BIOLOGIA (LAHORE)
December, 2019, 65 (II), 277-283

PKISSN 0006 – 3096 (Print)
ISSN 2313 – 206X (On-Line)

# Alterations, complexity and efficient strategies suitable for genetic studies

**[1]*AFIRA WAQAR, [2]ADNAN RAUF, [3]BUSHRA CHAUDHRY &[4]IKRAM-UL HAQ**

*[1]Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Karachi*
*[2]Liaquat National Hospital and Medical College, Karachi*
*[3]Aga Khan University, Karachi*
*[4] Institute of Industrial Biotechnology, Government College University, Lahore*

**ARTICLE INFORMAION**

Received: 02-08-2019
Received in revised form: 25-09-2019
Accepted: 19-12-2019

***Corresponding Author:**

AfiraWaqar:
afira.waqar@szabist.edu.pk

**Review Article**

**ABSTRACT**

Structural design of genetic studies has progressed over the last decade into the most powerful tool for probing the complex human diseases. There are many different approaches for the identification of genetic risk factors of the complex diseases. New pharmacological therapies may get established by understanding the underpinning of biological genetic effects. We focus here on the applications and strategies that are currently best suitable for genetic studies which provide preliminary genetic information for subsequent statistical analyses. Here we review the basic concept of genetic alterations including monogenic and polygenic characteristics of complex diseases to determine underlying genetic complexity and the most appropriate methods for genetic analysis.

**Keywords:** SNP, GWAS, Candidate gene, Polygenic

## INTRODUCTION

Alterations in the form of epigenetic changes are inheritable either by mitosis or meiosis. These epigenetic changes are drawn in the regulation of gene expression without the involvement of alterations at the sequence level of gene (Berger *et al.,* 2009). They are mostly involved in the basic processes of embryogenesis (Li, 2002) or by cell cycle to increase the expression of gene (Zaehres & Scholar, 2007). In addition, alterations are accountable for biological diversity amongst individuals and indicate acquitted phenotype discrepancies though, with the same genotypes. Epigenetic alterations include alterations in histone, expression of non-coding small RNAs and CpG (cytosine methylation of DNA at the guanine-phosphate-cytosine dinucleotide sites) are the most important and frequent (Fraga *et al.,* 2005).

The advancement of cellular repair mechanism stimulates the signal for prevention of DNA deterioration and activates the mechanism to repair the alterations. The frequency of the DNA alterations can be increased by the failure of the DNA repair mechanism neoplasm and can be considered at a large or small scale, depending on the size of the effective locus of DNA (Harper & Elledge, 2007).

### Single Nucleotide Polymorphism

A certain sequence variation, one of the diverse forms of DNA alterations that take place seldom in one percent of population, in a normal standard sequence is known as mutation though, most usual ones are recognized as polymorphisms (Fig. 1). Generally, the most studied mutation is single nucleotide polymorphisms (SNPs). However, more than 90 percent of the genomic alterations are conferred by SNPs (Maniatis, 2007). These 90 percent alterations are biallelic (wild or normal allele and lethal or rare allele) and can be surrogated either by transversion, purine to pyrimidine or vice versa or transition, purine to another purine or pyrimidine to another pyrimidine (Mimi *et al.,* 2018). Hence, nearly two-thirds of SNPs are substituted by transition between pyrimidines or purines (Guo & Jamison, 2005).

SNPs are aligned according to their location in genome and might be functional or non-functional. Coding SNPs are located in the exons and may be synonymous or non-synonymous (Burton *et al.,* 2007). Synonymous SNPs are silent that do not change the coding sequence of amino acids but alter the sequence of DNA whereas non-synonymous SNPs are found in a coding region

**Author's Contribution: A. W**., Proposed, designed and wrote the manuscript, **A. R**., Compiled and reviewed the literature, **B. C**. **& I. H**., proofread the final manuscript

that change the amino acid sequence by varying the sequence of DNA. Myriad of SNPs are positioned in the non-coding regions, introns and these intronic SNPs have unidentified function so far but are, by and large, involved in the regulatory functions and are termed as rSNPs. These non-functional SNPs are located in silencers or promoters. Nevertheless, all types of SNPs can cause genetic alterations leading to disease (Carlson *et al.,* 2004). There are total six probable ways of genotyping, which have included candidate association based polymorphism testing, diagnostics and risk outline, gene discovery and mapping without assumptions or hypothesis, pharmacogenetics, prediction of environmental risks, testing for homogeneity, and epidemiological study designing to understand the molecular foundation of SNPs (Palmer & Cardon, 2005; Schork *et al.,* 2000).

