Understanding the basic statistics behind Genome Wide Association Studies (GWAS)

Daniela Baracaldo-Santamaría

Introduction

Genome wide association studies (GWAS) are a method of identifying genetic variants statistically associated with a biological trait of interest. Humans share more than 99% of the genome, however 0.1% of the genome will be different among humans as a consequence of genetic variants. The most common genetic variants causing these differences are called single nucleotide polymorphisms (SNPs), defined as a single nucleotide change that is present in more than 1% of the population. The majority of SNPs are responsible for the observed biological difference among humans, but others may don't even affect the coding of a protein. However, in the minority of cases some SNPs may predispose individuals to develop certain illnesses (1). Therefore, the main goal of GWAS is identifying genetic variants (SNPs) that are more likely to be present in individuals who express a qualitative trait or a quantitative trait more strongly than would be anticipated from the population.

GWAS test hundreds of thousands of genes using microarray technology in controls and in patients with the disease or the trait of interest. With this method an individual can be genotyped for between 500,000 and 4 million SNPs (2). The regions of the genome that are being studied are those that have the most variation and can provide enough data to infer the nucleotides found in other parts of the genome. This is possible due to a feature of the genome referred to as linkage disequilibrium (LD). LD is defined as the non-random assortment of alleles at different loci (3,4). This means that some portions of DNA are more likely to be passed on together during the process of meiosis than would be expected by chance. GWAS studies often identify several SNPs which are in LD with each other, and this creates a difficulty in identifying which is the truly causal SNP and which are the SNPs that are in LD with the causal one. Nonetheless, GWAS report blocks of correlated SNPs that show statistical association with the trait of interest, and therefore they are extremely useful providing hints into disease biology (5).

p-values and multiple testing

From a statistical point of view GWAS are hypothesis studies, in which the null hypothesis is that the marker (SNP) has no effect on the trait (disease), while the alternative hypothesis is that the marker does affect the trait because it is in LD with a quantitative trait locus (meaning they are associated) (6).

The usual p-value for determining statistical significance in epidemiological studies is 0.05, however this applies when performing single comparisons. In GWAS we are testing millions of SNPs simultaneously making the possibility for false-positives very high. For these types of studies, the p value can be adjusted using some methods such as the Bonferroni correction. For this method, the correction for multiple, *K*, comparisons is:

$$\alpha_{\text{corrected}} = \frac{\alpha}{K} \text{ or } \frac{0.05}{\text{\# of independent common SNPs}}$$

However, the use of Bonferroni correction is widely criticized in these studies for not taking into account that the tests are not independent (the tests are correlated because of LD). Many modifications to the Bonferroni method have been proposed including methods that account for LD, false discovery rate, and false-positive probability (7). Essentially the current accepted genome wide significance threshold is $5x10^{-8}$ (2).

Testing for associations: linear and logistic regression

Commonly used tests of association between SNPs and phenotypic traits used in GWAS include linear and logistic regression models. Linear regression is used when analyzing continuous (quantitative) variables:

$$\hat{Y} = \alpha + \beta X + \varepsilon$$

Where $\hat{Y}$ is the score on the phenotype, X is the genetic variant or SNP and β is the effect of the SNP on that outcome. This would tell us if a genetic variant is associated with a continuous trait (phenotypic score). The genotype of the participants is plotted in the x axis depending on whether the patient is homozygote for the minor allele (TT or 2), heterozygote (TG or 1) and homozygote for the major allele (GG or 0).

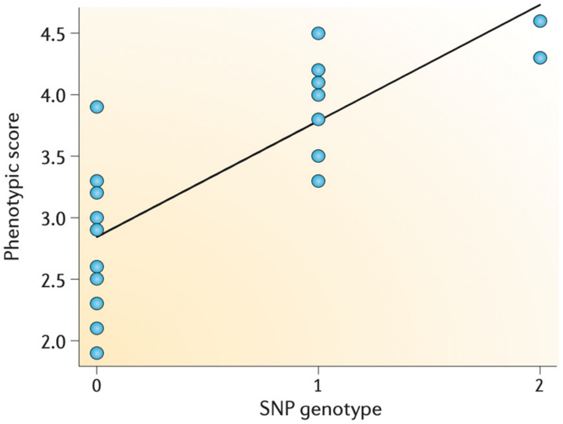

Figure 1. In this case there is a positive correlation between the genetic trait and the phenotype score. Image taken from: (8)

