# GENOME-WIDE ASSOCIATION STUDY AS A TOOL IN COMMON DISEASE RESEARCH

Joanna Zyla
Faculty of Automatic Control, Electronic and Computer Science, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

**Abstract**

The aim of this paper is to present main aspects related with the Genome-Wide Association Study commonly known as GWAS, which over last ten year become powerful tool in genomic research. That type of study is based on genome analysis for a large set of data (e.g. genotypes of single nucleotide polymorphisms), where the main goal is to find correlation/interaction between genomic differences and common diseases (like Alzheimer) or health issues (like response to radiation). To obtain reliable results the mathematical analysis has to be perform and its basic rules are include at this work.

## 1. Introduction

Nowadays, after huge progress in computer science and genomic knowledge, in one experiment we are able to asses hundreds of thousands potential effectors to commonly known diseases. Genome-Wide Association Study that analyzes DNA sequence variations from across the human genome in an effort to identify genetic risk factors for diseases that are common in the population. The final aim of GWAS is to use genetic risk factors to make predictions about who is at risk. Such approach not only allow for better understanding the disease and its prevention but GWAS make and important role in developing new pharmacologic therapies. Additionally, today Genome-Wide Association Study is not only used for investigation the risk factor in common diseases but it start to explain many social issues e.g. alcoholism, drug addiction or sensitivity to the sunlight.

However, first GWAS occur in 2005 at *Science* magazine, when the Klein et al. published work about Age-related Macular Degeneration (AMD) and proved that occurrence homozygote "CC" at polymorphism rs380390 increase the risk of AMD (OR=4.7)[1]. After 2005 the number of GWAS grew up every year and become more significant to biological knowledge. In January 2008, NIH (National Institute of Health) decided to create the catalogue of GWAS studies to organize all research and increase their usefulness in public health [2]. In their database they weekly screen PubMed to find and catalogue new Genome-Wide Studies. Now they label as a GWAS 1695 publications, which include 11342 interesting polymorphisms. Fig.1 present the number of published works from 2005 to 2012.

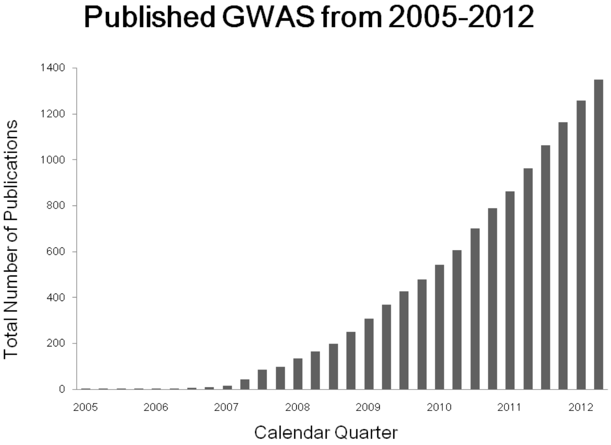

**Fig.1. View to the number of published GWAS from 2005 to 2012 [3].**

The research is approved as a GWAS when it meets several criteria. First of all it have to investigate more than 100,000 polymorphisms (SNPs) at first stage of study. Second statistical significance for SNP is $p<1.0*10^{-5}$ in initial and replication stage (If a study does not include a replication stage, significant SNPs from the discovery stage will be reported). Third, research has to be published in English and available in PubMed resources [3].

As every bioinformatics research the GWAS need the specific approach in analysis. The goal of this paper is to introduce and review GWAS technology, study design and analytical strategies.

## 2. GWAS - General information

### 2.1 Single Nucleotide Polymorphisms

The basic part of genome, which are use in Genome-Wide Association Study, are called single nucleotide polymorphisms (SNP). SNPs are point base-pair changes in the DNA between members of a biological species or paired chromosomes in a human with minor allele frequency>1% (What distinguishes it from mutations). Each individual has many single nucleotide polymorphisms that together create a unique DNA pattern for that person. SNPs can occur in every region, within coding sequences of genes, non-coding regions of genes, or in the intragenic regions between genes. All of those positions can have functional consequences from causing amino acid changes (nonsynonymous SNP), changes to mRNA transcript stability, and changes to transcription factor binding affinity, which make SNPs by far the most abundant form of genetic variation in the human genome [4-5].

### 2.2 Common disease, common variant hypothesis

The main hypothesis in GWAS study says that common diseases are likely caused by genetic variation that is also common in the population. By common it is understand that >1-5% must have the genetic variant. This statement led to important conclusions, which have to be considering in GWAS. First says that if SNP has only a small effect it explains only a small proportion of all genetic variance in disease but if SNP is heritable then common alleles multiply the risk of common disease. Second consequences say that by definition common alleles cannot have high effect to disease [6]. Fig.2 present graphical interpretation of above statements.

