universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„An evolutionary perspective on three psychiatric disorders: major depressive disorder, bipolar disorder, and attention-deficit/hyperactivity disorder"

verfasst von / submitted by

Ana Jelicic

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2019 / Vienna 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 013
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Joint Degree Programme MEi :CogSci Cognitive Science UG2002
Betreut von / Supervisor:	Mag. Dr. Helmut Schaschl, Privatdoz

Acknowledgments

This master thesis would not have been possible without all the people who supported me throughout the whole process in different ways. I would like to thank:

my supervisor, Helmut Schaschl, for having the time and enthusiasm to share his knowledge, and for encouraging me to learn and explore new ideas.

Irena for stimulating discussions and being there through all the difficulties.

Kevin for understanding certain things better.

Ryan for proofreading and all the cheeseburgers.

Maike for her support and helping me not be lost in translation.

My family for supporting me whenever I need it.

Table of Contents

1. INTRODUCTION 6

1.1 Aims and hypotheses of this study 9
1.2 Interdisciplinarity 10
1.3 Evolutionary Theory 10
1.4 Population Genetics and Natural Selection 14
1.4.1 Selection and patterns of linkage disequilibrium (LD) 16
1.4.2 Genome-Wide Association Studies (GWAS) used in human genetic studies 19

2. THE GENETICS OF PSYCHIATRIC DISORDERS 23

2.1 Major Depressive Disorder 25
2.1.1 State of the art 26
2.2 Bipolar Disorder 27
2.2.1 State of the art 29
2.3 Attention-Deficit/Hyperactivity Disorder 30
2.3.1 State of the art 31
2.4 Comorbidity and Co-occurrence 32
2.4.1 Phenotypic overlap 32
2.4.2 Genetic overlap 33

3. MATERIALS AND METHODS 34

3.1 Data 34
3.1.1 Genetic Data 34
3.1.2 Genome-Wide Association Studies 34
3.2 Data Analysis 43
3.2.1 Statistics 43
3.2.2 Bioinformatics 48

4. RESULTS 50

4.1 Genes under Natural Selection 50
4.1.1 Results for MDD 50
4.1.2 Results for BPD 52
4.1.3 Results for ADHD 54
4.1.4 Overlap of risk genes between disorders 56
4.2 EHH plots 58
4.3 Biological pathways 60
4.3.1 Results for MDD 60
4.3.2 Results for BPD 61
4.3.3 Results for ADHD 62

5. DISCUSSION 65

5.1 Loci under positive selection 65
5.2 Biological pathways 67
5.3 Evolutionary perspective 67
5.4 Limitations 69
5.5 Conclusion 70
5.6 Ethical consideration 70

REFERENCES 72

APPENDIX 80

SUPPLEMENTARY MATERIAL 82

1. INTRODUCTION

The question of nature and nurture is one that will always bring heated discussions to the table, but making generalizations that it is one or the other given the abundance of research done in various fields such as biology, psychology, anthropology, should be regarded as dogmatic. Undeniably, there are certain phenotypic traits that are controlled by nature and that cannot be changed, like having two ears or a beating heart. However, when we talk about cognitive capabilities while denying the influence of the environment, or denying the genetic basis of it, we can only reach the wrong conclusions. Additionally, evolutionary research shows that the environment and nurturing from one generation will influence nature in some subsequent generations, so nature does not remain unchanged either. How much is one or the other responsible for a certain characteristic and to what extent they influence each other is also relevant for the development of new medical treatments [1]. Pharmacogenomics looks like a promising way for personalization of treatments (precision medicine), as drugs could be chosen based on the person's genotype, and not just phenotype [2, p564]. Although generalized treatments seem to be effective in the case of certain medical diseases, for the ones that affect our cognitive ability it seems that more personalized treatments could yield stronger results. Psychotherapy can also lead to epigenetic changes that occur through the learning process in the therapeutic relationship; however, genetic differences may also contribute to differences in susceptibility to psychotherapy [1].

The aim of this thesis is to investigate how natural selection influences the distribution of risk alleles of cognitive disorders in human populations by applying population genetics models of natural selection. By considering different populations it is possible to hypothesize how certain differences might be caused by different environments and the evolutionary forces behind them.

Novel research in genetics shows that the development and function of the human brain are directed by a genetic blueprint - the genome. A recent transcriptome analysis shows that 74% (n=14 518) of all human proteins (n=19 613) are expressed in the brain and many these genes show an elevated expression in brain compared to other tissue types [3]. While humans were migrating from Africa (the out-of-Africa expansion hypothesis) and settling in new and different geographical and biological environments across the globe, the genome was changing as a response to the new environments [4]. Moreover,

about 9 000 – 13 000 years ago humans were switching from hunter-gatherer societies to agricultural societies which brought about substantial changes in diet and lifestyle associated with plant and animal domestication. This period is known as the Neolithic revolution, and it started in the area of today's Syria, northern Mesopotamia, and part of southeast Turkey [2, p374-375]. Additionally, in the same period, human population densities increased very quickly. The new population densities, as well as the exposure of farmers to livestock pathogens, increased the spreading of novel infectious diseases [but see 5]. Every one of these changes imposed new selective pressures which resulted in alterations to the population's genotypes so that they would be better suited to the new environments. Some examples of this are changes in response to malaria infection, changes of lactase (LCT) gene in response to dairy farming, salt sensitivity variant in response to climate, and changes of genes that are involved in brain development [6]. Thus, different additional selection pressures are assumed to act on a persistent basis in humans, and other species, which can be the result of sexual competition, viability selection and resistance to evolving pathogens [7].

Phenotypic evolution in populations is the result of different evolutionary forces such as **mutation**, **genetic drift**, **gene flow**, and **natural selection** acting on heritable genetic variation. These forces are part of the so-called Modern Synthesis, which emerged in the first half of the 20th century and has been the dominant conceptual framework in evolutionary biology since then. However, the field of evolutionary biology has continued to evolve with new theoretical and empirical findings which led to broadening of the initial theory in what is now called the Extended Evolutionary Theory. The extended theory includes findings from evolutionary developmental biology, developmental plasticity, inclusive inheritance, and niche construction, and while it does not deny anything from the Modern Synthesis theory, it is not just its extension, but rather a "distinctively different framework for understanding evolution, which, alongside more traditional perspectives, can be put to service constructively within the field" [8]. Thus, phenotypic evolution can occur via plastic changes, genetic changes, or a combination of the two processes [9].

The human brain is an unusually large and complex organ and its function is shaped by these numerous evolutionary processes. Not only environmental, but potentially social and cultural conditions as well influence at different intensity the evolution of the human brain. Evolutionary hypotheses have been proposed to account for the observation that phenotypic traits with mildly deleterious effects on fitness such

as disorders have not been removed by natural selection [10]. Risk alleles of psychiatric disorders should be under strong negative (purifying) selection because of their negative effects on human life and reproduction. Psychiatric disorders carry a lifetime burden for the people experiencing them as they often appear in young adulthood (some also in childhood, like autism spectrum disorder and attention-deficit/hyperactivity disorder) and therapy may be only partially effective as a form of treatment. These disorders affect negatively different cognitive functionalities such as emotional regulation and perception, executive function, and motivation, and often appear in times that are important for the individual's life, e.g. during education, career development, or formation of personal relationships. Additionally, these disorders can lead to substance abuse and lower access to medical treatment for other illnesses. However, it appears that at least some psychiatric disorders are highly polygenic, suggesting that their genetic predisposition should be due to the additive result of hundreds to thousands of genetic variants with small effect sizes [10]. Moreover, these genetic risk variants may differ among human populations. Different authors have suggested that human adaptation in response to natural selection of polygenic phenotypes may be driven via subtle allele frequency changes at many loci [11, 12, 13]. Thereby, selectively beneficial traits may trade off with one another, with consequences that are mediated by the form and strength of natural selection and the genetic bases of phenotypic traits. Furthermore, the strength of natural selection may be dependent on various environmental factors driving population-specific frequencies of selected alleles (i.e., local adaption). Signatures of polygenic adaptation have been identified for some human traits including anthropometric traits, immune responses, and metabolic traits [12, 14, 15]. Moreover, the results of a recent study suggest that certain autism spectrum disorder (ASD) risk alleles were under positive directional selection during human evolution due to their involvement in neurogenesis and cognitive ability [16].

The World Health Organization’s 2016 Global Burden of Disease report [17] estimates that depressive disorders are currently the third ranking cause off years lost due to disability and that self-harm, which is often caused by depressive disorders, is the 18th most common cause of death.

1.1 Aims and hypotheses of this study

Despite intensive research, there are still major gaps in our understanding of the evolution of the human brain with its different structures and functions. In particular, we want to understand how the brain evolved, if and how evolutionary forces have driven human cognitive abilities, and how the existence of different disorders, which can cause substantial impairments in people's lives, fits into the evolutionary narrative. Part of the puzzle lies in the identification of all the relevant genetic loci and their functions involved in brain development, as well as understanding how environmental, social, and cultural factors can drive potential selective events at these loci.

Recently, it has been suggested that some genes under positive selection are linked to increased susceptibility to certain psychiatric disorders, and that these disorders emerged as humans developed higher cognitive abilities, such as language, intelligence, and creativity [1, 10, 18, 19, 96]. We can apply population genetic models of natural selection to identify genes that are under positive selection in different human populations. This can allow us to pinpoint functionally important genomic regions that have participated in the genetic adaptation of humans to their different environments. The genetic dissection of the intensity and type of selection acting on genomic regions can be used to predict involvement in different forms and severities of human diseases [20].

The aim of this study was to analyse three psychiatric disorders **Major Depressive Disorder**, **Bipolar Disorder**, and **Attention-Deficit/Hyperactivity Disorder** that have overlapping observable characteristics and test the hypothesis that genetic loci associated with these disorders are under some form of positive selection. These disorders have been shown to be polygenic and highly heritable, and it has also been demonstrated that they share a certain number of genes [21, 22, 23, 24].

Specifically, we hypothesized that genetic loci associated with these psychiatric disorders have been targets of **recent positive directional selection**. We integrated genomic data from different human populations, data from genome-wide association studies (GWAS), and biochemical pathway analysis in order to pin down the genetic targets of natural selection, and to understand the natural history of these disorders. We performed a genome-wide scan for recent positive selection using haplotype and single nucleotide polymorphism (SNP) based selection models. Furthermore, we tested the

hypothesis that the selected loci assemble into a specific network (pathways) of genes that have a correlated evolutionary history.

1.2 Interdisciplinarity

The interdisciplinarity of this thesis is reflected in the use of human population genetics, evolutionary theory, anthropology, and bioinformatics in order to try to tackle the question of development of the human brain and its cognitive abilities. Three disorders investigated here affect cognitive abilities in different ways, making it more or less difficult for an individual that experiences the disorder to live, even leading in some cases to the individual taking their own life. Uncovering the underlying mechanism causing these disorders is important for understanding them, developing better therapeutic method, and destigmatizing them.

1.3 Evolutionary Theory

Nothing in biology makes sense except in the light of evolution.
Theodosius Dobzhansky

Evolution is a process in which the genetic makeup of the individuals in a population changes over time. Depending on the time frame in which it is observed, it can be divided into microevolution, which happens over shorter intervals of time that include a couple of generations, and macroevolution, which happens over geological timescales (thousands to millions of years) and focuses on the evolution of species and higher levels (general, families, etc.) [2, p3]. The main driving forces behind evolution are natural selection, mutation, migration (gene flow) and genetic drift. The neutral theory of evolution was proposed by Motoo Kimura (1968, 1983) and holds that most nucleotide changes that become fixated in a population have a neutral effect on its fitness and that the most important evolutionary force on the level of DNA sequences is genetic drift.

Kimura also claimed that the rate of molecular evolution is equal to the mutation rate [25, p252]. The theory has been highly debated since its publication but in the end, a consensus was reached that the theory does not completely explain genetic variation and adaptation nevertheless, the theory is still used as a null hypothesis for testing evidence of natural selection [2, p160]. On the other hand, the selectionist theory, strongly supported by John Gillespie (1991) claims that mutations that are advantageous "are common enough that they cannot be ignored" [25, p251]. The theory predicts that the rate of substitution for many genes in most populations, "will reflect the action of natural selection on advantageous mutations" [25, p251].

Natural selection is the process in which individuals that have the best chances of survival in a certain environment will be the ones that reproduce, and therefore transfer their genetic material to the next generation. This ability of the individual to survive and reproduce in an environment is referred to as fitness. Selection occurs on the level of the individuals, but the consequences are observed on the level of populations [25, p90]. It acts on phenotypes, meaning the observable characteristics of individuals, but what is changed in the process are the gene frequencies [25, p90]. The problem also lies in the choice of word "selection" as it can imply that there is an intention behind it [25, p93] and that nature is looking forward and picking what will be good for survival. However, "natural selection is not forward looking" [25, p91], rather it is adapting the organisms to the conditions that existed and not the ones that will exist in the future [25, p91]. Take for example the case of the peppered moth which has two forms, a dark-coloured and a light-coloured form. At some point in history, the majority of moths were light-coloured, and the minority, about 1%, were dark-coloured because the tree trunks in the environment where they lived were lightly coloured and therefore the dark-coloured ones would be more visible, making them an easy target for the birds to see and eat them. With the coming of the industrial revolution and the consequent drastic increase in pollution, the environment changed, the trees became darker and so the light-coloured moths were easier for birds to spot and therefore more likely to be eaten. This led to an increased number of dark-coloured moths because they were the ones surviving and therefore reproducing. This example demonstrates very clearly how a change in phenotypes, based on the changes in the environment, is driven by natural selection. Another misconception about natural selection is that it is random, however, "it is, by definition, the non-random superiority at survival and reproduction of some variants over others" [25, p93]. Selection acts on individuals and not for the good of the species; there is no altruistic behaviour.

Even if there are singular examples of it like prairie dogs giving alarms when predators approach, drawing attention to themselves, if a certain trait does not increase the individual's (the one who has the trait) fitness when compared to other individuals, the trait will not evolve. If there was an allele that caused purely altruistic behaviour, the bearer of that trait would have less chance of survival and reproduction and therefore, it would quickly disappear from the population [25, p94].

Mutation is the "ultimate source of genetic variation" and without it, there would probably be no evolution [25, p213]. Mutation represents random changes in the genetic code that can include a change in single DNA bases, insertion and deletion, and other rearrangements of DNA sequences [26, p19]. As mutation constantly supplies new alleles, natural selection removes the deleterious alleles and keeps the beneficiary ones, but some of the deleterious persist. If the deleterious alleles are removed by selection at the same rate as they are being created by mutations, the allele frequency is in equilibrium and this is called mutation-selection balance [25, p214]. Mutation rate varies between organisms and it is very low; in human about 1.20×10^{-8} mutations per nucleotide site per generation. This rate is too low for a mutation to reach high frequency or even to become fixed in a population in a reasonable evolutionary time [2, p54]. The probability of fixations of a new mutation under neutral evolution is calculated with the equation $p_t = (1-\mu)^t p_0$ where p_0 is the initial frequency of allele A, p_t is the frequency of allele A after t generations, and μ is the mutation rate [28, p53]. However, natural selection can drive the mutation to higher frequencies or even to fixation at a very fast rate [2, p54].

Genetic drift, a further important evolutionary force, is the process by which allele frequencies change due to random sampling. It is a non-adaptive mechanism of evolution because it leads to a loss of genetic diversity. The amount of genetic drift is larger in small populations (where the size of the population is defined by the number of reproductive adults and not the total number of individuals [26, p112]) over a period of many generations, and smaller in large populations. Its effect, fixation, or loss of alleles depends on the (effective) population size (Ne) [25, p251]. Fixation means that the allele frequency will be 100% - only one allele of a gene will remain in the population, and it happens faster in smaller than in larger population (as $t=4Ne$). Fixations do not occur very often, and it is more likely that the new allele will be lost [2, p143,144]. Fixation of deleterious recessive alleles can lead to a reduction in the fitness of individuals in a population [25, p276].

Migration or **gene flow** represents the moving of alleles between gene pools (collections of all genes in a population) of different populations. The mechanisms behind it can be various, “from long-distance dispersal of juvenile animals to the transport of pollen, seeds, or spores by wind, water, or animals” [25, p225]. It can introduce new alleles into a population, making different populations more similar genetically to each other [26, p19] and eliminating adaptive differences that have been introduced by natural selection to the population [25, p276]. Overall, migration can result in a change of the allele frequencies within a population [2, p154].