Now, we can calculate the coefficient of determination (R^2) in each model, which tells us the proportion of the variance in the dependent variable (phenotype) which can be explained by the independent variable (genotype). Therefore, the higher the R^2 (meaning the closer to 1) the stronger the association. R^2 can be calculated as follows in the case of linear regression:

$$R^2 = 1 - \frac{SS\ error}{SS\ total} \text{ or, } 1 - \frac{\sum_{i=1}^{n}(\hat{Y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Where SS error is the sum of squared distances from the fit ($\hat{Y}$), and SS total is the sum of SS distance from the mean ($\bar{y}$). Now, when performing a regression analysis, you can obtain another value termed the F-value (also called the F-statistic). The F value is used to check whether the variances of two populations are significantly different and can be expressed as:

F = Mean sum of squares regression / Mean sum of squares error,

= (SSR/DFssr) / (SSE/DFsse)

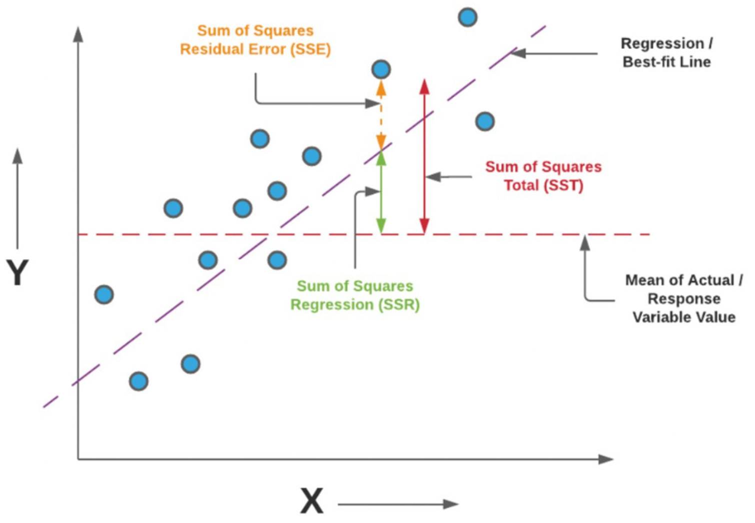

Figure 2. Explanation of SSE, SST and SSR in a linear regression. Image taken from: (9)

Where SSE is the sum of squares of the error, SSR is the sum squares of the regression, DFssr and DFsse are the degrees of freedom for the regression and for error respectively. The p-value can then be found from the F value by comparing the F-value calculated from your regression to a probability distribution of F-values assuming there is no relationship between the genotype and the phenotype (assuming the H_0 is true). The p-value is therefore the probability that if the null hypothesis is true, we observe a value as extreme as the F we observed or even more extreme.

On the other hand, when analyzing the association of genetic variants with dichotomous traits (absence or presence of disease, or cases and controls) a logistics regression model is more appropriate. This will tell us if the genetic variant increases or decreases the likelihood that an individual would possess the trait. In this way the logistic regression is:

$$\text{Ln}\ (P/1\text{-}P) = \alpha + \beta X + \varepsilon$$

Where β is the log odds for cases compared to controls, and the $e^{(\beta)}$ gives us the odds ratio.

Calculating R^2 in the case of logistic regression can be done by different methods, one of them is by calculating the McFadden’s pseudo-R squared, expressed as:

$$R^2_{\text{McFadden}} = 1 - \frac{log(L_c)}{log(L_{\text{null}})}$$

The p-value in the case of a logistic regression will be found by calculating a likelihood ratio Chi -squared test (LR chi2) (10) which is then compared to a probability distribution of chi squared values assuming no relationship between the genotype and the phenotype (assuming the H_0 is true). Then the p-value would be the probability that if the null hypothesis is true, we observe a value as extreme as the chi-square value we observed or even more extreme.