## Copy number variations

Human genome has several kinds of variations that range from large chromosome anomaly that can be seen through microscope to single nucleotide variants. One of the recent discovery is copy number variants that are abundantly present, sub-microscopic in nature, and range from kilo bases to mega bases (Mb) in size. Copy number variants (CNVs) are the most occurring and complex but less renowned form of genetic alterations and are defined by either gain or loss of more than 1000 base pairs. These types of variations are observed among individuals by the unaccepted number of copies of small region of DNA. CNV is defined as a segment of DNA which is 1 kb or larger than 1kb, exists at variable number of copies as compared to reference genome. This can be simple including tandem duplication and can be complex by gaining or losing of homologous sequences at different loci in the genome (Redon *et al.,* 2006).There are some variations in the gene of 50bp or more than 50bp that are collectively called as Structural Variants (SVs). SVs consist of many forms such as insertions, deletions translocations, duplication, and inversions (Hall & Quinlan, 2012). CNVs are unstable SVs that can affect copy number of DNA and can include deletions or every form of duplications such as tandem du-plication, tri-plication, and other amplifications. A broad range of processes or mechanisms can produce SVs and they are responsible for the varied SVs distribution in terms of location and size all across the genome (Monlong *et al*, 2018).

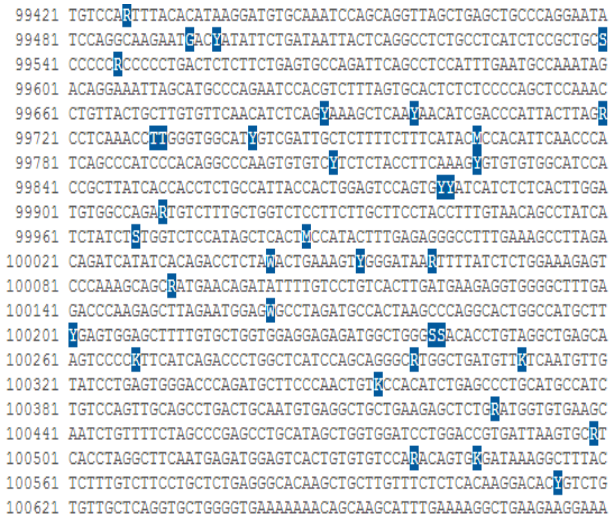


**Fig. 1:** Single nucleotide Polymorphism

Highlighted alphabets in a human TCF7L2 gene sequence are showing the ambiguity code for nucleotides considered as SNPs provided by International Union of Pure and Applied Chemistry (IUPAC) and indicated the possible nucleotides that can occur at a given position (Ensembl Release 98,2019). *Y is for pyrimidine either T or C, R is for purine either G or A, K is for keto (G or T), S is for strong bonds (G or C), W is for Weak bonds (A or T), M is for amino (A or C).

## Genetic complexity

All the characteristics and diseases of human with the heritable constituent can generally be categorized in two main sets, based on the genetic complexity.

## Monogenic

The Mendelian patterns of recessive or dominant inheritance are followed by monogenic diseases or characteristics. Typically, in humans, it has been recognized by various studies that diseases with Mendelian inheritance have multiple affected members within a family. Mode of inheritance can examine the trait or any disease that can be transmitted on throughout the progenies within family. To define a disease or a particular characteristic or trait with Mendelian inheritance, it must be illustrated to confirm either a recessive or dominant pattern within families. Recently information, about traits or diseases that show the existence of Mendelian inheritance is organized in the “Online Mendelian Inheritance of Man”, OMIM database. Genes that specify any ailment, disease, or affected phenotypes following inheritance of Mendelian, are also enlisted in OMIM. Cystic

fibrosis (CF) was identified as the first monogenic disease due to the presence of underlying gene. For patients, it will be an added advantage if a disease is noticed with Mendelian inheritance and referred as monogenic. Though, the involvement and interaction of multiple genes at same time will result in an extremity. This will further leads to complexity of disease along with the existence of same mutation in the candidate gene. Comparatively, a small number of diseases or traits have been considered as monogenic. As a result, an enormous amount of commonly known human traits, ailments or any disease with inbred capability are polygenic (Scriver & Waters, 1999).