The extended Evolutionary Synthesis (EES) is a novel approach in contemporary evolutionary thinking that incorporates new concepts that have appeared throughout the years: evolutionary developmental biology, developmental plasticity, inclusive inheritance, and niche construction theory.

Evolutionary developmental biology, also referred to as evo-devo, aims to determine the developmental principles that are responsible for the phenotypic differences within and between populations, species, and higher taxa, by using comparative and evolutionary biology. One of the most important insights from evo-devo is that changes in gene regulations that influence the timing, location, and amount or type of gene product are influencing phenotypic variation [8].

Developmental (phenotypic) plasticity refers to the changes in the phenotypes of an organism that occur as a response to the environment. Plasticity can influence phenotypic evolution by facilitating colonization of novel environments, affecting population connectivity and gene flow, contributing to temporal and spatial variation in selection, and through its potential to increase the chance of adaptive peak shifts, radiations, and speciation events. [8].

Inclusive inheritance refers to other mechanisms that an offspring inherits from the parent that are not genetic, but a “variety of developmental resources that enable reconstruction of developmental niches”. This non-genetic inheritance has similar evolutionary consequences as plasticity, cultural inheritance, and niche construction e.g. by biasing the expression and retention of environmentally induced phenotypes which can then affect the rate and direction of evolution. Additionally, there is more proof that adaptive evolution may happen because of selection on epigenetic variants and not just genes [8].

Niche construction explains how the metabolism, activities, and choices of organisms modify or stabilize environmental states, which then can influence selection.

Niche construction affects development by incorporating environmental factors into it in a way that can become as dependable as genomic factors. Accumulation of the environmental changes is referred to as ecological inheritance and it is brought to the new generation by the niche-constructing activities of the previous generation. If organisms live in the same environment and make changes to it, they can influence each other which leads to direct or diffuse coevolution and can have a valuable effect "on the stability and dynamics of ecosystems on both micro- and macro-evolutionary timescales" [8].

EES is often not accepted by the traditional evolutionary biologists as they see evolutionary processes as those processes that directly alter gene frequencies, and phenomena such as niche construction or developmental plasticity do not do so. However, for other evolutionary biologists, developmental processes should not be discarded as they contribute to creating novel variants, contribute to heredity, generate adaptive fitness, and in that way influence the direction of evolution [8].

From the EES perspective, there are two key mechanisms: constructive development and reciprocal causation. The basis for constructive development is the view that the genome is not a static set of instructions but senses and responds to the signals that affect it. Organisms have the ability to respond to and alter internal and external states and in that way influence their development. It is not a simple mapping between the genotype and phenotype, rather there is condition-dependent gene expression and the developmental system responds to internal and external inputs through the physical properties of cells and tissues and 'exploratory behaviour' among microtubular, neural, muscular, and vascular systems. Reciprocal causation refers to the idea that organisms are not just a product of evolution, but also the cause of it [8].

1.4 Population Genetics and Natural Selection

Population, like many other concepts, does not have a unified, final definition, instead, it can mean different things depending on the context and intended use. Hedrick defined population as a "group of interbreeding individuals that exist together in time and space" [29], they occupy a particular area over a relatively long period of time, share a common gene pool and have children [29]. There are a few criteria to be considered when we want to define a population: geographical proximity, a common language, shared

ethnicity, culture, religion. These are not the only criteria that should be taken into account, nor they are necessary, rather they are some that can be considered [2, p327]. Population genetics is the study of the genetic composition of populations (or sets of populations, or species) and how it changes over time [26, p2]. Genetic variation can be observed within a population (differences and similarities of individuals within a population) and between different population (the average difference between two or more populations) [26, p3]. Population genetics can be used to investigate human variation and evolution by answering questions such as: what can genetic variation tell us about our history as a species, can it be used to trace human migrations, why do some human populations have high frequencies of harmful alleles (such as sickle cell allele), are certain genes resistant to some diseases and disorders, how come there is a substantial difference between a small population and its neighbours [26, p4].

Population size may vary but in the context of evolution, it is important to consider the (long-term) effective population (N_e) size [27]. In contrast to the census population size (which consists of number of individuals at a certain time), the long-term effective population size presents an ideal population size that allows real populations to be modelled as Wright-Fisher populations with the equivalent amount of genetic drift [30]. The effective population size is usually much lower than the census size of populations and may vary across different locations in the genome of a species (due to differences in the modes of transmission of different chromosomes such as X-chromosome and Y-chromosome inheritance). The concept of N_e provides thus the basic concept to compare different populations (or species) to each other and to model the amount of genetic drift expected in an idealized Wright-Fisher population. This allows for quantification of genetic drift in determining the effectiveness of mutation, recombination, migration, and selection [10].

Natural selection and genetic drift are the key evolutionary forces that causes allele frequencies to change over time. Genetic drift as a stochastic factor differs from selection based on the fact that in each new generation some chromosomes leave more descendants by chance than others. However, the influence of both evolutionary factors on allele frequencies depend N_e. This population parameter determines the rate at which populations lose alleles through random genetic drift and, together with the mutation rate, it determines the number of alleles and the level of linkage disequilibrium (LD) expected in populations. Genetic drift is inversely related to N_e, while the opposite is true for natural selection. The probability that a new neutral mutation is ultimately fixed is $1/(2N_e)$. Most

neutral mutations are lost, but some persist. Neutral mutation that persists, i.e., goes to fixation, will do it within about $4N_e$ generations. Human population when compared to other primates have a rather low N_e of about 10.000 individuals (in chimpanzee and gorilla N_e is about 20.000) [31]. Because of this considerably low N_e slightly deleterious mutations may therefore be less efficiently removed from human populations [10].

The different forms of natural selection acting on the DNA are **positive**, **negative** (purifying) and **balancing selection**. Positive selection favours beneficial mutations (i.e., the ones that increase the fitness of an individual) while negative selection eliminates deleterious mutations because of their selective disadvantage. Strong positive selection is called "selective sweep". Negative selection is more common and eliminates the mutations which have detrimental effects on fitness [25, p262]. Balancing selection maintains genetic polymorphism (different alleles at the selected loci) in populations either through heterozygosity advantage or negative-frequency dependent selection of alleles [32].

Mutation-selection balance occurs when deleterious alleles are removed by selection at the same rate at which new copies of the same alleles are created by mutations, and it can explain why some deleterious alleles persist in the population. This means that the allele frequency is at equilibrium and it is calculated by the formula (in the case of recessive alleles:

$$q = \sqrt{u/s} \quad (1)$$

where q is the equilibrium frequency, u is the mutation rate, and s is the selection coefficient which represents the strength of selection against the allele with values between 0 and 1. The equilibrium frequency will be high when the selection coefficient is small and the mutation rate is high, and the equilibrium frequency will be low when the selection coefficient is high and the mutation rate is low [25, p220, 221].

1.4.1 Selection and patterns of linkage disequilibrium (LD)

Linkage disequilibrium (LD) is the tendency of two alleles to appear on the same chromosome above the random rates in a population [2, p83]. It is "the difference between the observed frequency of a two-locus haplotype and the expected frequency (based on observed allele frequencies) if the alleles were randomly segregating" 2, p84, Box 3.5].

If we take two loci, A (alleles A and a) and B (alleles B and b), and their haplotype AB, the observed frequency of that haplotype is P_{AB}. The expected frequency of the AB haplotype is P_A x P_B where P_A is the observed frequency of allele A and P_B is the observed frequency of allele B. LD, marked with *D*, as defined by Lewontin is:

$$D = P_{AB} - P_A P_B \quad (2)$$

When *D* is significantly different from zero, regardless of the sign (positive/negative), it is said that LD exists. The downside of this measure is that since it depends on allele frequencies, comparison of different values of *D* is of limited use. Another measure, labelled *D'*, can be put to greater use. It is defined as "the absolute value of *D* divided by its maximum possible value given the allele frequencies at the two loci" [2, p84, Box 3.5]. *D'* is equal to 1 only in the case when two alleles were not separated by recombination during the history of the sample being analysed. This is a complete LD and no more than three out of four possible two-locus haplotypes are observed (figure 1) [2, p84].

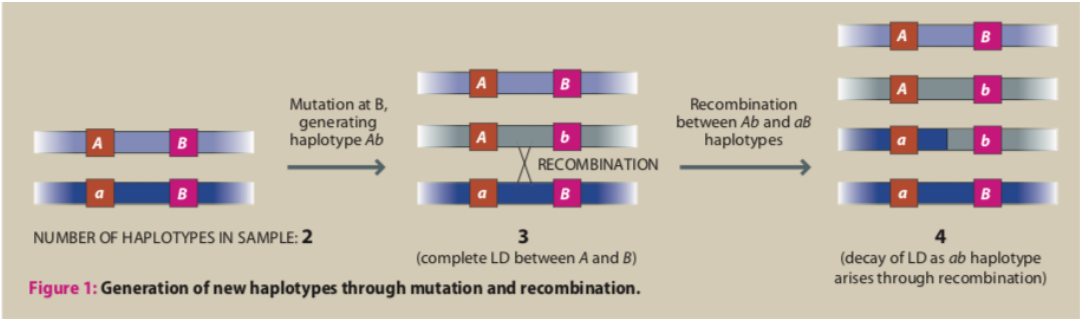

Figure1: Generation of new haplotypes through mutation and recombination [2. p84]

LD can also be seen as "measure of recombination at the population level". A new mutation that appears on a chromosome is linked to every other locus on that same chromosome and it forms a single haplotype which means that it will be associated only with one allele at each of the loci, meaning it is in complete LD with them (*D* is at its maximum). The frequency of the new mutation can increase over several generations and recombination will introduce the new allele into the copies of the chromosome with different alleles at other loci (Figure 1), which results in a LD decrease. The LD decrease can be tracked after some number of generations (*t*) if we know the recombination rate per generation (*r*) between the newly mutated locus and a given locus with the equation:

$$D_t = (1 - r)^t \times D_0$$

(3)

where D_t is the present value of D and D_0 is the initial value of D. As time increases, D_t tends towards zero (linkage equilibrium). Along the chromosome, the interlocus recombination rate increases with the distance from the newly mutated locus, decreasing D_t to zero even more quickly [2, p140].

A haplotype (derived from "haploid" and "genotype") represents a group of genes that were inherited together from one of the parents. It can refer to a pair of genes or all the genes that were inherited together from one parent on one chromosome. The genes are inherited together because of genetic linkage, which is to say that genes are often inherited together because they are located near each other on the same chromosome. A haplotype can also be used to describe the inheritance of a cluster of SNPs. Haplotypes can be useful for identifying patterns of genetic variation associated with a certain disease. For example, when a haplotype has already been associated with a certain disease, we can investigate the DNA around the haplotype in order to identify the gene or genes that are causing the disease [33]. Haplotypes can be used to detect positive selection in populations, as well as distinguish different forms of selective sweeps [34].

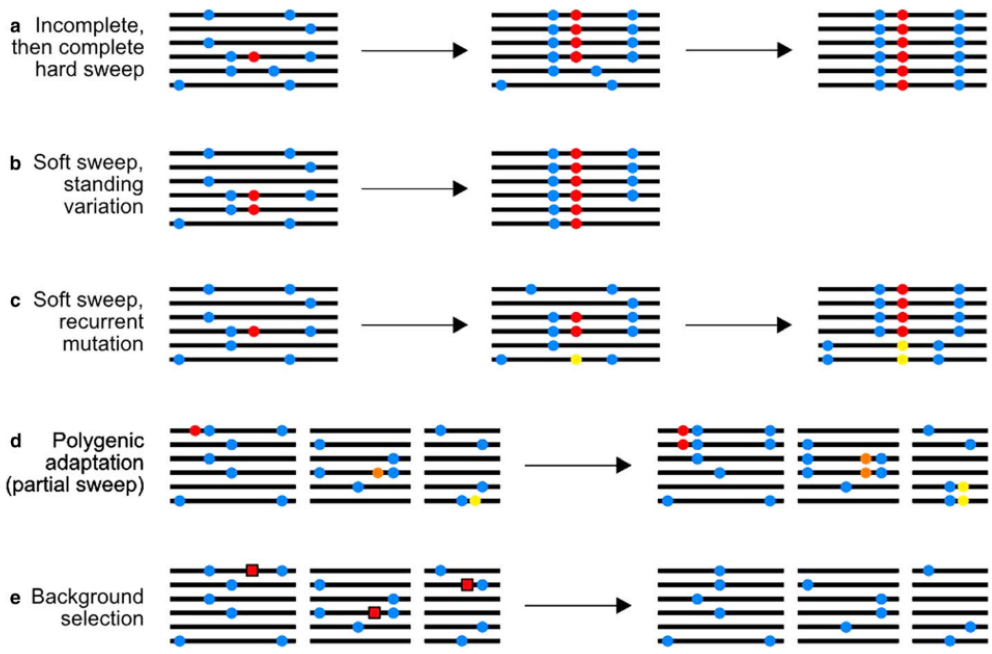

Figure 2: a. hard/classic sweeps b.c. soft sweeps d. incomplete/partial sweeps e. Background selection

Hard sweeps (Figure 2a) are the most studied sweeps. They focus on cases in which a new mutation that has beneficial effects rapidly increase in frequency until it reaches fixation in the population. During the sweep, the beneficial mutation carries with itself the portion of the haplotype on which it appeared. This reduces the levels of neutral diversity in the surrounding area. Soft sweeps are the ones in which several haplotypes could be led to high frequency because of their beneficial effects, but none of them becomes fixed. Figure 2b shows the case in which an allele is associated with multiple haplotypes, and therefore those haplotypes also increase in frequency. Figure c shows the case in which the same new allele appears but in different mutation events. Incomplete/partial sweeps (Figure 2d) are the ones in which the beneficial allele appears and increases in frequency but does not reach fixation, in which case "there will be some loss of linked neutral diversity". Background selection (Figure 2e) happens when natural selection removes deleterious mutations which then leads to linked neutral alleles being removed as well [34].

During the sweeps, a favoured allele carries linked variants on the same chromosome. If there is a hard sweep of a single new beneficial mutation and no recombination or mutation happens, the haplotype will become completely dominant in that population. This situation shows that the absence of haplotype diversity or an increase in LD between alleles at different sites can signal the effect of positive selection. "In the case of soft sweeps, more than one haplotype may be elevated to a high frequency, and in the case of incomplete and partial sweeps, a single haplotype may be at a higher frequency than expected under null models" [34].

Heritability is "the proportion of variance in a trait that is explained by genetic factors" [2, p483, Box 15.3]. We can distinguish broad-sense heritability and narrow-sense heritability. The former refers to the proportion of variance that is caused by all the genetic effects: additive (due to homozygous alleles), dominant (due to heterozygous alleles), and epistatic (due to the interaction between genes). The latter refers to the proportion of variance that is caused only by the additive effects [2, p483, Box 15.3].

1.4.2 Genome-Wide Association Studies (GWAS) used in human genetic studies

Genome-Wide Association Studies (GWAS) are conducted to identify the most common single nucleotide polymorphisms (SNPs) in the human genome and to find out the distribution of the SNPs across different populations. These studies are used to

discover the association between individual SNPs and disorders (or other complex traits). They have an impact on medical genetics as the results can be used to estimate a person's risk for acquiring a certain disorder, as well as personalize treatment [35]. In order to associate a gene with a disease, hundreds of thousands to millions of SNPs have to be typed across the whole genome, and this has been made possible through the development of SNP chips [2, p552].

Associations are made by comparing the distribution of SNPs between people with and without the trait in question (case-control studies). SNP is associated with a disease if people who have the disease also have the SNP more or less often than people who do not have the disease. Association strength is usually calculated by the odds ratio (OR) which can be determined from the results of the study, if it represents the wider population and matches cases of other factor that could influence the association. The factor does not have any effect on risk if the odds ratio is 1, and it rarely goes above 1.5 (usually it's closer to 1.1) [28 p477, 609]. Associations are quantified by their *p* values.

In complex diseases, SNPs contribute only partially, there are other genetic (e.g., epigenetic) and environmental factors that also contribute to the onset of complex disease [35]. GWAS only detects a region of linkage disequilibrium (LD) that contains the causal variants, further (fine-mapping and functional) studies are needed to narrow that region and identify the causal variants. These follow-up studies can be very time consuming and lab demanding, therefore, knowing the most important SNPs can make them easier, faster as well as more relevant. GWAS are useful for investigating links between genotype and phenotype. They are able to detect small genetic effects as well as loci connected to complex diseases that have multiple loci associated with them. GWAS Catalog is a resource of GWAS and their results that are publicly available. It contains a large amount of easily accessible data [36].