In GWAS the p-value is calculated for every SNP analyzed, which is then visualized in a Manhattan plot:

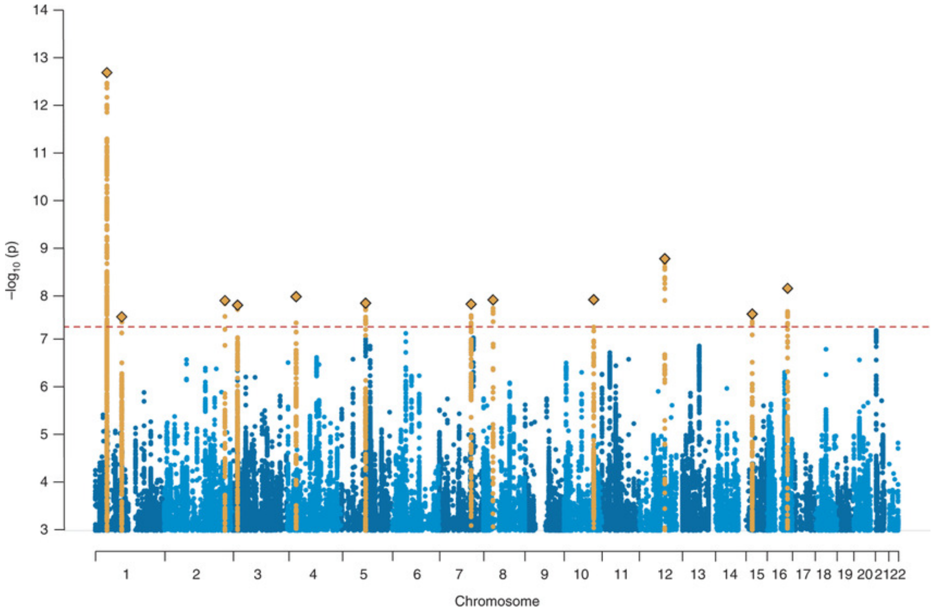

Figure 3. Manhattan plot. Figure taken from: (11)

In the Manhattan plot across the X axis, we can visualize the position of each SNP within the chromosome and the p-values for each SNP in the Y axis after being multiplied by $-\log_{10}$. The red dotted line represents the genome wide significance which as mentioned previously is generally considered to be $5x10^{-8}$.

Example

We are going to interpret the results of a GWAS used to determine genetic variants associated with obesity. Cotsapas et al carried out a GWAS of 775 bariatric surgery patients with a mean BMI of 50.6 compared with 3197 controls. They analyzed 655 130 SNPs and after quality control data an association study was performed with 457 251 SNPs. The authors then used logistic regression to calculate association between the SNPs and extreme obesity and then performed a chi-squared test to obtain the p-value, obtaining the following results:

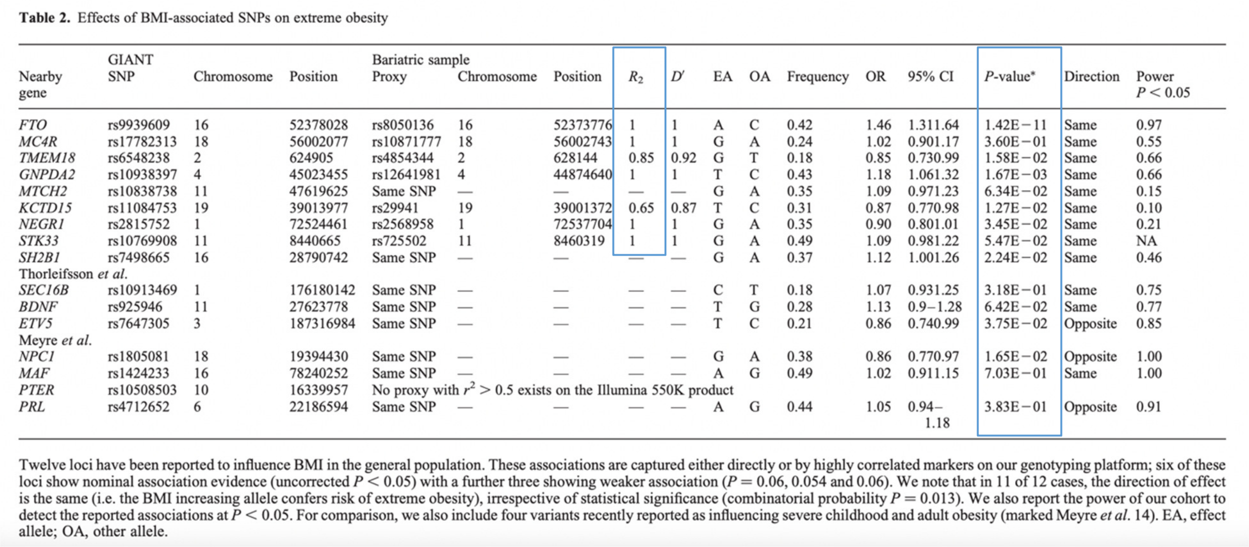

Table 2. Effects of BMI-associated SNPs on extreme obesity

	GIANT			Bariatric sample												
Nearby gene	SNP	Chromosome	Position	Proxy	Chromosome	Position	R_2	D'	EA	OA	Frequency	OR	95% CI	*P*-value*	Direction	Power $P < 0.05$
FTO	rs9939609	16	52378028	rs8050136	16	52373776	1	1	A	C	0.42	1.46	1.311.64	1.42E−11	Same	0.97
MC4R	rs17782313	18	56002077	rs10871777	18	56002743	1	1	G	A	0.24	1.02	0.901.17	3.60E−01	Same	0.55
TMEM18	rs6548238	2	624905	rs4854344	2	628144	0.85	0.92	G	T	0.18	0.85	0.730.99	1.58E−02	Same	0.66
GNPDA2	rs10938397	4	45023455	rs12641981	4	44874640	1	1	T	C	0.43	1.18	1.061.32	1.67E−03	Same	0.66
MTCH2	rs10838738	11	47619625	Same SNP	—	—	—	—	G	A	0.35	1.09	0.971.23	6.34E−02	Same	0.15
KCTD15	rs11084753	19	39013977	rs29941	19	39001372	0.65	0.87	T	C	0.31	0.87	0.770.98	1.27E−02	Same	0.10
NEGR1	rs2815752	1	72524461	rs2568958	1	72537704	1	1	G	A	0.35	0.90	0.801.01	3.45E−02	Same	0.21
STK33	rs10769908	11	8440665	rs725502	11	8460319	1	1	G	A	0.49	1.09	0.981.22	5.47E−02	Same	NA
SH2B1	rs7498665	16	28790742	Same SNP	—	—	—	—	G	A	0.37	1.12	1.001.26	2.24E−02	Same	0.46
Thorleifsson *et al.*																
SEC16B	rs10913469	1	176180142	Same SNP	—	—	—	—	C	T	0.18	1.07	0.931.25	3.18E−01	Same	0.75
BDNF	rs925946	11	27623778	Same SNP	—	—	—	—	T	G	0.28	1.13	0.9–1.28	6.42E−02	Same	0.77
ETV5	rs7647305	3	187316984	Same SNP	—	—	—	—	T	C	0.21	0.86	0.740.99	3.75E−02	Opposite	0.85
Meyre *et al.*																
NPC1	rs1805081	18	19394430	Same SNP	—	—	—	—	G	A	0.38	0.86	0.770.97	1.65E−02	Opposite	1.00
MAF	rs1424233	16	78240252	Same SNP	—	—	—	—	A	G	0.49	1.02	0.911.15	7.03E−01	Same	1.00
PTER	rs10508503	10	16339957	No proxy with $r^2 > 0.5$ exists on the Illumina 550K product												
PRL	rs4712652	6	22186594	Same SNP	—	—	—	—	A	G	0.44	1.05	0.94–1.18	3.83E−01	Opposite	0.91

Twelve loci have been reported to influence BMI in the general population. These associations are captured either directly or by highly correlated markers on our genotyping platform; six of these loci show nominal association evidence (uncorrected $P < 0.05$) with a further three showing weaker association ($P = 0.06$, 0.054 and 0.06). We note that in 11 of 12 cases, the direction of effect is the same (i.e. the BMI increasing allele confers risk of extreme obesity), irrespective of statistical significance (combinatorial probability $P = 0.013$). We also report the power of our cohort to detect the reported associations at $P < 0.05$. For comparison, we also include four variants recently reported as influencing severe childhood and adult obesity (marked Meyre *et al.* 14). EA, effect allele; OA, other allele.