Many benefits can be attained by diagnosing a monogenic disease that involves aimed and specific unconventional treatment depends on the susceptible gene and genome sequencing of family member testing for genetic identification that often leads to prior recognition and prevents misdiagnosis. Although, there is a continuous rise and increase in clarity of benefits with the diagnosis of monogenic disease but still many boundaries and challenges are encountered (Misra & Owen, 2018).

**Polygenic**

In the recent literature, most frequently referred complex genetic traits are the polygenic traits and are also referred as complex genetic traits. Opposing to monogenic trait, the name of this complex genetic trait shows that it is predisposed by two or more than two genes, and also does not exhibit the Mendelian pattern of disease transmission within pedigree. The greater part of extensive human diseases or human traits with the capability of inheritance has a complicated genetic make-up (Wang *et al.,* 2005).

The most frequently used method is to study monozygotic and dizygotic twins for resolving the heritable constituent of diseases or traits that does not explain Mendelian inheritance. In comparison between both twins, dizygotic twins share less phenotypes whereas the extent of sharing in monozygotic twins is higher. This method of estimation provides a good and primary assessment of the transmissible constituent by affecting the disease or trait being studied (Boomsma *et al.,* 2002). Polygenic traits are expressive among number of different traits for instance, a recent study concluded that foundations of polygenic risks are self-regulating however, they differ with phenotypic relationship (Weiner *et al.,* 2017). These conclusions of polygenic analyses lead to further contribution and perception of complex genetic traits with the increase in sample size, rise in clinical symptoms or environmental risk factors (Wray *et al.,* 2014).

**Genetic analyses**

There are generally two methods that have been used in genetic analyses to identify the hereditary machinery of disease or any trait. They are referred to as the linkage analysis and the association analysis. Both methods have the common contributed factors of utilizing recognized genetic markers including SNPs or microsatellites moreover, both share the common principle of finding one or many markers that are in correlation with the standard inherited gene loci which are associated with the specific trait of concern (Wang *et al.,* 2018).

**Linkage**

During meiosis, the action of recombination events occur in which short segments of DNA are exchanged between two chromosomes. The standard number of total recombination events during meiosis is approximately 38 and 24 for females and males, respectively (Cheung *et al.,* 2007). Recombination is a chief mechanism that can be utilized to map loci which is in linkage to a disease or a trait. The incidence and process of recombination events is directly related to the distance between loci. Loci can be inherited together if they are present adjacently to one another and are referred to be in mutual association.

Linkage analyses are used to determine the measurement of coinheritance by assessing how the already-established genetic marker is inherited together with the particular disease or trait within families. In linkage analysis, microsatellites as genetic markers have been typically used. Recombination among family members can also be followed by linkage analysis to find the causative loci. Nevertheless, this analysis is inadequate and restricted as it depends on the actual presence of the number of meiosis among families related to the family size or the number of generations represented. Due to this restriction and inadequacy, the poor genetic declaration of the linkage is being perceived between the marker and the interested region of gene, even they are at enormous distance apart. Studies used to discover linkage of genetic loci to a specific disease or trait following Mendelian inheritance pattern are extremely useful, done by

genetic linkage analysis (Jimenez-Sanchez *et al.,* 2001). On the other hand, detection of genetic machinery of common and familiar diseases by means of linkage analysis has not been considered as practically thriving (Hirschhorn & Daly, 2005).