The advantages of GWAS are hypothesis-free approach (no need to set a candidate gene beforehand) and low cost per sample which allows for a large number of subjects to be analysed. These advantages make it possible to identify genes with low prior probability and susceptibility SNPs with small effect size, respectively. The studies are not biased by assumptions made in advance about the identity of the locus of interest. The disadvantage is that a precise significance level is required, typically $p < 5 \times 10^{-8}$ in order to identify a real association of a SNP allele with a disease and not just association that appeared because enough SNPs were tested [37, 2 p552]. GWAS also "make the assumption that the SNPs being tested will either include the causative allele itself or a

SNP in strong linkage disequilibrium (LD) with (tagging) the causative allele" [2, p552]. This assumption is supported by the fact that there exists a haplotype block structure in the human genome "because much of the common genetic variation within a haplotype block will be tagged by alleles at a few SNPs" [2, p552]. The number of SNPs that can tag a block depends on the LD structure of the block and how a haplotype block is defined. A rule that is usually followed (at least in European populations) is that one-third of all common SNPs within a block can tag most of the common variants [2, p552].

GWAS are good at identifying new associations between variants and traits, they can identify new biological mechanism, and there is a variety of ways their results can be applied in clinical practice (as this is one of the goals for genetic research) e.g. identification of risk for a certain disease, classification of diseases, drug development. GWAS can give us more understanding of how the genetics of complex traits differ between populations. They are important for discovering low-frequency and rare variants, and they can discover new genes for monogenic and oligogenic diseases. Apart from being used to study single nucleotide variants (SNVs) they can also be used for detection of copy number variants (CNVs), haplotypes, variable number, tandem repeats, retrotransposon insertion, polymorphisms, insertions, or deletions (indels), and inversions. Data generation, management, and analysis of GWAS are straightforward; the data can be easily shared and is publicly available [38].

GWAS also have certain limitations. One of the limitations is the need for high level significance in order to justify multiple analyses [38]. GWASs are very good at detecting the common alleles connected to a trait, however, the case of complex traits, these alleles are not the only ones associated with it which leads to a problem called "missing heritability" where we know that a trait is heritable, but we cannot provide a full explanation of the heritability. For example, only 10% of the variation in human height is explained by the combined effects of all associated variants from a GWAS of 184 000 people. However, as already stated, GWAS was designed to represent the most common variants in the human genome and those are the variants that it can find. Any other variants that are connected to the trait are rare and therefore not detectable by GWAS [2, p558]. GWAS results do not necessarily give us the exact causal variants and genes, so additional steps are needed to pinpoint them. Furthermore, they are not able to detect all the associations of complex traits (though, admittedly, none of the current methods and technologies can do so) and they have not been able to detect gene-gene interactions (epistasis). Last, the clinical predictive value is limited as for complex traits (which most

disorders go under) heritability is not explained in its entirety and the identified variants have small effect size [38].

2. THE GENETICS OF PSYCHIATRIC DISORDERS

Psychiatric disorders belong to the group of complex disorders, meaning that there is not one gene that is responsible for their symptoms but rather multiple genes including the impact of environment and lifestyle (epigenetic changes in the genome). On the other hand, one gene does not have to be specific for one disorder but can be associated with many of them. Additionally, recent studies in psychiatric genetics have demonstrated that there is a strong connection between genes and environment - the environment, as well as human behaviour, can modify gene expressions. There are two ways in which genes and environment connection can be viewed: gene-environment interaction and gene-environment correlation. The former refers to the genotype of the person affecting the influence of the environment and vice versa, and it can be used to explain different individual reactions to the same environment. The latter "refers to the genetic differences determined by exposure to particular environments. According to this model –widely used in evolutionary biology– animals modify their environment through genetic programming with the aim of favouring adaptive phenomena" [1]. These two models are not mutually exclusive [1].

In complex disorder, a combination of alleles is responsible for causing the disorder and if we were to investigate how one allele contributes to it, we would need to determine the genotype of the disorder. If the alleles are not among genotyped variants, they can be used as proxies. This can be determined by genotyping technologies and statistical methods that consider the frequency of each genotype in cases and controls. The chance that the association will be discovered depends on the effect size of the allele as well as the sample size. Describing the phenotype can be difficult because there are similar ones that have different causes. This can be partially solved by using endophenotypes which are more readily definable quantitative parameters, and which represent parts of diseases phenotypes. As already mentioned, the environment also influences the phenotype and therefore endophenotypes too, and it can be individual-specific, shared within families, or shared within populations. Another point that needs to be considered is that alleles are interacting among themselves but also with the

environment. The frequencies of the alleles are influenced by evolutionary forces (natural selection, mutation, genetic drift) which cause variations between individuals of the same population but also among populations. Taking an evolutionary perspective can help in explaining the appearance of complex disorders and this approach is called evolutionary medicine [2, p541].

It has been shown that psychiatric disorders are influenced by genetic variation, however since many of them have clear selective disadvantages because the individuals who are affected by them are less likely reproduce and sometimes commit suicide, the question can be posed how are their susceptibility alleles surviving? There are three possible explanations for this. Ancestral neutrality says that the susceptibility alleles became disadvantageous too recently for purifying selection to act on them, and they were not disadvantageous in the past. Positive and/or balancing selection says that apart from the disadvantageous effects, susceptibility alleles also have beneficial effects which are responsible for their survival at a relevant frequency. Mutation-selection balance says that susceptibility alleles are indeed being removed by purifying selection because they are disadvantageous, but recurrent mutations are responsible for repeatedly generating new susceptibility alleles. If there was any positive selection of these alleles, their positive effects would have to be very strong to balance out the negative effects. The higher cognitive abilities that developed recently in modern humans are associated with the persistence of psychiatric disorders insofar as those disorders have features that are favoured with in society, for instance autism spectrum disorder has been connected with high intelligence, depressive disorders with creativity [2, p555], and schizophrenia with the development of language [4]. Determining which one of these possible explanation is the correct one is not an easy task because it demands knowing information from the past that we do not have since there was not much patient data collected at that time or at least not any one place or in connection to genes.

A number of GWAS have been conducted so far which have uncovered a range of genes associated with psychiatric disorders. Moreover, for some of these genes, it was shown that they were under positive selection during human evolution. One such example are the genes associated with autism spectrum disorder, and it has been assumed that they are under positive selection because of their involvement in neurogenesis and cognitive ability [39].

The disorders chosen for this study are the major depressive disorder (MDD), bipolar disorder (BPD), and attention-deficit/hyperactivity disorder (AD(H)D). These

disorders were chosen because they share certain phenotypic characteristics and the idea was to determine whether they also share genes (and biochemical pathways), whether these genes are under positive selection and whether there are differences between populations.

By analysing different populations, we hope to determine whether the environment has influenced the selection of the alleles, and how do these effects differ between the environments. If there are overlaps between the populations, it can mean that there was a single event that caused the selection, in which case the same haplotype should be present in each population [40].

2.1 Major Depressive Disorder

Major depressive disorder (MDD) belongs to the group of depressive disorders and it represents the classic condition in this group. The symptoms need to be present for at least a 2-week period, almost all day and nearly every day, and differ from the previous way of functioning. At least five of the following nine symptoms have to be present and either depressed mood or anhedonia need to be one of them. The symptoms are: depressed mood which is present most of the day and reported either by the person experiencing it or observed by others, anhedonia in all or most activities which is present most of the day, significant weight loss or weight gain without being on a diet (or decrease/increase in appetite), insomnia or hypersomnia, psychomotor agitation or retardation (observed by others and not just self-reported), fatigue, feelings of worthlessness or excessive/inappropriate guilt, decreased concentration or indecisiveness (self-reported or reported by others), and recurrent thoughts of death [41 p160, 161].

There needs to be clinically significant distress or impairment in social, occupational, or other important areas of functioning as a consequence of the symptoms and they should not be invoked by some substances or other medical conditions [41, p160, 161].

Takin into account the broad spectrum of symptoms, it is possible that two people have the same diagnosis but completely different symptoms. This heterogeneity also represents a part of the reason why there is no unified treatment for depression and why there are people that have a depressive episode but do not seek professional help [42,

p488]. Episodes usually do not last longer than 2 years, however, there is a great chance that they will recur if it is not treated and the more time they appear the greater the chance that they will appear again. It is twice as common in women as it is in men [43, p673]. However, some studies show that it is still up for debate whether that is the real state of the disorder or whether men simply under-report depressive episodes [42, p489]. It can appear at any age, and the course may be quite different; children and adolescents might experience irritable mood rather than sadness.

Suicide risk exists during the whole period of major depressive episode. Functional consequences can vary from very mild when it is not noticeable by others to complete inability to attend to basic self-care needs. Individuals with MDD suffer from more pain and physical illness and greater decreases in physical, social, and role functioning [41, p167].

2.1.1 State of the art

Several studies, including a study done in Sweden that included 42 161 twins, of which 15 493 were complete pairs [44], have shown that the heritability of MDD is higher in woman than in men, and it can be between 0.37 and 0.39. However, despite such high heritability, there are no consistent genetic findings that could explain the cause of the disorder. This could be due to the heterogeneity of the disorder and the groups used in the studies that try to identify the genes, as well as the polygenic nature of the disorder. Another twin study that included 1 029 female twin pairs showed that twins prone to depression share depressive symptoms which provides support for the theory that the genes associated with the depressive disorder depend upon the subtype of the disorder [42].

Linkage studies, including both common and rare genetic variants, identified some genomic regions that are reported at least twice, however the exact genes or causal variants that drive the linkage are still unknown. The key genomic regions identified so fare by linkage studies are: 75-80 Mb on chromosome 11, 37-42 Mb on chromosome 15, 87-92 Mb chromosome 15, chromosome 3, 4-9Mb, and 64-68 Mb chromosome 2 [42].

Recent genome-wide meta-analysis of depression, that included 807 553 individuals (246 363 cases and 561 190 controls) has identified 102 independent genetic variants, 269 putative genes, and 15 gene pathways associated with depression and an independent replication sample of 1 306 354 individuals (414 055 cases and 892 299

controls), showed that “87 of the 102 associated variants were significant after multiple testing correction” [95].

2.2 Bipolar Disorder

Bipolar disorder (BPD) is also a recurrent mood disorder characterized by repeated or mixed episodes of mania and depression which is why it is also called manic-depressive disorder [43, p673]. A manic episode consists of an abnormally and persistently elevated, expansive, or irritable mood, and abnormally and persistently increased goal-directed activity or energy that lasts almost all day, every day for at least a week. At least three of the following symptoms (if the mood is only irritable than four) need to be present and the individual needs to manifest a visible difference from their usual behaviour The symptoms are: magnified self-esteem or grandiosity, decreased need for sleep, more talkative than usual or pressure to keep talking, flight of ideas or the subjective experience that thoughts are racing, distractibility (self-reported or observed by others), increase in goal-directed activity or psychomotor agitation (self-reported or observed by others), and engagement in activities that have a high potential for painful consequences.

Mood change needs to cause noticeable impairment in social or occupational functioning or hospitalization because of possibilities for harming oneself as well as others, or psychotic features must be present. Manic episodes can be caused by some substances but if it only persists while the substance is being used, it is not considered to be part of bipolar I disorder. However, if it persists even after the substance is not being used anymore, it can be diagnosed as a manic episode belonging to a BPD. A hypomanic episode is a milder form of a manic episode. The symptoms are the same, but they are not severe enough to cause marked impairment in social or occupational functioning, hospitalization, or psychotic features. The depressive episode is the same as described before in the MDD section [41, p123, 125].

There are two types of bipolar. Type I is characterized by at least one manic episode, with or without a major depressive episode [41, p126]. Type II is characterized by at least one hypomanic episode and at least one major depressive episode, without the presence of a manic episode. Individuals with bipolar II usually seek help during the

depressive episodes as the hypomanic episodes do not cause them impairments, and they might not see the hypomanic episodes as problematic although people around them might. Therefore, it is useful to get information about a person's behaviour from family and friends as well in order to distinguish it from MDD or some other depressive disorders. However, a hypomanic episode should be distinguished from euthymia and restored energy or activity after a major depressive episode. Bipolar II should not be viewed as milder type of bipolar I especially since the individuals that experience it have more chronic health problems and depressive episodes last longer on average. However, there are not always clear boundaries between the episodes as depressive symptoms can co-occur with a hypomanic episode and vice versa, which also means that the individuals that experience the symptoms cannot always distinguish between them and might report a certain period as a depressive episode with increased energy or irritability, when it was actually a hypomanic episode [41, p134, 136]. Onset for the first manic, hypomanic, or major depressive episode is usually 18 years [41, p130].

Both types of bipolar are connected with elevated suicide risk and even though many of the individuals with BPD return to a fully functional level between episodes, about 30% of those affected by bipolar I and 15% of those affected with bipolar II show some inter-episode dysfunction, while 20% of those affected with bipolar II do not even have the inter-episode recovery but switch immediately into another mood episode. In both cases, even when the symptoms stop there is a certain time that needs to pass before the individual is functionally recovered as well, which can result in a lower socioeconomic status for the individual even when they have the same educational level as the unaffected part of the population. Individuals with bipolar I and II also do not perform as well as the unaffected individuals on cognitive tests. Cognitive impairment connected with bipolar II can contribute to vocational difficulties while the ones connected to bipolar I additionally include interpersonal difficulties and persist during the entire lifetime even when the person is in the euthymic period (period that is not manic or depressive but is still different than the usual state of people that are unaffected by BPD). Being unemployed for a longer period relates to more depressive episodes, older age, increased rates of current panic disorder, and lifetime history of alcohol use disorder [41, p131, 138].

2.2.1 State of the art

BPD is highly heritable and there have been consistent study results demonstrating that it aggregates in families, with relatives of people with BPD having an increased risk of the disorder compared to controls, and it seems that it doesn't differ between sexes. Additionally, some studies showed that risk of BP II is higher than BP I [45].

A study done in the United States that included 447 probands with BP I, BP II, or MD and 2 082 of their living first-degree relatives investigated separately mania and depression. The study concluded that mania and major depression are inherited separately in families, bringing into question whether or not the major components of BP have a common underlying pathway [46] which is in agreement with a previous study that concluded that mania and depression are separate processes that are highly comorbid [47].

A range of twin studies have shown that the heritability of BP can be from 73% to 93% and those numbers even increase when BP is observed on a broader spectrum that includes both BP I and BP II [45].

A systematic review of recent findings from GWAS of BP found that among 27 studies, 12 studies had genome-wide significance. The hallmark study was conducted by Psychiatric Genomic Consortium (PGC) in 2011. It had a large sample size and a replication analysis with a large number of independent samples. Two susceptibility genes found in this study were a SNP in the gene calcium voltage-gated channel subunit alpha1 C *CACNA1C* and a SNP in teneurin transmembrane protein 4 (*TENM4*) gene, however they both have a small effect size (1.14 for *CACNA1C* and 0.88 (=1/0.88 = 1.14) for *ODZ4*). These findings also imply that BPD is a polygenic disorder, because many SNPs that have a small effect size are causing it. The newest study by the PGC is the largest one yet consisting of approximately 30 000 BD cases and 169 000 controls; 30 susceptibility loci were reported, of which 20 were novel and a SNP in tetratricopeptide repeat and ankyrin repeat containing *TRANK1* gene (which was also previously reported) was the most significant one. This systematic review concluded that the current susceptibility SNPs are not enough to make a diagnosis, and additionally, the influence of the environment is probably affecting the development of BP [37].

2.3 Attention-Deficit/Hyperactivity Disorder

Attention-Deficit /Hyperactivity Disorder (AD(H)D) is categorized by the DSM-V as a neurodevelopmental disorder, meaning that it has its onset in the developmental period. It manifests in early development, often before the child starts school, and in order to be diagnosed at least several symptoms should be present before the age of 12. Neurodevelopmental disorders frequently co-occur. AD(H)D is characterized by "a persistent pattern of inattention and/or hyperactivity-impulsivity that interferes with functioning or development" [41, p59]. The child is inattentive in such a way that they are wandering off tasks, do not have persistence or focus, and are disorganized. Hyperactivity means that the child has excessive motor activity when it is not appropriate, or fidgeting, tapping, or talking. "In adults, hyperactivity may manifest as extreme restlessness or wearing others out with their activity" [41, p61]. Impulsivity, on the other hand, means that some actions occur without thinking about it and they can have a harmful effect on the individual, and may reflect a need for immediate rewards or inability to delay gratification. These symptoms do not have anything to do with disobedience or lack of comprehension of the tasks and instructions. Manifestation of the symptoms has to be present in more than one environment (e.g. home, school, work) but symptoms vary depending on the context of the environment. Symptoms may be minimal or even absent if the individual "is receiving frequent rewards for appropriate behaviour, is under close supervision, is in a novel setting, is engaged in especially interesting activities, has consistent external stimulation (e.g., via electronic screens), or is interacting in one-on-one situations (e.g., the clinician's office)" [41, p61]. This disorder affects about 5% of children and 2.5% of adults in most cultures, and the heritability is substantial [41, p61]. "AD(H)D is associated with reduced behavioural inhibition, effortful control, or constraint; negative emotionality; and/or elevated novelty seeking. These traits may predispose some children to AD(H)D but are not specific to the disorder" [41, p62].