Based on the previously explained concepts, here we can see they report the R^2 obtained from the logistic regression. The R^2 of some SNPs is 1, this means that there is a strong association between the SNP and obesity, however when we see the p-values for some of them we can conclude not all reach the threshold of genome-wide statistical significance. Indeed, the authors conclude that the variants near the FTO gene are associated with increased risk of extreme obesity, but that 12 others show nominal associations but don't reach the threshold.

Useful resources

1. Introduction to GWAS statistics: https://www.youtube.com/watch?v=Hjv_otXAkh0
2. Introduction to GWAS statistics part 2: https://www.youtube.com/watch?v=g1fQCC92WO0&t=511s
3. Linkage disequilibrium: https://www.youtube.com/watch?v=_xbpGvQHQAA&t=11s
4. F value: https://www.statisticshowto.com/probability-and-statistics/f-statistic-value-test/
5. Interpreting F statistics in linear regression: https://vitalflux.com/interpreting-f-statistics-in-linear-regression-formula-examples/
6. Linear and logistics regression in GWAS: https://www.youtube.com/watch?v=Du2MNYZGCiA&t=188s
7. McFadden's pseudo-R squared: https://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/
8. Statistical methods for population association studies: https://www.nature.com/articles/nrg1916

References

1. Clifford RJ, Edmonson MN, Nguyen C, Scherpbier T, Hu Y, Buetow KH. Bioinformatics Tools for Single Nucleotide Polymorphism Discovery and Analysis. Ann N Y Acad Sci [Internet]. 2004 May 1 [cited 2022 Nov 24];1020(1):101–9. Available from: https://onlinelibrary.wiley.com/doi/full/10.1196/annals.1310.011

2. Attia J. Genetic association and GWAS studies: Principles and applications - UpToDate [Internet]. 2022 [cited 2022 Nov 24]. Available from: https://www.uptodate.com/contents/genetic-association-and-gwas-studies-principles-and-applications?search=genome%20wide%20association%20studies&source=search_result&selectedTitle=1~150&usage_type=default&display_rank=1#H3898615991
3. Alvarenga AB, Rovadoscki GA, Petrini J, Coutinho LL, Morota G, Spangler ML, et al. Linkage disequilibrium in Brazilian Santa Inês breed, Ovis aries. Scientific Reports 2018 8:1 [Internet]. 2018 Jun 11 [cited 2022 Nov 24];8(1):1–11. Available from: https://www.nature.com/articles/s41598-018-27259-7
4. Qanbari S. On the Extent of Linkage Disequilibrium in the Genome of Farm Animals. Front Genet. 2020 Jan 17;10:1304.
5. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers 2021 1:1 [Internet]. 2021 Aug 26 [cited 2022 Nov 24];1(1):1–21. Available from: https://www.nature.com/articles/s43586-021-00056-9
6. Hayes B. Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). Methods Mol Biol [Internet]. 2013 [cited 2022 Nov 24];1019:149–69. Available from: https://pubmed.ncbi.nlm.nih.gov/23756890/
7. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nature Reviews Genetics 2014 15:5 [Internet]. 2014 Apr 17 [cited 2022 Nov 24];15(5):335–46. Available from: https://www.nature.com/articles/nrg3706
8. Balding DJ. A tutorial on statistical methods for population association studies. Nature Reviews Genetics 2006 7:10 [Internet]. 2006 Oct 22 [cited 2022 Nov 24];7(10):781–91. Available from: https://www.nature.com/articles/nrg1916
9. Kumar A. Interpreting f-statistics in linear regression: Formula, Examples - Data Analytics [Internet]. [cited 2022 Nov 24]. Available from: https://vitalflux.com/interpreting-f-statistics-in-linear-regression-formula-examples/
10. Williams R. Logistic Regression, Part III: Hypothesis Testing, Comparisons to OLS. [cited 2022 Nov 24]; Available from: https://www3.nd.edu/~rwilliam/
11. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nature Genetics 2018 51:1 [Internet]. 2018 Nov 26 [cited 2022 Nov 24];51(1):63–75. Available from: https://www.nature.com/articles/s41588-018-0269-7