**Association**

Association analysis is used for the detection of relationship among alleles and disease by relating frequencies of alleles between case subjects and coordinated healthy control. There are various reasons for the analysis based on association. First, the inspected allele might be connected to the specific disease, causing a direct association. Second, the allele might be in linkage disequilibrium with the responsible allele and displayan indirect association. The third cause for association is population structure in which the occurrence of the inspected allele varies in the influenced group compared to the control group. This is not because of the examined trait but because they originate from different population cohorts. It is the most powerful tool to inspect the genetic make-up of complex diseases (Burton *et al.,* 2007).

**Association analysis: Linkage disequilibrium**

Linkage disequilibrium (LD), usually used in association analysis and is the connection amongst genetic loci and genetic markers. There are two main categories of statistical measurements, D´ and $r^2$ that are explained by the LD. The estimation of the D´ measurement shows the occurrence of the amount of recombination events connecting the two loci and their significance ranges from 0 to 1. Alleles of the two loci can be inherited together and are believed as in complete genetic linkage if the value of D´ equals to one. However, the frequency of marker alleles varies among different populations due to the presence of different mutations at different instances in the genetic history of the populations. The statistical measurement, $r^2$ value is used to express the correlation between the two loci thus, the marker alleles consist of the equivalent frequencies if $r^2$ equals to one. This relationship is employed in choosing the useful SNPs referred to as tag SNPs with the intention of maximizing cost-benefits in an association study (Carlson *et al.,* 2004).

**Haplotypes**

A set of discrete genetic loci, affiliated to the same chromosome, are inherited together and are termed as haplotypes (Reynisdottir *et al.,* 2003)(Fig.2). Haplotypes are used in various genetic analyses and can be constructed by the measures of linkage disequilibrium of a specific segment of DNA and its population-based data are freely accessible in the Hap Map database (Frazer *et al.,* 2007).

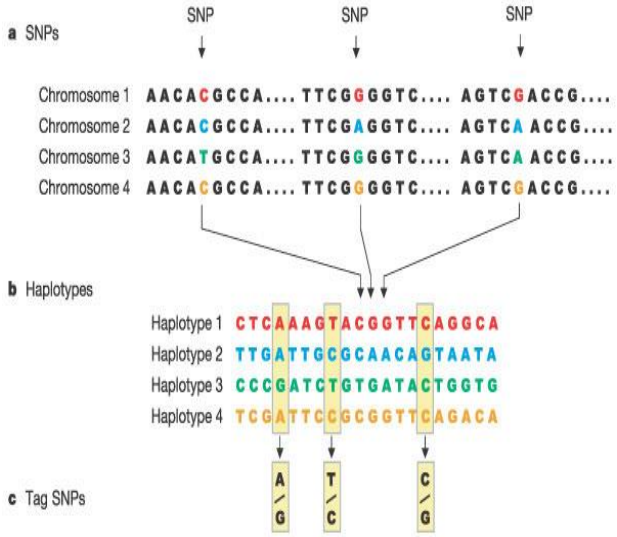


**Fig. 2:** Construction of the haplotypes

This figure illustrates the formation of haplotypes occurring in three steps. (a) Represents the SNPs recognized in DNA samples from various individuals. (b) Represents the close SNPs that are inherited together are assembled into haplotypes. (c) Represents the tag SNPs within haplotypes is identified that exclusively identify these haplotypes. By genotyping, the three tag SNPs can identify which of the four haplotypes are present in each individual (adopted from www.hapmap.org).

**Genome-wide association studies**

Genome-wide association studies (GWAS) have become progressively common to identify associations between single nucleotide polymorphism and phenotypic traits (Marees *et al*., 2018). Many complex human diseases with the contribution of many genetic and environmental determinants have increased in incidence during the past 2 decades. During the same time, extensive effort and expense have been consumed in conducting whole-genome screening of individuals. The major aim of these studies is recognition of genetic loci contributing to the vulnerability of complex human diseases. However, the achievement of genome wide association

studies (GWAS) based on inherited studies is the identification of commonly occurring genetic susceptible disease risk (Martin *et al.,* 2017).