AD(H)D consequences are "reduced school performance and academic attainment, social rejection, and in adults, poorer occupational performance, attainment, attendance, and a higher probability of unemployment as well as elevated interpersonal conflict" [41, p63]. Children with AD(H)D have greater chances than those without it to develop conduct disorder in adolescence and antisocial personality disorder in adulthood, which then also increases the likelihood for substance use disorder and incarceration.

Individuals with AD(H)D are perceived by society as lazy, irresponsible, or uncooperative, and personal relationships might be characterized by discord and negative interactions. They are often rejected, neglected, and teased by their peers. There is great variability in how the individual with AD(H)D deals with the symptoms and what the consequences are but in its most severe form, the disorder is impairing, affecting social, familial, and scholastic/occupational adjustment. Inattention mostly leads to academic deficits, school-related problems, and peer neglect, while hyperactivity leads to peer rejection and accidental injuries [41, p63]. Two-thirds of children diagnosed with AD(H)D will continue to have impairments into adulthood [48]. There is also a possibility of AD(H)D emerging in adulthood.

2.3.1 State of the art

It has been shown through family, twin, and adoption studies that AD(H)D is highly heritable, about 74% of which a third is due to the polygenic nature of the disorder. Linkage studies did not show any genome-wide significant results which suggests that there are probably no common DNA variants that have a high impact on AD(H)D. On the other hand, candidate gene association studies have found six relevant genes for AD(H)D: the serotonin transporter gene (*5HTT*), the dopamine transporter gene (*DAT1*), the D4 dopamine receptor gene (*DRD4*), the D5 dopamine receptor gene (*DRD5*), the serotonin 1B receptor gene (*HTR1B*) and a gene coding for a synaptic vesicle regulating protein (*SNAP25*). For adult AD(H)D it has been shown that the gene brain-specific angiogenesis inhibitor 1-associated protein 2 (*BAIAP2*) has significant association [48].

Genome-wide association meta-analysis of 20 183 AD(H)D cases and 35 191 controls identified significance at 12 independent loci which "are enriched in evolutionarily constrained genomic regions and loss-of-function intolerant genes, as well as around brain-expressed regulatory marks". These loci are connected to the genes forkhead box P2 (*FOXP2),* dual specificity phosphatase 6 *(DUSP6),* semaphorin 6d *(SEMA6D), ST*3 beta-galactoside alpha-2,3-sialyltransferase 3 *(ST3GAL3),* long intergenic non-protein coding RNA 461 *(LINC00461),* transmembrane protein 161b *(TMEM161B),* and *MEF2C Antisense RNA 1 (MEF2C-AS1)* [49]. *CACNA1C* has also been associated with AD(H)D [50].

The genetic complexity surrounding AD(H)D is probably due to the gene-gene interactions, gene-environment interactions, or gene-environment correlations. The

environment is included in the AD(H)D etiology which can be seen from the twin studies, as well as the fact that many of the SNPs are located in the regulatory regions rather than the coding regions and it is suspected that they work through epigenetic mechanism, however this is yet to be investigated [48].

2.4 Comorbidity and Co-occurrence

2.4.1 Phenotypic overlap

Certain symptoms are shared between the three disorders described previously, and sometimes the only way to distinguish the disorders is by the duration of the symptoms and whether they appear in episodes or not. This makes it interesting to compare them to one another and investigate potential shared genetic information between them. Additionally, the three disorders occasionally co-occur as well.

Depressive episodes are part of BPD, however major depressive episodes that have a very visible irritable mood can be hard to differentiate from manic episodes with irritable mood. On the other hand, when we compare MDD with AD(H)D they both share distractibility and low frustration tolerance, and in cases where both are present, AD(H)D can be diagnosed in addition to MDD [41, p167]. Inability to concentrate is another overlapping symptom, however, in MDD it is present only during a depressive episode while in AD(H)D it is almost always present [41, p64]. Diagnosis should be made carefully in children with AD(H)D as mood disturbance is characterized by irritability rather than sadness or loss of interest [41, p167].

AD(H)D can co-occur with Bipolar I [41, p132], but it can also be misdiagnosed as Bipolar I or Bipolar II, particularly in adolescents and children. Rapid speech, racing thoughts, distractibility, and less need for sleep are symptoms that overlap between the manic episode of bipolar I, bipolar II and AD(H)D. Differentiation is made based on the occurrence of the symptoms. If they appear in episodes then they belong to a manic episode of bipolar I, while an apparent increase over baseline needs to be present for bipolar II [41, p131, 132, 138, 139]. Mood changes within the same day can be present in children with AD(H)D, however, in order to be associated with BPD, those changes need to persist for at least 4 days. In certain cases, MDD can include also hypomanic or

manic symptoms but not all of them, and for a shorter period of time than it is the case in mania or hypomania. In order to correctly determine the disorder, whether it is Bipolar I, Bipolar II or MDD, it is important to check the individual's history of past episodes. This may depend on self-report from the affected person and in the event that they are in the midst of one of the episodes while they are reporting, it might be difficult for them to correctly report past experiences. Irritability can be connected with both MDD and BPD [41, p131, 138]. Bipolar I and Bipolar II are distinguished based on past episodes of mania [41, p131] The co-occurrence of BPD and AD(H)D happens more often than it would be expected by chance, between 9.5% and 28% of individuals with BPD also have AD(H)D, while for individuals with AD(H)D, the occurrence of BPD is around 20% [22].

2.4.2 Genetic overlap

CACNA1C has been associated with BPD and major depression (as well as some other psychiatric disorders) through both GWAS and clinical studies [21, 39, 50] and with AD(H)D through GWAS [50]. It is hypothesized that association of *CACNA1C* with multiple psychiatric disorders is a result of its broad expression in the limbic regions and neuronal circuits that are important for emotion, motivation, and cognition. Changes to gene expression that can arise during individual's development and adulthood can cause different behavioural symptoms and susceptibility to stress [39].

There seems to be a significant and consistent number of SNPs associated with AD(H)D and BPD, as well as overlap between the two, however there may be a difference when early onset BPD is compared to late onset. These genome-wide significant regions are centrosomal protein 85 like (*CEP85L*), tata-box binding protein associated factor 9b pseudogene 2 (*TAF9BP2*), adenylate cyclase 2 (*ADCY2*), and actin binding lim protein 1 (*ABLIM1*) [22].

Genome-wide analysis of risk loci that have shared effects on five major psychiatric disorders [21], including AD(H)D, BPD, and MDD, found that most strongly associated SNPs are within two L-type voltage-gated calcium channel subunits, *CACNA1C* and calcium voltage-gated channel auxiliary subunit beta 2 (*CACNB2*).

A meta-analysis of the genome-wide association data of BPD and MDD [23] found that *CACNA1C* and ankyrin 3 (*ANK3*) SNPs exceed the genome-wide significance, and one study that compared genome-wide analysis of schizophrenia, BPD, and MDD, have found that adrenomedullin (*ADM*) seems to be shared between these disorders [51].

3. MATERIALS AND METHODS

3.1 Data

3.1.1 Genetic Data

We used the phased genetic data of the 1000 Genomes Project phase 3 data (FTP server: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). From this database we included single nucleotide polymorphism (SNP) data from 12 human populations (3 from Africa, 3 from East Asia, 3 from South Asia and 3 from Europe) with the following genetics ancestries and number of subjects (n): Luhya in Webuye, Kenya - LWK (99), Gambian in Western Divisions in the Gambia - GWD (113), Esan in Nigeria - ESN (99), Finnish in Finland - FIN (99), Toscani in Italia - TSI (107), British in England and Scotland - GBR (91), Punjabi from Lahore, Pakistan - PJL (96), Bengali from Bangladesh - BEB (86), Indian Telugu from the UK - ITU (102), Han Chinese in Beijing, China - CHB (103), Japanese in Tokyo, Japan - JPT (104), Kinh in Ho Chi Minh City, Vietnam - KHV (93) [52].

3.1.2 Genome-Wide Association Studies

We used data from the GWAS catalog available online (https://www.ebi.ac.uk/gwas/ , accessed on the 9th of February 2019.) to identify SNPs that have been associated with the disorders in question (BPD, MDD, and AD(H)D). For further analysis, we took only the SNPs with the p-value equal or smaller than 5×10^{-8}.

In total we obtained twenty-two GWAS for MDD (Table 1), thirty-one GWAS for BPD (Table 2), and eight GWAS, for AD(H)D (Table 3).

Table1: Genome Wide Association Studies used to identify SNPs for Major Depressive Disorder

Author	Name of the study	Genes
Direk N	An Analysis of Two Genome-wide Association Meta-analyses Identifies a New Locus for Broad Depression Phenotype.	FHIT
Amare AT	Association of the Polygenic Scores for Personality Traits and Response to Selective Serotonin Reuptake Inhibitors in Patients with Major Depressive Disorder.	XKR6, PLEKHM1, PRAG1, PLXNC1, MSRA, ELAVL2, YEATS4, BRUNOL4, LOC105379222, LOC105369912, LOC105369913, LOC107987055, LOC105369298, LYZ, CELF4
Ji Y	Citalopram and escitalopram plasma drug and metabolite concentrations: genome-wide associations.	CYP2C9, CYP2C19, SEPT, LOC100132273, LOC388906, CYP2D6, CYP2D6*4, CYP2D6*10, CYP2D6*14 TRIML1, LOC102723706, SEPT3, LOC102723722, LOC105377607, RPL7AP27
GENDEP Investigators	Common genetic variation and antidepressant efficacy in major depressive disorder: a meta-analysis of three genome-wide pharmacogenetic studies.	MYO10, LINC01021 - LOC105374696, LOC105377684
Li X	Common variants on 6q16.2, 12q24.31 and 16p13.3 are associated with major depressive disorder.	RBFOX1, LOC105377628, VRK2, LOC105370772, RPS15P8, PCDH9, LOC391334, MIR1281, LOC105378797, RN7SL618P, ZNF646P1
Mbarek H	Genome-Wide Significance for PCLO as a Gene for Major Depressive Disorder.	PCLO
Hyde CL	Identification of 15 genetic loci associated with risk of major depression in individuals of European descent.	TMEM161B, MEF2C, VRK2, L3MBTL2, NEGR1, RERE, HACE1, LIN28B, SORCS3, OLFM4, PAX5, MEIS2, TMCO5A, RSRC1, MLF1, SLC6A15, KIAA0020, RFX3, AS1, CHADL, LOC105378797, HMGB1P18, HNRNPA1P64, LOC100996447, LOC105377703, LOC107983986, LOC105378799, CARM1P1

Huang J	Cross-disorder genomewide analysis of schizophrenia, bipolar disorder, and depression.	ADM, CAND1.11
Zhou H	Genetic Risk Variants Associated With Comorbid Alcohol Dependence and Major Depression.	SEMA3A
Okbay A	Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses.	LOC105378797, LOC105372648, ARFGEF2, LOC105375974, LOC105375975, DCC, KSR2
Wray NR	Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression.	RERE, SLC45A1, NEGR1, LINC01360, DENND1B, VRK2, LINC01876, NR4A2, TOPAZ1, TCAIM, RSRC1, LOC100996447, SLC30A9, LINC00682, DCAF4L1, LINC00461, MEF2C, LOC101927421, TENM2, FBXL4, C6orf168, TMEM1068, VWED, PUM3, LINC01231, ASTN2, DENND1A, LHX2, SORCS3, DKFZp686K1684, PAUPAR, ELP4, PAX6, SOXS, ENOX1, LACC1, CCDC122, OLFM4, LINC01065, LRFNS, SYNE2, MIRS48H1, DLST, PROX2, RPS6KL1, BAGS, APOPT1, RBFOX1, SHISA9, CPPED1, PMFBP1, DHX38, CRYBA1, MYO18A, NUFIP2, MIR924HG, DCC, MIR4258, RAB27B, CCDC68, TCF4, MIR4529, L3MBTL2, EP300-AS1, CHADL
Power RA	Genome-wide Association for Major Depression Through Age at Onset Stratification: Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium.	C3orf70, VPS8, EHHADH, MAP3K13, AS1
Maciukiewicz M	Genome-wide association studies of placebo and duloxetine response in major depressive disorder.	STAC, RFC3P1
Ikeda M	Genome-wide environment interaction between depressive state and stressful life events.	BMP2 , LINC01428, LOC101929288

Howard DM	Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways.	HCG9, TRIM31, AS1, UBQLN1P1, MICC, TMEM183AP1, SNORD32B, ZKSCAN3, ZSCAN12, OR5V1, HCG4, LOC105375013, LOC105378797, LOC101929006, LOC105375005, OR2N1P, OR2J2, GPX6, THSD7A, TMEM106B, HIST1H4L, LOC107986585, ZSCAN12P1, ZNF165, TRA-AGC4-1, TRA-CGC2-1, HCG16, LOC105375001, TRS-AGA2-3, TRF-GAA4-1, TRA-AGC2-1, GRM5, TRT-CGT5-1, TRI-TAT2-3, LOC107984363, TYR, LOC105379109, ABT1, CTTNBP2, LOC105375470, B3GALTL, TRS-GCT1-1, RNU2-62P, ASTN1, TRI-AAT6-1, TRNAM-CAU, NBAS, BTN3A1, BTN3A2, LINC00240, LOC100420530, MUC21, LOC105375015, HLA-B, SORCS3, KCNQ5, LOC105375983, LOC105370959, MRPL46, SGIP1, LOC105378800, LOC105378799 , KRT8P21, LOC105375975, LOC101929446
Hall LS	Genome-wide meta-analyses of stratified depression in Generation Scotland and UK Biobank.	CRTAP
Smoller JW	Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis.	ITIH3, DPYD, MIR137, CCDC68, CNNM2, ANK3, SYNE1, DCP1B, CACNA1C, NT5C2, MMP16, ODZ4, CSMD1, CACNB2, PCGEM1, LOC105372125, ZNRD1, LOC107984265, AS3MT, C10orf32-ASMT, LOC105375629, LOC105375631, TENM4, LOC107985969
Liu Y	Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder.	CACNA1C
Peterson RE	Molecular Genetic Analysis Subdivided by Adversity Exposure Suggests Etiologic Heterogeneity in Major Depression.	SLC25A37, LPGAT1, C1ORF95, ITPKB

Biernacka JM	The International SSRI Pharmacogenomics Consortium (ISPC): a genome-wide association study of antidepressant treatment response.	HPRTP4, LOC101060084
Kohli MA	The neuronal transporter gene SLC6A15 confers risk to major depression.	SLC6A15, LOC107984536
Wong ML	The PHF21B gene is associated with major depression and modulates the stress response.	OR2T12, TMEM150B, PRR5-ARHGAP8, MUC5B, LOC105373062

Table 2: Genome Wide Association Studies used to identify SNPs for Bipolar Disorder

Author	Name of the study	Genes
Ikeda M	A genome-wide association study identifies two novel susceptibility loci and trans population polygenicity associated with bipolar disorder.	FADS2, C11orf9, C11orf10, FEN1, FADS1, NFIX, TRANK1, MAD1L1, ODZ4, DAGLA, DDN, LOC100652964, MLL2, RHEBL1, SYNE1, LOC105377030, LOC105377031, TENM4, KMT2D
Baum AE	A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder.	DGKH
Willour VL	A genome-wide association study of attempted suicide.	FAM110C, LOC105373324
Liu X	A genome-wide association study of bipolar disorder with comorbid eating disorder replicates the SOX2-OT region.	SOX2-OT
Amare AT	Association of Polygenic Score for Schizophrenia and HLA Antigen and Inflammation Genes With Response to Lithium in Bipolar Affective Disorder: A Genome-Wide Association Study.	HCG4, ADAMTSL3, ERCC4, CCNH, EPHX2, ZNF804A, DMA, FAM177A1, MEF2C, ADCY1, CMAHP, MYO1H, MAIP1, FAM178B, GRAMD1B, EFTUD1P1, DNM1P41, SHISA9, LOC105371093, LOC105379066, LOC105379069, BRD2, LINC00461, LOC100132239, LOC107986530, LOC105373833, SPATS2L
McElroy SL	Bipolar disorder with binge eating behavior: a genome-wide association study implicates PRR5-ARHGAP8.	PRR5, ARHGAP8, LOC105373536, LOC105373538, BCKDHB
Ferreira MA	Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder.	ANK3,CACNA1C
Huang J	Cross-disorder genomewide analysis of schizophrenia, bipolar disorder, and depression.	ADM, CAND1.11