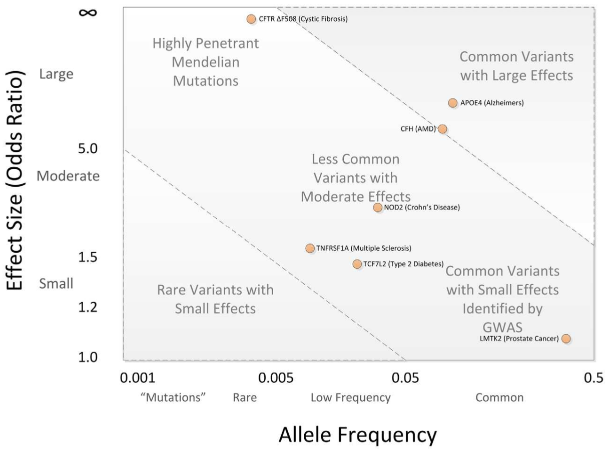

**Fig.2. View to the number of published GWAS from 2005 to 2012 [6].**

All this information has to be considered in construction of study group which size have to be large to get statistical significances.

### 2.3. Study Designs in GWAS

In GWA Study, three types of population under investigations are popular. For the most common belong case-control study, where two groups are required, first with recognitions of interest disease or health issue (case) and control, population where investigated problem do not occur. Very similar to case-control is cohort study, where population under investigation at the beginning of research is uniform and then under some stimulation (e.g. radiation) the disease outcomes are investigated against normal. Fig. 3 present difference between above mention studies.

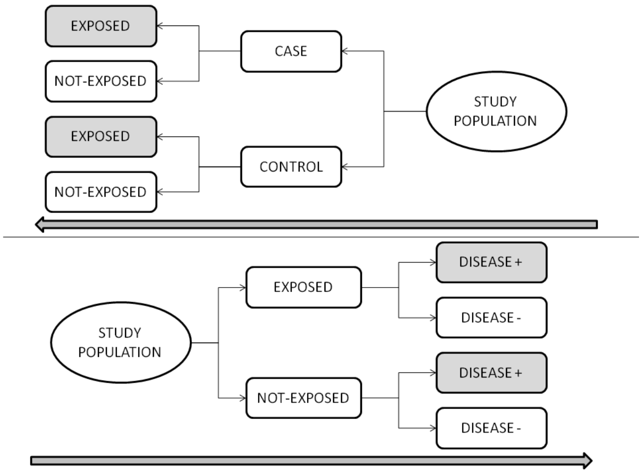

**Fig.3. Interpretation of difference between case-control and cohort study.**

Third type of study is cold Trio or case + parents, where biological material from investigated group and parents is genotyped. Such kind of study allows to assessment the frequency, which an allele is transmitted to an affected offspring from heterozygous parents. in assumptions disease-related alleles are transmitted in excess of 50%. Disadvantage of this study is highly sensitivity to genotyping errors.

In constriction of high quality GWA Study there is need to remember about ethnical differences, sex, age etc. Under those issue GWA Study are mostly performed on one ethnical group, with equal proportion of male and female and uniform age distribution. Also to obtain high statistical power large population are used, where the biggest research for now were performed on 100,000 persons [7].

## 3. Mathematical aspects

### 3.1 Hardy-Weinberg principle

Before starting every analysis in GWAS Hardy-Weinberg equilibrium (HWE) should be test. The principle states that *allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences* [8]. If we assume that allele "A" frequency is equal to p, and allele "a" frequency is equal to q, then according to HWE, genotype frequencies look as follow: $AA=p^2$, $Aa=2pq$ ans $aa=q^2$ (Fig.4). To assess the HWE the $X^2$ test, Fisher exact test or G test could be used. When the ratios of homozygous and heterozygous genotypes significantly differ from Hardy-Weinberg Equilibrium assumptions, it can indicate genotyping errors, batch effects or population stratification. In GWAS analysis SNPs, which do not follow Hardy-Weinberg equilibrium should be excluded from analysis.

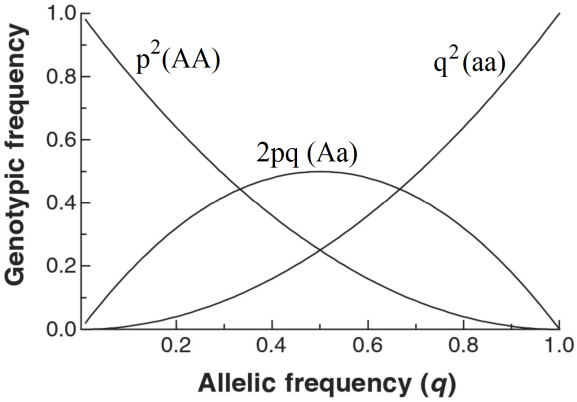

**Fig.4. Interpretation of Hardy-Weinberg Equilibrium [9].**

## 3.2 Linkage Disequilibrium (LD)

During the GWA Study, two significant SNPs findings are possible. First is when genotyped SNP is statistically significantly to phenotype of investigated disease - direct assassination or commonly named functional SNP. Second option is when the influential SNP is not directly genotyped, but instead it lay in high LD region with the influential SNP that is typed and statistically associated to the phenotype - indirect association (Fig.5).