GWAS have developed over the last ten years into an influential tool for the investigation of genetic architecture of complex human diseases. GWAS have rapidly become a standard technique for the discovery of disposed genes. A considerable number of recent GWAS show that for most disorders, only limited common variants are associated and the associated Single Nucleotide Polymorphisms (SNPs) clarify only a small division of the genetic hazard. A standard method for disease gene discovery by GWAS provided an effective unbiased approach to reveal the risk alleles for genetically complex disorders. The principle of the GWAS strategy is that widespread common variations in the human genome, revealed by SNPs with occurrences greater than 1%, is accountable for the possibility of most genetically complex disorders (Peter *et al.,* 2017).

There are three essential elements that are involved in conducting GWAS; a large sufficient study samples from populations, polymorphic alleles that can be reasonably and proficiently genotyped, and powerful statistical analytic methods that can be utilized to detect the genetic associations in an unbiased manner. A generous numbers of SNPs throughout the genome were identified by the HapMap Project and are sited in simple accessible databases including *Ensembl* at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBI-EBI) (Woicik *et al.,* 2019).

Large-scale sequencing was formerly excessively expensive but is now more financially practical today and is plausible to become a communal choice in the future. This will provide an added evidence regarding the genetic cause of complex disorders. Hardy & Singleton in 2009 elucidates several of the significant GWAS concepts. It concludes that if the genetics of complex diseases is placed in a position of jigsaw puzzle, GWAS put the boundaries and corners in position and now just need to have a framework to pursue the etiology of complex genetic disorders. Following a GWAS, association results can be allocated to routes with the suitable computational tools and *in silico* databases (Bustamante *et al*., 2011).

## CONCLUSION

In conclusion, although these genetic approaches have limitations and analytical challenges, they also provide ways to prioritize genes for bioinformatics and laboratory studies focused on detecting underlying variants and their biological roles. In addition, post-GWAS studies can probably provide a clearer picture of common SNPs in complex disorders. There have been many improvements in the field of complex human traits or diseases, although sufficient challenges remain. Improving awareness of polygenic or monogenic diseases and convenience of economical high value services of genetic testing remain a key challenge. Much remains to be done to facilitate diagnostic pathways that are clinically safe, at a national level. We are at the cusp of scientific revolution by applying genetic strategies to multiple different population stratification and can easily reinforce these efforts to be authentically effective.

## REFERENCES

Berger, S. L., Kouzarides, T., Shiekhattar, R., Shilatifard, A., 2009.An operational definition of epigenetics. *Genes Dev.,*23(7):781-3.

Boomsma, D. Busjahn, A. and Peltonen, L.,2002. Classical twin studies and beyond. *Nat Rev Genet.,* 3(11):872-82.

Burton , P. R., Clayton, D.G. and Cardon, L. R. , 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447:661–78.

Burton, P. R., Clayton, D. G. and Cardon, L.R., 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–78.

Bustamante, C.D., Francisco. M. and Burchard, E. G., 2011.Genomics for the world. *Nature*; 475(7355): 163–165.

Carlson C. S., Eberle, M. A., Rieder, M. J,, Yi, Q, Kruglyak, L. and Nickerson, D. A., 2004. Selecting a maximally informative set of single nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet*, 74(1): 106-120.

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q,, Kruglyak, L. and Nickerson, D, A., 2004.Selecting a maximally informative set

of single nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet.,*74(1): 106-20.

Cheung, V.G., Burdick, J. T., Hirschmann, D. and Morley, M., 2007. Polymorphic variation in human meiotic recombination.*Am J Hum Genet.,*80(3): 526-30.

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ling C, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu YZ, Plass C and Esteller M., 2005. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci*.,102(30):10604-9.

Frazer, K. A., Ballinger, D.G., Cox, D. R., Hinds, D. A., Stuve, L. and Gibbs, R.A., 2007. A second generation human haplotype map of over 3.1million SNPs. *Nature,*449(7164): 851-861.

Guo, Y., and Jamison, D. C., 2005.The distribution of SNPs in human gene regulatory regions. *BMC Genomics.* 6:140.

Hall, I. M. andQuinlan, A. R., 2012.Detection and interpretation of genomic structural variation in mammals. *Methods Mol Biol.,* 838:225-48.

Harper, J. W.andElledge, S. J., 2007. The DNA damage response: ten years after. *Mol Cell.* 28(5):739-45.