Charney AW	Evidence for genetic heterogeneity between clinical subtypes of bipolar disorder.	TRANK1, FER1L5, LMAN2L, CNNM4, ANK3, CACNA1C, ADD3, LOC100505933, XPNPEP1, SYNE1, LOC105377030, LOC105377031, ITIH1, RHEBL1, DHH
van Hulzen KJE	Genetic Overlap Between Attention-Deficit/Hyperactivity Disorder and Bipolar Disorder: Evidence From Genome-wide Association Study Meta-analysis.	ADCY2, TAF9BP2, CEP85L, ABLIM1, SLC35F1, LOC644303
Hou L	Genetic variants associated with response to lithium treatment in bipolar disorder: a genome-wide association study.	PPIAP22, LOC107985490
Fabbri C	Genetics of long-term treatment outcome in bipolar disorder.	DFNB31, TRAF3IP2-AS1, LOC105376229
Lencer R	Genome-wide association studies of smooth pursuit and antisaccade eye movements in psychotic disorders: findings from the B-SNIP study.	IPO8, PCDH12, LMO7, GIGYF1, POP7, LOC105372897, RPN2, STX2, ZNF740, C11orf21, CABLES1, LTN1, PLCB4, LINC01098, LOC107985255, TSPAN32, RNA5SP173, LOC105377563
Athanasiu L	Genome-wide association study identifies common variants associated with pharmacokinetics of psychotropic drugs.	LOC101927995, LOC105377013, NRG1, RNU6-663P
Cichon S	Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder.	NCAN
Song J	Genome-wide association study identifies SESTD1 as a novel risk gene for lithium-responsive bipolar disorder.	PLET1, PTS, SESTD1, CCDC141
Kerner B	Genome-wide association study in bipolar patients stratified by co-morbidity.	MARK1, PDE10A, RPL23AP39, RPL21P17, LOC105372929
Chen DT	Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder.	TRANK1, LMAN2L, ANK3, PTGFR, LOC105377031, LOC101929054, PPM1M, LOC652549, ADGRL4

Wellcome Trust Case Control Consortium	Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.	DCTN5, PALB2, NDUFAB1
Winham SJ	Genome-wide association study of bipolar disorder accounting for effect of body mass index identifies a new risk allele in TCF7L2.	TRIM42, TCF7L2, LOC102724068
Greenwood TA	Genome-wide association study of temperament in bipolar disorder reveals significant associations with three novel Loci.	FBLN1, INTS7, MDM1, LOC105369819, LOC100507195
Muhleisen TW	Genome-wide association study reveals two new risk loci for bipolar disorder.	ANK3, ODZ4, ADCY2, POU3F2, MIR2113, TRANK1, NCAN, TENM4, EIF4EBP2P3, LOC105377031, MAU2
Sleiman P	GWAS meta analysis identifies TSNARE1 as a novel Schizophrenia / Bipolar susceptibility locus.	MAD1L1, SDCCAG8, ITIH1, CSMD1, GPR89P, TRV-AAC1-5
Qi G	Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits.	LOC105377030, LOC105377031
Smoller JW	Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis.	ITIH3, DPYD, MIR137, CCDC68, CNNM2, SYNE1, DCP1B, CACNA1C, ANK3, NT5C2, AS3MT, MMP16, ODZ4, CSMD1, CACNB2, PCGEM1, TRIM26, LOC105372125, ZNRD1, LOC107984265, C10orf32-ASMT, LOC105375629, LOC105375631, TENM4, LOC107985969
Sklar P	Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4.	CANCNA1C, ODZ4, TENM4
Liu Y	Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder.	ANK3, CACNA1C
Ruderfer DM	Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia.	CACNA1C, TRANK1, MAD1L1, PIK3C2A, IFI44L, GPR89P, TRV-AAC1-5, LOC105377030, LOC105377031, LOC652549, ADGRL4

Table 3: Genome Wide Association Studies used to identify SNPs for Attention-Deficit/Hyperactivity Disorder

Author	Study	Genes
Martin J	A Genetic Investigation of Sex Bias in the Prevalence of Attention-Deficit/Hyperactivity Disorder.	MEF2C, SEMA6D, LOC105379147, LOC102467226
Yang L	A new locus regulating MICALL2 expression was identified for association with executive inhibition in children with attention deficit hyperactivity disorder.	MICALL2, INTS1, PSMG3
van Hulzen KJE	Genetic Overlap Between Attention-Deficit/Hyperactivity Disorder and Bipolar Disorder: Evidence From Genome-wide Association Study Meta-analysis.	ADCY2, TAF9BP2, ABLIM1, CEP85L
Sun X	Genetic variant for behavioral regulation factor of executive function and its possible brain mechanism in attention deficit hyperactivity disorder.	CCDC170, ESR1
Qi G	Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits.	LOC105377030 , LOC105377031
Smoller JW	Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis.	ITIH3, DPYD, MIR137, CCDC68, CNNM2, ANK3, SYNE1, DCP1B, CACNA1C, NT5C2, AS3MT, MMP16, ODZ4, CSMD1, CACNB2, PCGEM1, TRIM26, LOC105372125, ZNRD1, LOC107985969
Lesch KP	Molecular genetics of adult ADHD: converging evidence from genome-wide association and extended pedigree linkage studies.	MOBP, GPC6
Yang L	Polygenic transmission and complex neuro developmental network for attention deficit hyperactivity disorder: genome-wide association study of both common and rare variants.	LOC100289360, NCL

3.2 Data Analysis

In order to analyse the data in an easy and reliable way, we have implemented Linux-based bash scripts that use several bioinformatic tools and statistics. In these scripts we specified all the parameters that we needed (described below) in a way that allowed us to perform batch processing of the large genomic data (see Supplementary script 1 and 2). We also implemented several R scripts [53] in order to obtain the Manhattan plots, haplotype plots, and biological pathways plots shown in the Results section.

3.2.1 Statistics

There are different analytic methods to detect signatures of selection in DNA sequences
[54]. In this thesis, we used two different approaches: the linkage disequilibrium–based methods, extended haplotype homozygosity (EHH) and integrated haplotype score (iHS), and a population differentiation–based method, the Wright's fixation index (F_{st}) approach.

Extended haplotype homozygosity (EHH) and integrated haplotype score (iHS)

Extended haplotype homozygosity (EHH) is "a probability that two haplotypes that carry the mutation of interest (usually called the core region) are identical for all other alleles within a specified genomic region surrounding the core" [28, p299]. This also means that EHH registers the transmission of an extended haplotype without recombination. "The EHH measures the decay of identity, as a function of distance, of haplotypes that carry a specified ''core'' allele at one end. For each allele, haplotype homozygosity starts at 1, and decays to 0 with increasing distance from the core site" [6]. Core haplotypes with a high EHH and a high population frequency show that there was a mutation which became significant in the human gene pool faster than expected under neutral evolution [55]. Therefore, selected alleles will have a larger area under the EHH curve than the neutral allele in plots that show EHH versus distance (Figure 3). The

integral of the observed decay of EHH away from a specified core allele, until EHH reaches 0.05, needs to be calculated and that is called iHS [6]. EHH of a sample with n chromosomes, from the locus of interest x_0 to the x_i (i-th marker either upstream or downstream from x_0) is:

$$EHH(x_i) = \frac{\sum_{h \in C(x_i)} \binom{n_h}{2}}{\binom{n}{2}} \quad (4)$$

where C is a set of all possible distinct haplotypes at x_0 and $C(x_i)$ is a set of all possible distinct haplotypes from x_0 to x_i and n_h is the number of observed haplotypes of type $h \in C(x_i)$. We can also calculate EHH of a subsample of chromosomes that carry the core haplotype c to marker x_i as:

$$EHH_c(x_i) = \frac{\sum_{h \in H_c(x_i)} \binom{n_h}{2}}{\binom{n_c}{2}} \quad (5)$$

where $H_c(x_i)$ is a partition of $C(x_i)$ that contains all distinct haplotypes carrying the core haplotype ($c \in C$) at x_0 extending to x_i, n_h is the number of observed haplotypes of type $h \in H_c(x_i)$and n_c is the number of observed haplotypes carrying the core haplotype ($c \in C$) [56].

Decay of haplotypes in a single region in which a new selected allele is sweeping to fixation can be presented with a plot shown in the Figure 4. The left side of the figures shows the decay (starting from 1 and going to 0, with increasing distance from the core site) of haplotype homozygosity. The blue line represents the ancestral allele and the red line represents the derived allele. The further the red line is from the blue line the greater the strength of the signal of selection. The right side of the figures contains the colormap plot, which shows sorted haplotypes at different distances from the selected SNP (the one with the highest absolute value of iHS) at the centre (0.0). SNP positions are marked below the plot with blue for SNPs with intermediate allele frequencies (minor allele .0.2), and red otherwise. The distance over which the haplotypes are spread is shown above the plot. The upper part of the plot shows ancestral allele (blue), and the lower part shows derived allele (red). Haplotypes that have the same colour and that are next to each other carry the same genotype between the SNP and the selected site (in the centre). The left and right side are sorted separately, and they are no longer plotted after the point they become unique. The red haplotype is sweeping to fixation. The colour changes at some

distance from the core SNP when a chromosome has a different allele compared to the chromosomes with whom it shared a common haplotype up until that point. The new colour is a new haplotype from that point on (http://haplotter.uchicago.edu/instruction.html).

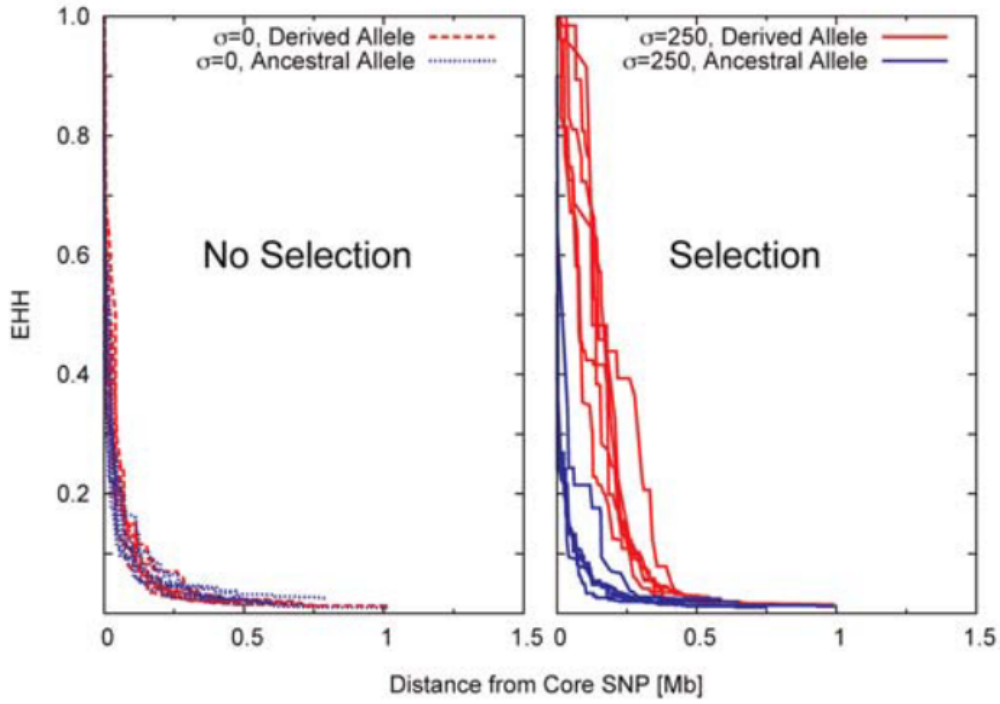

Figure 3: Decay of haplotype homozygosity for ten replicate simulations. Left side shows that the haplotype homozygosity decays similarly for ancestral and derived alleles when the core SNP is neutral. The right side shows that the haplotype homozygosity decays much slower for the derived alleles than for the ancestral alleles if the derived one is favoured [6]

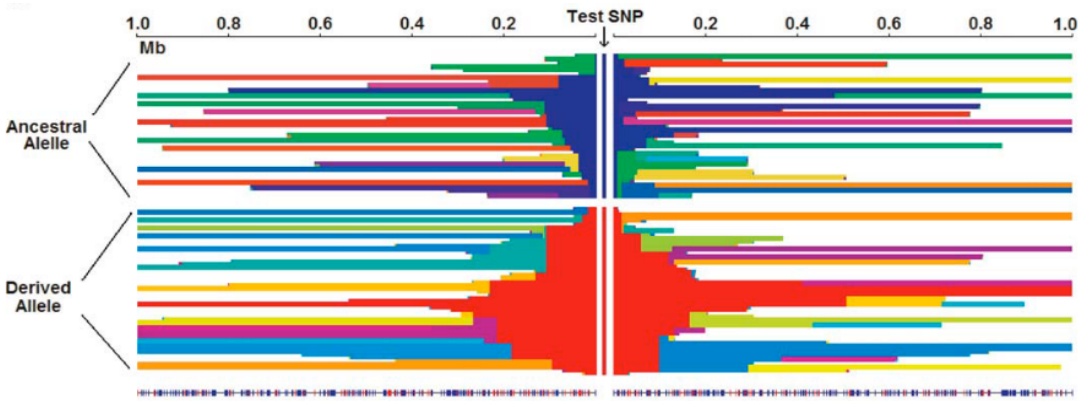

Figure 4: Decay of haplotypes in a single region [6]

Integrated haplotype score (iHS) is a statistic that is used to "compare the EHH for one allele at a locus with the EHH for the other allele at that locus (assuming a bi-allelic locus, which is usually the case for SNP data)" [28, p299]. iHS is used for detecting local selection, however, it is not as effective when the selected allele is at or near fixation because the EHH for the non-selected allele cannot be measured accurately [28, p300]. It is "designed to use phased genotypes to identify putative regions of recent or ongoing positive selection in genomes" [56] based on a model of a hard selective sweep. In this model, *de nov*o adaptive mutations arise on haplotypes and the haplotypes quickly sweep to fixation which then decreases the diversity around the locus. When selection is strong, fixation occurs quickly so recombination and mutation cannot break up the haplotype. This makes it possible for observing a signal of high haplotype homozygosity that extends from an adaptive locus [56]. The calculation is based on the equation (5) and to calculate the iHS at a locus, first we need to calculate integrated haplotype homozygosity (iHH) for the ancestral and ancestral haplotypes:

$$iHH_c = \sum_{i=1}^{|D|}[\frac{1}{2}(EHH_c(x_{i-1}) + EHH_c(x_i))g(x_{i-1}, x_i)] + \sum_{i=1}^{|U|}[\frac{1}{2}(EHH_c(x_{i-1}) + EHH_c(x_i))g(x_{i-1}, x_i)]$$

(6)

where D is the set of markers downstream from the current locus such that x_i represents the i-th closest downstream marker from the locus of interest (x_0). U and x_i represents the i-th closest upstream marker from the locus of interest (x_0), $g(x_i\text{-}1, x_i)$ is the genetics distance between the markers. Unstandardized iHS is then calculated as

$$\ln(\frac{iHH_1}{iHH_0})$$

(7)

and then they are normalized in frequency bins across the entire genome:

$$iHS = \frac{\ln(\frac{iHH_1}{iHH_0}) - E_p\left[\ln(\frac{iHH_1}{iHH_0})\right]}{SD_p\left[\ln(\frac{iHH_1}{iHH_0})\right]}$$

(8)

where Ep is expectation and SD_p is standard deviation in frequency bin p [56].

Wright's fixation index

F_{st} (Wright's fixation index) is a statistic that "compares the variance of allele frequencies within and between populations." The F_{st} values range from 0 (no differentiation) to 1 (complete differentiation, for example in population 1 the allele A is fixed and in population 2 the allele a is fixed). If the F_{st} value is large at certain genomic regions (usually > 0.25) (i.e., relative to neutral regions), it means that there is a greater differentiation between population, and it may indicate directional selection. Small values, on the other hand, mean that the populations are similar to each other which can indicate a balancing or directional selection (or no selection at all; simply a demographic event). It can also detect other types of selection, like classic sweeps, sweeps on standing variants, and negative selection [57]. It is important for association mapping of human disease genes and "the same evolutionary process increases differentiation among population also increase the similarity among individuals within the population" [2, p147]. When allele frequencies are compared between cases and controls, F_{st} can be used to check that the differences are not just what is expected by chance [58]. The values of F_{st} are between 0 and 1 as shown in Table 4 [2, p147].