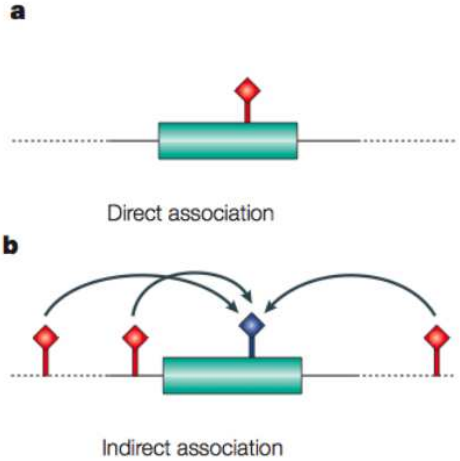

**Fig.5. Presentation of direct and indirect association [10].**

The indirect association is observed under Linkage Disequilibrium condition. By definition LD is situation where alleles laying nearby each other are non-independence because of a lack of recombination, which make their association non-random. In GWA Study such situation helps to find new significant SNPs even if they are not genotyped. To assess the linkage disequilibrium, two measures are used. First define as a capital D', describe recombination events between investigated SNPs. D' have range from -1 to 1 and if it is equal to 0 then we observe linkage equilibrium and cannot conclude about significance of discovered SNP. Second measure is defined as $r^2$ that is statistical measure of correlation of investigated locus. The range of $r^2$ is from 0 to 1 and the zero value states that locus is under linkage equilibrium [11].

## 3.3 Methods of analysis

Different study designs require various methods of analysis. In this paragraph method for case-control study will be present.

There are two most popular methods, first use the contingency table and second is based on logistic regression. To more conservative method belongs the contingency table that in example constructed as follow (tab.1):

**Tab.1. Contingency table to case control study, where a,b,c and d are counts of alleles.**

| | Allel A | Allel B |
|---|---|---|
| CASE | a | b |
| CONTROL | c | d |

To assess the significance of investigated SNP, chi-square test can be performed. Additional in GWA Study the odds ratio (OR) is measured as a effect size. The OR is a probability of occurrence the investigated event divided by not occurrence. Which in this case could be present like in formula 1.

$$OR = \frac{ad}{bc} \tag{1}$$

The value of OR=1 indicate no association to the diseases, OR > 1 show increase the risk of allele A, OR<1 increase the risk to allele B. However, the OR confidential interval (CI) cannot include 1.

Other method to analysis the case-control study is logistic regression, which use the binary outcomes. Let assume that $Y_i$ have to values, 0 for control and 1 for case. Then $X_i$ has 0 for AA allele, 1 for AB allele, and 2 for BB allele. In this conditions $\text{logit}(p_i)=\log_e[p_i/(1-p_i)]\sim\beta_0+\beta_1X_i$. The value of $\beta_1$ will estimate the significance of association with disease, which could be test e.g. by Wald Test ($H_0$: $\beta_1=0$). Also the CI of $\beta_1$ and $\beta_1$ alone are base to get the odds ratio that is estimate as follow (formula 2):

$$OR = e^{\beta_1} \tag{2}$$

Advantage of logistic regression is possibility of adding confounders like age, ethnic etc., which are define as other conditions in probability of logit function.

## 3. Summary

Above presented methods end problems are only a small part of what can be done in Genome-Wide Association Study. However in each analysis they should be consider and are base to more sophisticated solutions.

## Acknowledgement

The work was partially financially supported by SUT grant BK219/Rau1/2013/t.3 and DoktoRis-Scholarship program for Innovative Silesia.

## References

[1] Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J.: *Complement factor H polymorphism in age-related macular degeneration*. Science, 2005 308(5720):385-389

[2] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA. *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits*. *Proc* Natl Acad Sci USA. 2009, 106(23):9362-9367

[3] http://www.genome.gov/gwastudies/

[4] Nachman MW: *Single nucleotide polymorphisms and recombination rate in humans*. Trends Genet. 2001, 17(9):481-485

[5] Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al: *Integrating common and rare genetic variation in diverse human populations*. Nature, 2010, 467: 52–58.

[6] Bush WS, Moore JH: Chapter 11: *Genome-Wide Association Studies*. PLoS Comput Biol. 2012, 8(12) e1002822

[7] Hoffman TJ, Kvale MN, Hesselson SE at el. *Next generation genome-wide association tool: Design and coverage of a high-throughput European-optimized SNP array*. 2011, Genomics 98(2):79-89

[8] Hardy GH: *Mendelian Proportions in a Mixed Population*, 1908, Science, 28(7006):49-50

[9] http://4.bp.blogspot.com/-X0FjteNqWU4/TdGCZRSnuSI/AAAAAAAAAJU/_7Af2zg1NGA/s1600/0830060404006.png

[10] Hirschhorn JN, Daly MJ: *Genome-wide association studies for common diseases and omplex traits*. Nat Rev Genet, 2005, 6:95–108

[11] Devlin B, Risch N: *A comparison of linkage disequilibrium measures for fine-scale mapping*. 1995, Genomics 29: 311–322.

## Information about author

MSc. Joanna Zyla
Institute of Automatic Control, Silesian University of Technology
ul. Akademicka 16 p.535
44-100 Gliwice
tel. (032) 237 11 66
joanna.zyla@polsl.pl