Hirschhorn, J. N. andDaly, M.J., 2005.Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*., 6(2): 95-108.

Jimenez-Sanchez, G., Childs, B. and Valle, D., 2001.Human disease genes.*Nature,*409: 853-5.

Li, E., 2002. Chromatin modification and epigenetic reprogramming in mammalian develop-ment. *Nat Rev Genet.,*3(9):662-73.

Maniatis, N., 2007.Linkage disequilibrium maps and disease-association mapping. *Methods Mol Biol.*, 376:109-21.

Marees, A., T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C. and Derks, E.,M., 2018. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res.*, 27(2):1-10.

Martin, A. R., Gignoux C. R., Walters R. K., Wojcik G. L., Neale, B. M.andGravel, S., 2017.Human demographic history impacts genetic risk prediction across diverse populations. *Am J Hu Genet*., 100(4):635-649.

Mimi, X., Yao,M., Feng, S., Jian, J., Zhang, Y. and Peng, Z., 2018. Restriction site-associated DNA sequencing for SNP discovery and high-density genetic map construction in southern catfish (*Silurusmeridionalis*).*R Soc Open Sci.,* 5(5): 172054.

Misra, S. and Owen, K. R.,2018. Genetics of Monogenic Diabetes: Present Clinical Challenges. *CurrDiab Rep.,*18(12): 141.

Monlong, J., Cossette, P., Meloche, C. Rouleau, G. and Girard, S.L., 2018. Human copy number variants are enriched in regions of low mappability. *Nucleic Acids Res.,* 46(14):7236-7249.

Ensembl Release 98,2019. Human Species, gene TCF7L2.(www.ensemle.org).

Palmer, L. J., Cardon, L. R., 2005.Shaking the tree: mapping complex disease genes with linkage disequilibrium.*Lancet*,366(9492):1223-34.

Peter, M. V.,Naomi, R. W., Zhang, Q., Sklar, P., Mark, I.M., Matthew, A.B. and Yang, J., 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet.*, 101: 5–22.

Redon, R, Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D. and Fiegler, H., 2006.Global variation in copy number in the human genome. *Nature*, 444 (7118):444-54.

Reynisdottir, I., Thorleifsson, G., Benediktsson, R., Sigurdsson, G. and Emilsson, V., 2003.Localization of a susceptibility gene for type 2 diabetesto chromosome 5q34-q35.2.*Am J Hum Genet*.,73(2): 323-335.

Schork, N. J., Swapan, K. N., Fallin, D. and Chakravarti, A., 2000. Linkage Disequilibrium Analysis of Biallelic DNA Markers, Human Quantitative Trait Loci, and Threshold-Defined Case and Control Subjects. *Am J Hum Genet.,*67 (5): 1208-1218.

Scriver, C. R.and Waters, P. J., 1999. Monogenic traits are not simple: lessons from phenyl ketonuria.*Trends Genet.* 15(7):267-72.

Wang, W.Y., Barratt, B. J., Clayton, D. G.and Todd, J. A., 2005.Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet.,*6(2): 109-18.

Wang, X.,Felix, Boekstegers, F. and Brinster, R., 2018.Methods and results from the genome

wide association group at GAW20. *BMC Genet.,* 19(1): 110-117.

Weiner, D. J.,Emilie, M. W., Stephan, R., Walters, R. K. and Elise, B. R., 2017. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat Genet.,*49(7): 978–985.

Woicik, G. L., Graff, M., Katherine, K. N.,, Tao, R., Haessler, J., Christopher, R. G., Heather, M. H. and Yesha, M. P., 2019. Genetic analyses of diverse populations improves discovery for complex traits. *Nature,*570(7762): 514–518.

Wray, N. R. Lee, S. H., Mehta, D., Vinkhuyzen, A.A., Dudbridge, F. and Middeldorp, C.M., 2014. Research review: Polygenic methods and their application to psychiatric traits. *J Child Psychol Psychiatry*., 55(10):1068-87.

Zaehres, H.and Scholer, H. R., 2007. Induction of pluripotency: from mouse to human. *Cell*, 131(5):834-5.