F_{ST} values	Level of genetic differentiation
Less than 0.05	little
Between 0.05 and 0.15	moderate
Between 0.15 and 0.25	great
Greater than 0.25	very great

Table 4: Sewall Wright's qualitative guidelines for interpreting Fst

We have used the Weir & Cockerham F_{st} implemented in PLINK. The general formulae are:

$$\hat{\theta} = \frac{a}{a+b+c}$$

$$a = \frac{\bar{n}}{n_c}\left\{s^2 - \frac{1}{\bar{n}-1}\left[\bar{p}(1-\bar{p}) - \frac{r-1}{r}s^2 - \frac{1}{4}\bar{h}\right]\right\}$$

$$b = \frac{\bar{n}}{(\bar{n}-1)}\left[\bar{p}(1-\bar{p}) - \frac{r-1}{r}s^2 - \frac{2\bar{n}-1}{4\bar{n}}\bar{h}\right]$$

$$c = \frac{1}{2}\bar{h}$$

(9)

where $\bar{n}$the average sample size, n^2 with C^2 is the squared coefficient of variation of sample sizes, with $\bar{p}$the average sample frequency of allele A, s^2 is the sample variance of allele A frequencies over populations, r is the number of populations of the same size for which it is assumed they have descended separately from a single ancestral population that was in both Hardy-Weinberg equilibrium and linkage equilibrium, and $\bar{h}$ is the average heterozygote frequency for allele A.

3.2.2 Bioinformatics

We used the software programs PLINK 1.9 (www.cog-genomics.org/plink/1.9/) and PLINK 2 (www.cog-genomics.org/plink/2.0/) [59] and VCFtool v0.1.14 (https://vcftools.github.io/) [60] to process variant call format (VCF) files from the 1000 Genome Project database for all autosomal chromosomes. We excluded all structural variants and restricted our analysis to bi-allelic SNPs with minor allele frequency (MAF) > 0.05 and quality value (minQ) above 30. We used Variant Effect Predictor (https://www.ensembl.org/Tools/VEP; human assembly GRCh37.p13) [61] in order to retrieve the genes connected to the SNPs, location of the variants, consequence of the SNPs in the protein sequence (e.g. stop gained, missense, stop lost, frameshift), symbol and biotype (protein coding, pseudogene, processed pseudogene, antisense, miRNA, rRNA, scRNA, snoRNA, snRNA). The UCSC Genome Browser (http://genome.ucsc.edu/) was used to retrieve the genomic position of the thiamine transporter genes (including 5kbp up- and downstream of the gene) in accordance with the reference genome GRCh37/hg19.

iHS was calculated by using the software programme selscan version v1.2.0a [56] for every SNP from the 1000 genomes data and it measures the strength of selection that acts on or near every SNP. In order to calculate the iHS we first transposed the phased haplotype data (hapfiles) for each chromosome and population (with one row per haplotype, and one column per variant). The subsequent running parameter for the selscan programme were --ihs --hap <haplotype files> --map <mapfiles> --out <result file> The unstandardized iHS scores were normalized in frequency bins across the entire genome using the script *norm*, provided with the selscan programme We considered a SNP to have a candidate selection signal if it had -log10(p-value) ≥ 2.0 and was within a gene that had at least 20 SNPs with elevated iHS scores. We then manually looked in the files to see if these SNPs appear in a ‘cluster’ of at least 10 SNPs with values close or above |2|. The

reasoning behind this is that selective sweeps produce clusters of extreme iHS scores across the sweep region and even though large negative values of iHS are what is assumed to show that a derived allele has reached significant frequency, positive values can also indicate a sweep when ancestral alleles hitchhike with the selected site. It is also possible that selection favours an ancestral allele which was segregating in the population [6].

Genes identified in GWAS were compared to the genes that were identified as being under selection in the previous step and the overlap represents the genes that are connected to the disorders and under selection.

F_{st} was calculated with VCFtool using the parameter *–weir-fst-pop* and text files containing list of individuals (one per line) belonging to each population. We chose one population per continental group: GBR from European, LWK from African, PJL from South Asian, and JPT from East Asian, and calculated F_{st} between each two of these population for every chromosome.

ReactomePA library [62] for R scripting was used to perform enrichment analysis and find biological pathways of the genes that were found to be under positive selection and connected to disorders in order to gain novel insights into the molecular mechanism involved. This was done separately for every continental group (i.e., Africa, Europe, South Asia, and East Asia) and disorder.

GeneCards – the human gene database (www.genecards.org, accessed on the 29th of August 2019.) was used to obtain gene functions and the 1000 Genomes Project Phase 3 (www.ensembl.org, accessed on the 29th of August 2019.) was used to obtain the frequencies of the alleles that show positive selection signals.

For creating Manhattan plots, we used R package qqman [63] and from it a function manhattan() that takes a data frame with columns containing the chromosome number, chromosomal position, p-value, and the SNP rs number. We specified the SNPs that we wanted to highlight, in this case the ones that were found to be under positive selection. These plots show the statistics on the y-axis, chromosomal position of the SNP on the x-axis, and then normalized absolute iHS value. If a region has many highly associated SNPs in LD they will appear as “skyscrapers” [63]. We generated for each disorder a Manhattan plot containing the lead risk SNP loci of the GBR population. The risk loci which is under positive selection (i.e., smaller -2 or higher than 2) is coloured in green and the name is indicated with an arrow.

EHH plots for the core SNPs were created using selscan [56] with the parameters –ihh, --hap, and minor allele frequency --maf above 0.025.

4. RESULTS

4.1 Genes under Natural Selection

In this thesis, we highlight in particular the genes under positive selection in Europeans. Thus, we describe here the risk loci, their functions and allele frequencies that show evidence of positive selective in European populations GBR, FIN and TSI. The results for the other population can be found in the supplementary files.

4.1.1 Results for MDD

Out of 215 genetic loci that have been associated with major depressive disorder in the GWAS, 15 genes were found to be under positive selection in the African, seven in the European, 11 in the East Asian, and 13 in the South Asian populations. The following two genes were found to be under positive selection in all studied populations: major histocompatibility complex class I B (*MHC class I B* or also named as (human leukocyte antigen) *HLA-B*) and, RNA binding fox-1 homolog 1(*RBFOX1*). Table 5 presents the MDD risk genetic loci that were found to be under positive selection in the populations of European genetic ancestry (GBR, FIN and TSI). The SNPs found to be under positive selection are mainly located in the intron sequences. Figure 4. shows the Manhattan plot of iHS values in GBR population with SNPs associated with MDD that are under positive selection (marked with green circles).

The gene *CACNA1C* has a direct impact on the brain and important neuronal activities. The other genes are connected to important regulatory processes: *RBFOX1* is known to be evolutionary conserved as it regulates tissue-specific alternative splicing in metazoa, *FHIT* encodes a protein involved in purine metabolism, *RFX3* is a member of the regulatory factor X gene family, which encodes transcription factors that contain a highly-conserved winged helix DNA binding domain, spectrin repeat containing nuclear envelope protein 2 (*SYNE2*) encodes a nuclear outer membrane protein that binds cytoplasmic F-actin. DENN domain containing 1a (*DENND1A*) has a function in the internalization of proteins and lipids.

The F_{st} analyses revealed that some loci are 'outliers' with high n (> 0.25) F_{st}-values; for example, *RFX3* risk loci has a high F_{st} value, a high iHS value, and its risk allele T of the SNP rs6476437, which is the ancestral allele, occurs in higher frequency in Africans (0.78) compared to other continental groups: 0.17 (EUR), 0.21 (SAS), and 0.22 (EAS). The risk allele T of the SNP rs669131 within the intron 1 of the *DENND1A* gene has a high frequency of 0.78 in Africans, while in other continental groups it has much lower frequencies: 0.14 (EUR), 0.17 (SAS), and 0.08 (EAS). The *HLA-B* risk loci has the highest iHS value at 4.98234 for the SNP rs147324178 and this gene belongs to the HLA class of genes which are known to be under positive selection because they play an important role in the immune system and is expressed in nearly all cells [64].

Table 5: MDD variants that localize to regions with evidence of positive selection in Europeans. SNP column indicates the variant with the highest absolute normalized iHS value. F_{st} is calculated between GBR and LWK. Summaries for the genes were retrieved from genecards.org.

Risk Region	SNP	Position	Risk allele	Summary for the gene	Risk Loci	F_{st}	iHS
12p13.33	rs7295250	2776943	C	Encodes an alpha-1 subunit of a voltage-dependent calcium channel	*CACNA1C*	0.254375	2.86138
9q33.3	rs669131	126680489	T	Member of the connecdenn family, and as such functions as a guanine nucleotide exchange factor (GEFs) for the early endosomal small GTPase RAB35 (MIM 604199) and binds to clathrin and clathrin adaptor protein-2 (AP2; see MIM 601024). Thus, connecdenns link RAB35 activation with the clathrin machinery	*DENND1A*	0.47833	3.22609
3p14.2	rs12492742	61184224	G	This gene encompasses the common fragile site FRA3B on chromosome 3, where carcinogen-induced damage can lead to translocations and aberrant transcripts.	*FHIT*	-0.00345584	2.92529
6p21.33	rs147324178	31324742	A	Plays a role in the immune system by presenting peptides derived from the endoplasmic reticulum lumen and is expressed in nearly all cells	*HLA-B*	-0.0022935	4.98234

16p13.3	rs8046462	7725750	A	Evolutionarily conserved, and regulates tissue-specific alternative splicing in metazoa	*RBFOX1*	0.112663	-3.08489
9p24.2	rs6476437	3421997	T	A member of the regulatory factor X gene family, which encodes transcription factors that contain a highly conserved winged helix DNA binding domain	*RFX3*	0.57536	3.87486
14q23.2	rs7158401	64536713	A	The protein encoded by this gene is a nuclear outer membrane protein that binds cytoplasmic F-actin. This binding tethers the nucleus to the cytoskeleton and aids in the maintenance of the structural integrity of the nucleus	*SYNE2*	0.474776	2.75

4.1.2 Results for BPD

Out of 92 genes that have been associated with BPD in the GWAS, nine genes were found to be under positive selection in the African, four in the European, seven in the East Asian, and five in the South Asian populations. Two genes were found to be shared between the populations: *HLA-B*, and *PDE10*. Table 6 presents the risk genetic loci that were found to be under positive selection in the populations of European genetic ancestry (GBR, FIN and TSI). Figure 5. shows the Manhattan plot of iHS values in GBR population with SNPs associated with BPD and under positive selection marked with green circles.

Positive selection acts on these loci mainly on the intron sequences. SNP rs9457080 of the *PDE10A* gene had one of the highest iHS value (3.64532), and it has a frequency on the T allele in African populations of 0.78, while other populations (EUR, SAS, EAS) have around 0.40.

Table 6: BPD Variants that localize to regions with evidence of positive selection in Europeans. SNP column indicates the variant with the highest absolute normalized iHS value. F_{st} is calculated between GBR and LWK. Summaries for the genes were retrieved from genecards.org.

Risk Region	Lead SNP	Position	Risk allele	Summary for the gene	Risk loci	F_{st}	iHS
12p13.33	rs7295250	2776943	C	Encodes an alpha-1 subunit of a voltage-dependent calcium channel	*CACNA1C*	0.254375	2.86138
2q31.2	rs28812195	180064018	G	Protein coding gene. Gene Ontology (GO) annotations related to this gene include phosphatidylinositol-4,5-bisphosphate binding and phosphatidylserine binding	*SESTD1*	0.0274869	3.60537
6p21.33	rs147324178	31324742	A	Plays a role in the immune system by presenting peptides derived from the endoplasmic reticulum lumen and is expressed in nearly all cells	*HLA-B*	-0.0022935	4.98234
6q27	rs9457080	165783173	T	The protein encoded by this gene belongs to the cyclic nucleotide phosphodiesterase family. It plays a role in signal transduction by regulating the intracellular concentration of cyclic nucleotides.	*PDE10A*	0.19691	3.64532

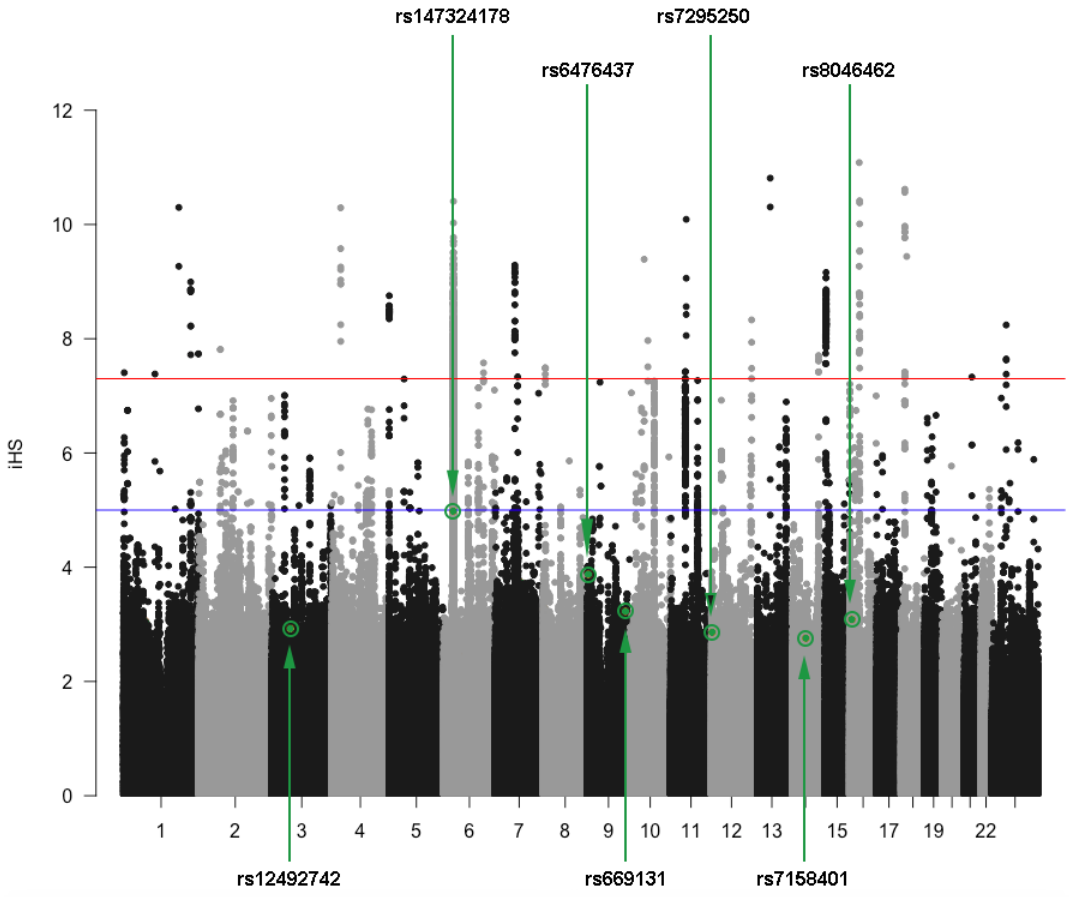

Figure 4: Manhattan plot of iHS values in GBR population with marked SNPs associated with MDD. Chromosome number is shown on the x- axis and the y-axis shows the absolute normalized iHS values for all the SNPs connected to MDD. The ones that have evidence of positive selection are coloured green with the name of the SNP indicated with a green arrow. The red line represents the genome-wide significance ($P \leq 5 \times 10{-}8$) and the blue line represents suggestive significance ($P \leq 1 \times 10{-}7$).

4.1.3 Results for ADHD

Out of 34 genes that have been associated with AD(H)D in the GWAS, five genes were found to be under selection in the African populations, two in the European, two in the East Asian, and three in the South Asian populations. Table 7 presents the risk genetic loci that were found to be under positive selection in the populations of European genetic ancestry (GBR, FIN and TSI). Figure 6. shows the Manhattan plot of iHS values in GBR population with SNPs associated with MDD and under positive selection marked with green circles.

As it was the case for MDD and BPD, positive selection acts on these loci mainly on the intron sequences. Risk loci *CCDC170* has high F_{ST}-value as well as high iHS value,

and the risk allele G of the SNP rs9383910 has a frequency of 0.88 in African populations and only 0.33 in European populations. However, its biological function is yet unknown.

