some primary arrange and analysis for some genotypes after the distinguish by BeadStudio and then it will be used for following analysis. In this way, the efficiency and powerfulness of GWAS is improved tremendously. First, we use Beadstudio to visualize the data of chips and then manually revise those incorrect SNPs locus and import them as text form. The main indexes include:

a. SNP call rate, which means the percentage of successful measurement of a specific SNP. It's normally required to be higher than 95%.

b. Minor allele frequency, MAF. It's normally bigger than 1%.

c. Examination of Hardy-Weinberg equilibrium. It can filter out unqualified SNPs. Deviation from Hardy-Weinberg equilibrium[15] can theoretically occur because of selection, mixture of genetically heterogeneous populations, cryptic relatedness, or genotyping errors[16] (either selective dropout of a given allele or misclassification of alleles).

d. Repeat examination of Samples to make sure the consistency of results and it should be higher than 99.5%. All of procedures can be done in PLINK, but can also be finished in other statistical packages.

In our research, researchers use PLINK retained over 639 thousand SNPs having greater than 90% genotyping call rate and greater than 5% minor allele frequency and all individuals having over 95% genotype call rate were retained.[14]

4. Population

As mentioned before, in GWAS, population stratification and multiple testing adjusting are the main reasons for causing the errors of analysis. One possible strategy is making use of investigation for association study which is based on pedigree.

Now, what is population stratification? Stratification alludes to that there exist some subgroups possessing different frequency alleles in the population. The linkage disequilibrium is effected by selection of nature, genetic drift and population stratification. Therefore, when we do the GWA studies, some unrelated genes can also perform to be associated with traits, which lead to the appearance of false positive association.

Now, many methods have been developed for correcting population stratification. EIGENSTRAT, which is one of the most frequently used package to correct for population heterogeneity, uses an approach based on principal component analysis.[17]

5. Association analysis

In GWAS, we normally use Logistic regression model to analyze qualitative traits. When it comes to the quantitative traits, we usually use linear regression models to do the association analysis. In Logistic regression model, genotype is dependent variant whereas the structure of population and phenotypes are independent variables. As for linear models, there are two of them. One is general linear model, and another is mixed linear model. A complex quantitative trait is usually influenced by multiple factors, so a mixed linear model can hence add fixed effect and random effect to it. Therefore, the mixed linear model is more popular for quantitative trait analysis.

In our research, GWAS were run using 639,663 SNPs in a case/control analysis approach on EMMAX software which corrects for population stratification and relatedness. Significance levels were generated using basic (adaptive) permutation testing in PLINK. The most highly associated SNPs had a permutated *p*-value less than 1 ×10−6 at 1,000,000 permutations.[14]

6. Results and Manhattan plot

Manhattan plot is a scatter plot and is always used to show data with a large number scale project, especially popular for GWAS. More specifically, a GWAS Manhattan plot, genomic coordinates is X-axis, with negative logarithm of the association P-valie for each SNP on Y-axis. Thus, a dot on the plot means a SNP. Here, in our research paper, the x-axis denotes chromosome 1 through 29 with SNP positions plotted in increasing genomic order. The y-axis plots the −log10 *p*-value as determined in an association analysis using the program EMMAX.[14]

Figure 2