Table 7: ADD Variants that localize to regions with evidence of positive selection in Europeans. SNP column indicates the variant with the highest absolute normalized iHS value. F_{st} is calculated between GBR and LWK. Summaries for the genes were retrieved from genecards.org.

Risk Region	Lead SNP	Position	Risk allele	Summary for the gene	Risk loci	F_{st}	iHS
12p13.33	rs7295250	2776943	C	Encodes an alpha-1 subunit of a voltage-dependent calcium channel	*CACNA1C*	0.254375	2.86138
6q25.1	rs9383910	151820799	G	The function of this gene and its encoded protein is not known	*CCDC170*	0.590223	3.72079

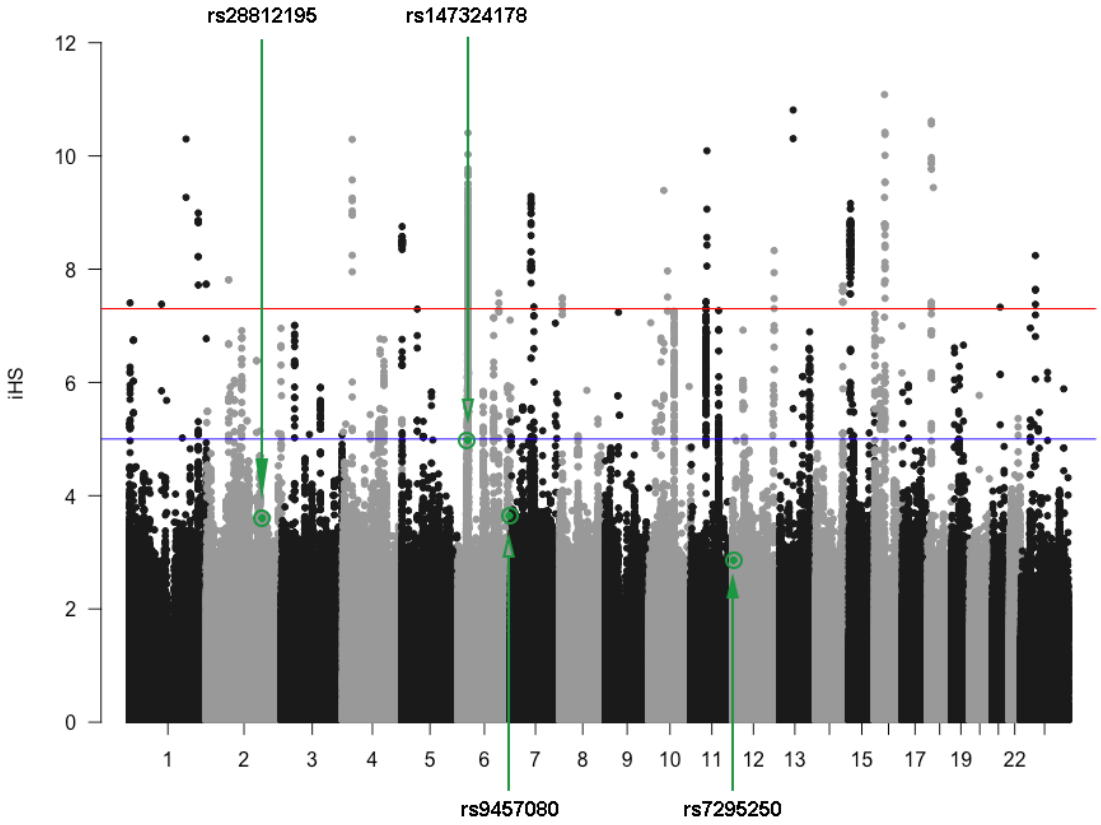

Figure 5: Manhattan plot of iHS values in GBR population with marked SNPs associated with BPD. Chromosome number is shown on the x- axis and the y-axis shows the absolute normalized iHS values for all the SNPs connected to BPD. The ones that have evidence of positive selection are coloured green with the name of the SNP indicated with a green arrow. The red line represents the genome-wide significance ($P \leq 5 \times 10{-}8$) and the blue line represents suggestive significance ($P \leq 1 \times 10{-}7$).

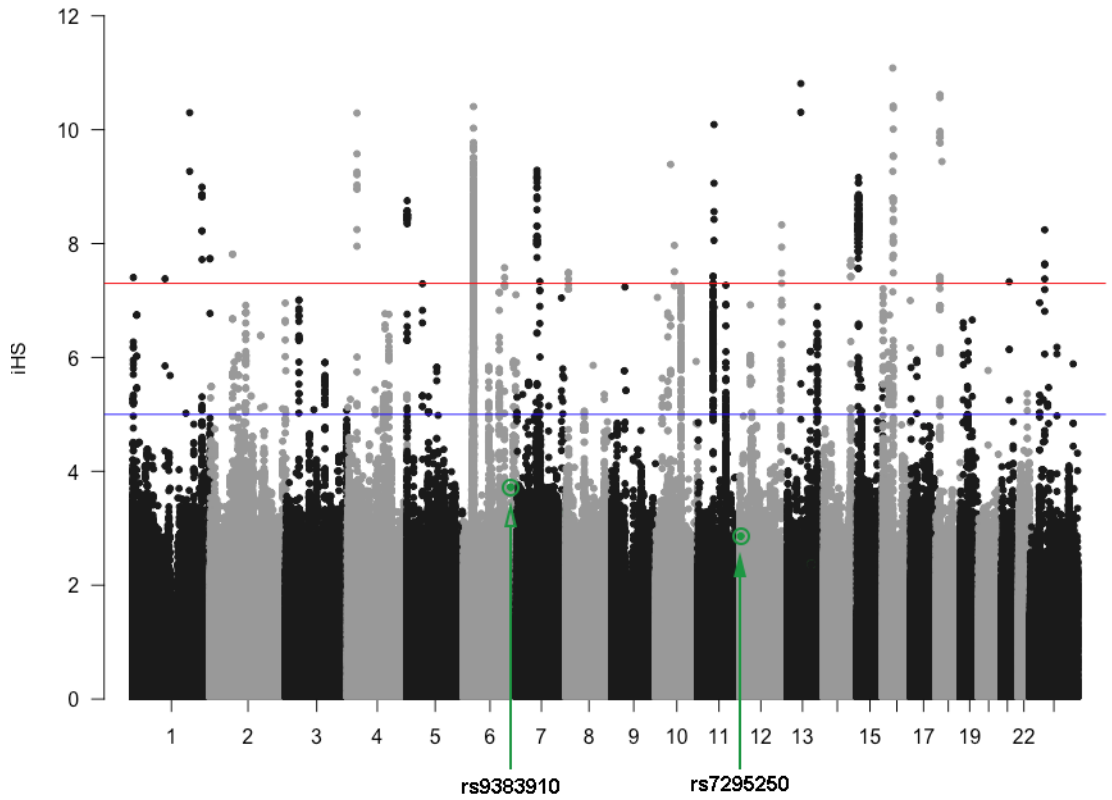

Figure 6: Manhattan plot of iHS values in GBR population with marked SNPs associated with AD(H)D. Chromosome number is shown on the x- axis and the y-axis shows the absolute normalized iHS values for all the SNPs connected to ADD. The ones that have evidence of positive selection are coloured green with the name of the SNP indicated with a green arrow. The red line represents the genome-wide significance ($P \leq 5 \times 10{-}8$) and the blue line represents suggestive significance ($P \leq 1 \times 10{-}7$).

4.1.4 Overlap of risk genes between disorders

Table 8 shows the overlap of risk loci under positive selection between the disorders per continental group. Most of the genes connected with AD(H)D are also connected with BPD, and BPD shares a few genes with ADD and MDD. All three disorders have genes associated uniquely to them, with MDD having the highest number of genes associated.

Table 8: Overlap of risk loci found under positive selection between disorders per continental group (a. European populations, b. East Asian populations, c. South Asian populations, d. African populations). Red colour represents loci shared between all disorders, blue colour loci shared between MDD and BPD, and yellow colour loci shared between AD(H)D and BPD.

a. European populations		
MDD	BPD	AD(H)D
CACNA1C	*CACNA1C*	*CACNA1C*
HLA-B	*HLA-B*	
DENND1A	*PDE10A*	*CCDC170*
FHIT	*SESTD1*	
RBFOX1		
RFX3		
SYNE2		

c. East Asian populations		
MDD	BPD	AD(H)D
CSMD1	*CSMD1*	*CSMD1*
	CEP85L	*CEP85L*
HCG4	*HCG4*	
HLA-B	*HLA-B*	
DENND1A	*IPO8*	
FHIT	*MYO1H*	
HCG9	*PDE10A*	
KCNQ5		
RBFOX1		
SNTG1		
SORCS3		
XKR6		

c. South Asian populations		
MDD	BPD	AD(H)D
CACNA1C	*CACNA1C*	*CACNA1C*
CSMD1	*CSMD1*	*CSMD1*
	CEP85L	*CEP85L*
HLA-B	*HLA-B*	
DENND1A	*PDE10A*	
GRM5		
KCNQ5		
RBFOX1		
RFX3		
SLC30A9		
THSD7A		
TYR		
XKR6		
ZNF165		

b. African populations		
MDD	BPD	AD(H)D
ANK3	*ANK3*	*ANK3*
AS3MT	*AS3MT*	*AS3MT*
CSMD1	*CSMD1*	*CSMD1*
DPYD	*DPYD*	*DPYD*
HLA-B	*HLA-B*	
LINC00461	*LINC00461*	
ASTN2	*PDE10A*	*GPC6*
DCC	*SHISA9*	
FHIT	*TCF7L2*	
KSR		
MMP16		
NBAS		
PCDH9		
RBFOX1		
SHISA9		
	ALL	
	MDD+BPD	
	AD(H)D+BPD	

4.2 EHH plots

Figures 7-11 show signals of positive selection for the risk loci that were studied in European populations. These EHH plots show the relative lengths of the haplotypes around the derived core SNP under selection.

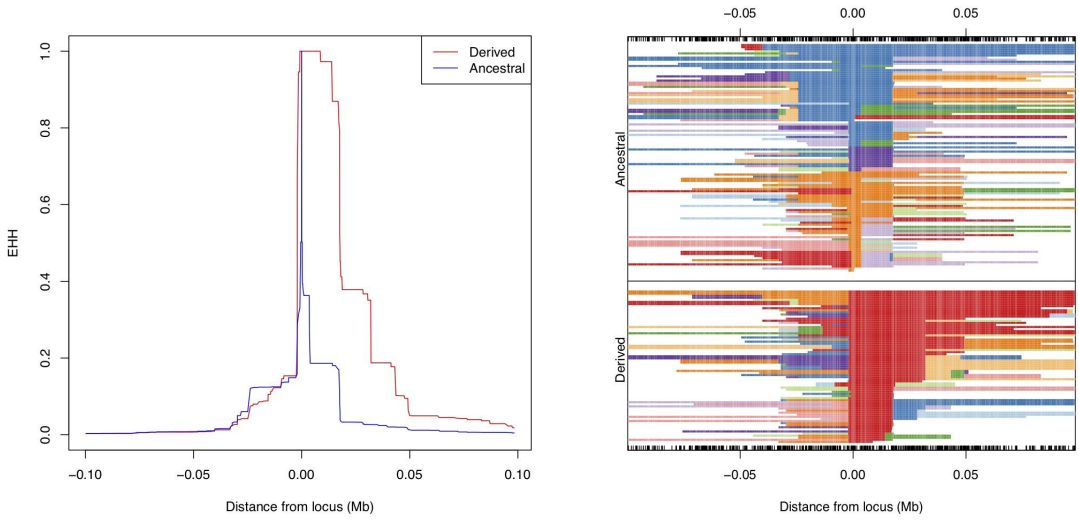

Figure 7: Signals of selection for *CACNA1C*. Left: decay of haplotype homozygosity. Right: haplotype plots

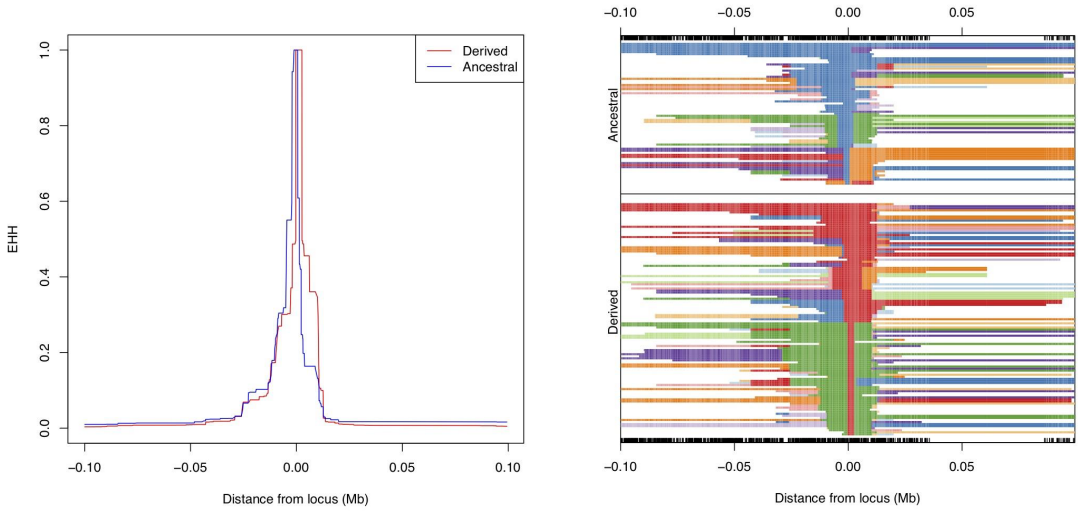

Figure 8: Signals of selection for *RBFOX1*. Left: decay of haplotype homozygosity. Right: haplotype plots

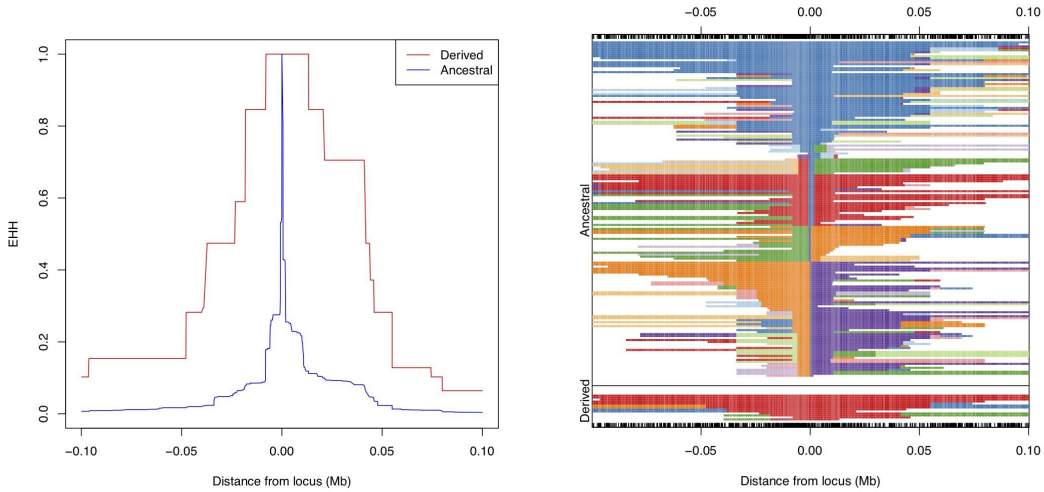

Figure 9: Signals of selection for *RFX3*. Left: decay of haplotype homozygosity. Right: haplotype plots

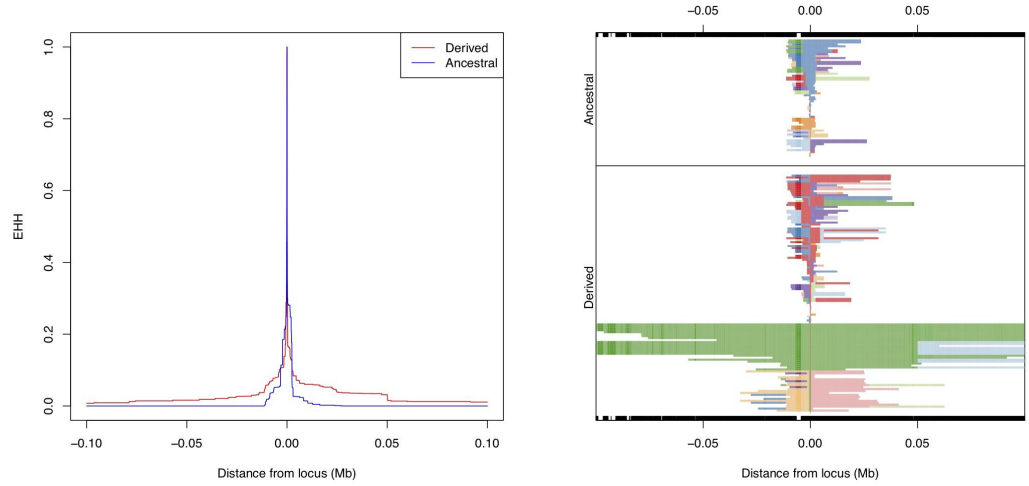

Figure 10: Signals of selection for *HLA-B*. Left: decay of haplotype homozygosity. Right: haplotype plots

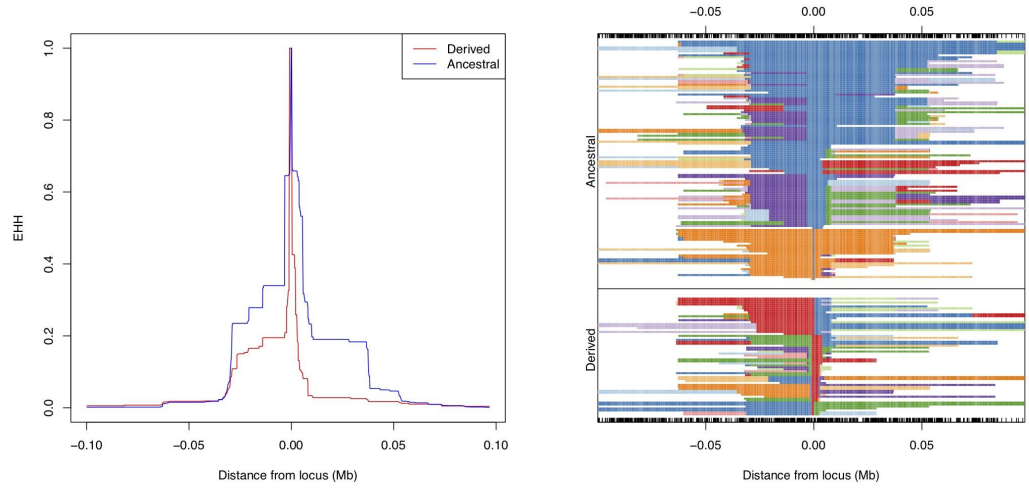

Figure 11: Signals of selection for *CCDC170*. Left: decay of haplotype homozygosity. Right: haplotype plots

4.3 Biological pathways

4.3.1 Results for MDD

The genes with evidence of positive selection in MDD are enriched for biological processes with a significant value $p<0.05$ only for East Asian populations (Figure 12). The detected pathways are related to immunity (MHC class I antigen generation (endosomal/vacuolar pathway) and cross-presentation as well as interferon signaling), cellular potassium channel functions and G-proteins binding pathways (RAB GEFs exchange GTP for GDP on RABs pathway). The potassium channels are found in all cell types and they are important for controlling the resting membrane potential in neurons. They also contribute to the regulation of action potentials in cardiac muscle and help the release of insulin from pancreatic beta cells. The voltage-gated potassium channels play a fundamental role in the excitability of heart, brain, and skeletal muscle cells.

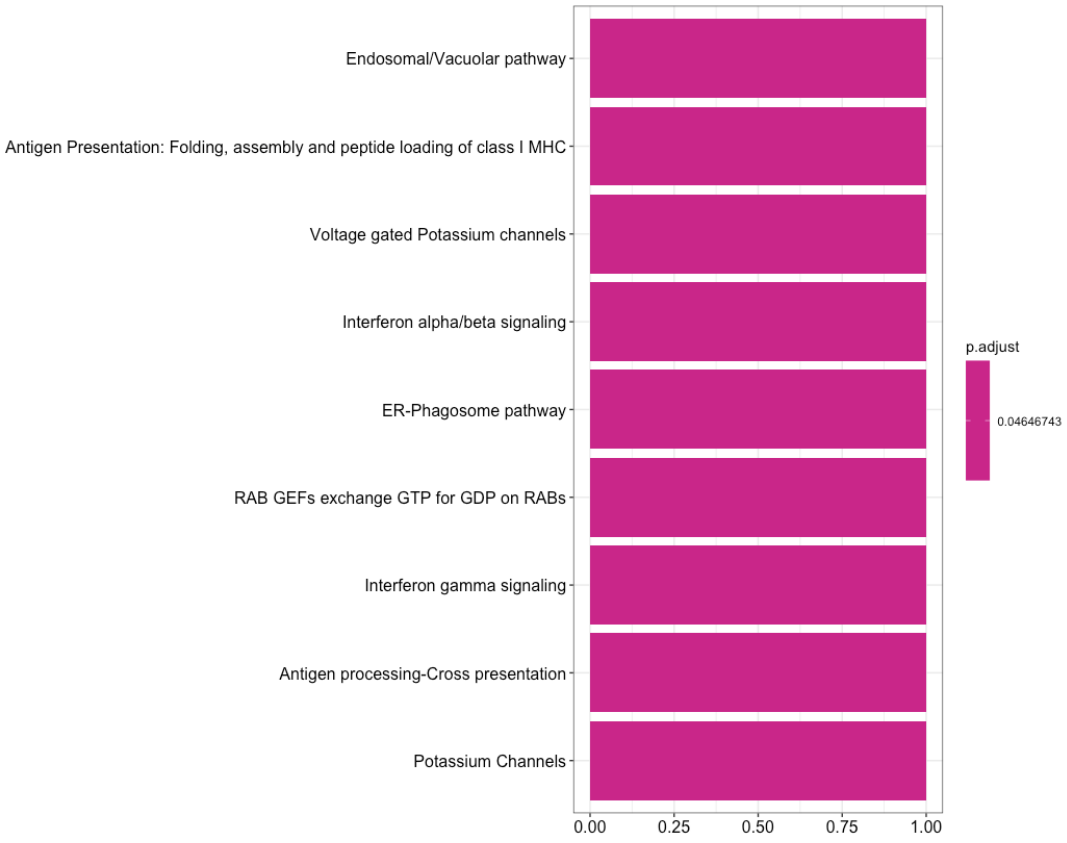

Figure 12: Pathways for East Asian population in MDD

4.3.2 Results for BPD

The genes with evidence of positive selection are enriched for biological processes with a significant value $p<0.05$ for East Asian, European, and African populations (Figure 13a and 13b).

Similar to the MDD disorder, the immunological pathways related to class I HLA function were detected as significant pathways also in the BPD disorder. Furthermore, the nitric oxide/cGMP signal transduction and platelet homeostasis pathways appear to be important in this disorder. In particular, signal transduction with nitric oxide (NO) is important for neurotransmission which facilitates functional interaction between neurons and other cell types [65]. In the South Asian and European populations in addition to this, the pathway NCAM signalling for neurite out-growth, which is involved in a variety of cellular processes important for the formation and maintenance of the nervous system, appears as significant for BPD.

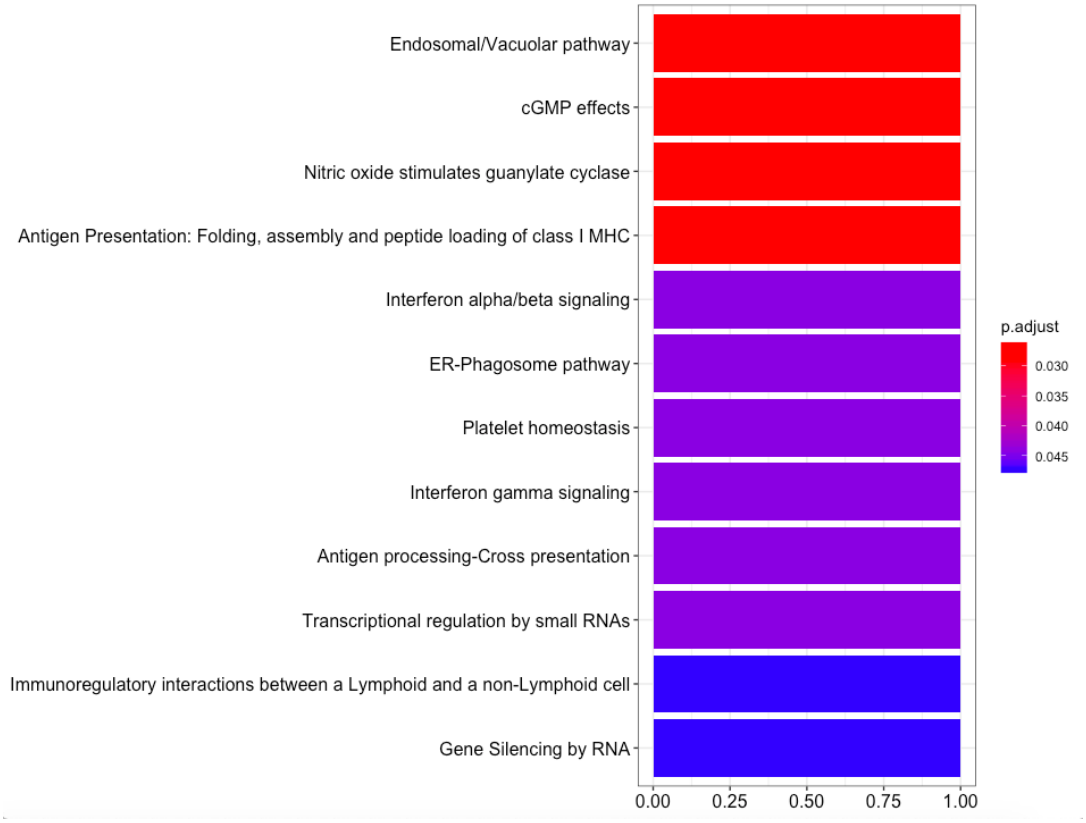

Figure 13a: Pathways for East Asian populations in BPD

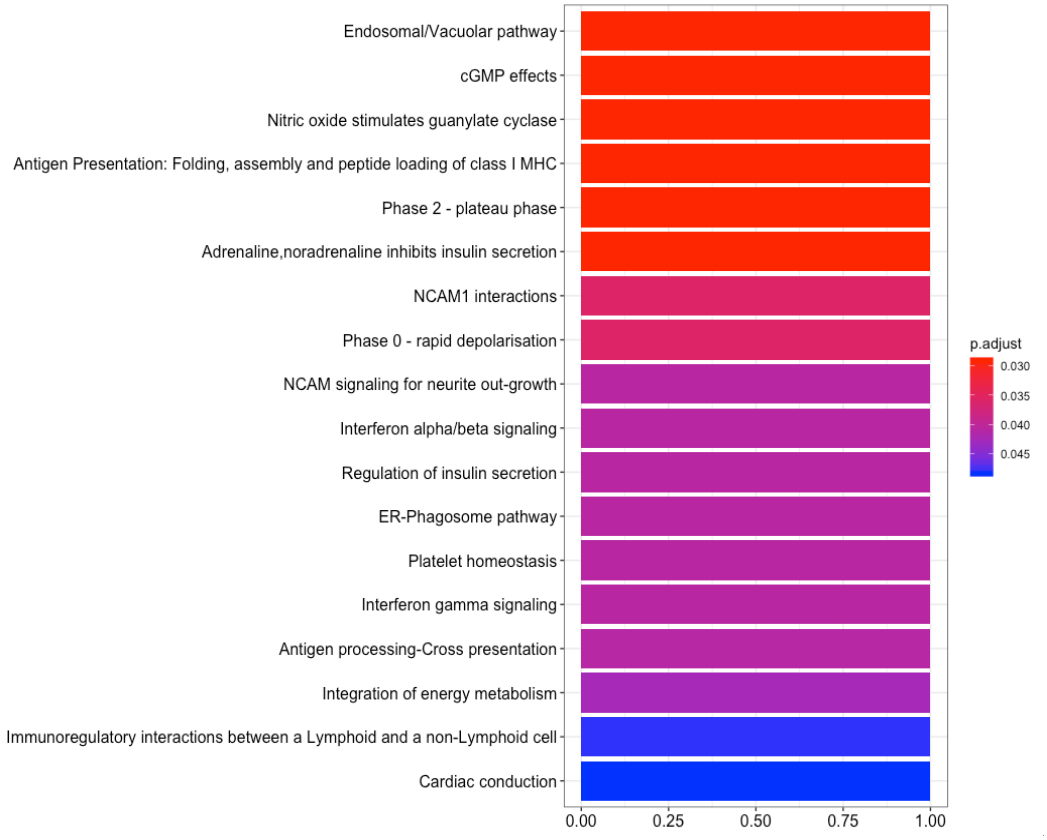

Figure 13b: Pathways for European and South Asian populations in BPD

4.3.3 Results for ADHD

The genes with evidence of positive selection are enriched for biological processes with a significant value $p<0.05$ for South Asian, European, and African populations (Figure 14a and 14b).

In European and South Asian populations pathway with evidence of positive selection NCAM signalling for neurite out-growth, is connected to nervous system because it has importance in cellular processes important for its formation and maintenance, and NCAM1 interactions is connected to the neural cell adhesion molecule NCAM1 which is considered to be a cell adhesion mediator and a signal-transducing receptor molecule. Pathways adrenaline, noradrenaline inhibits insulin secretion, regulation of insulin secretion, and integration of energy metabolism are involved in different physiological processes such as regulation of appetite, absorption, transport, and oxidation of foodstuffs influence energy metabolism pathways. Pathways cardiac conduction and muscle contraction are involved in regular heart and muscles functioning.

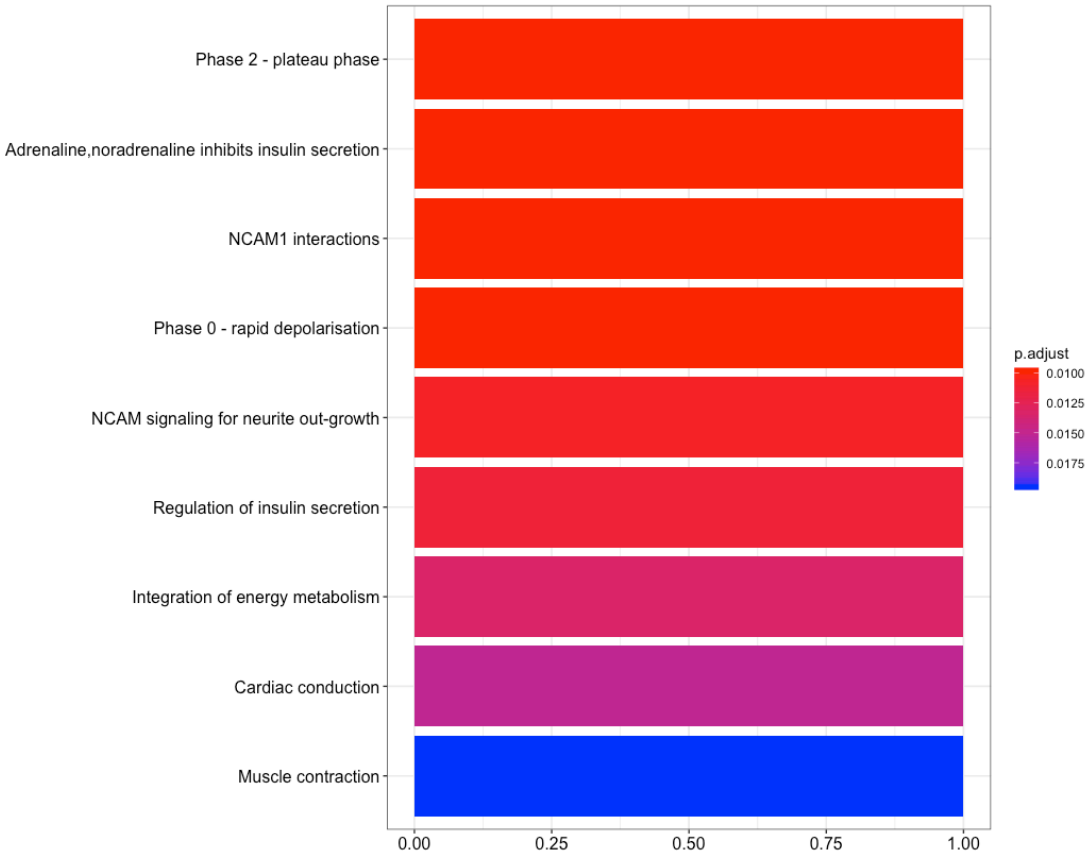

Figure 14a: Pathways for European and South Asian populations in ADHD

In African populations, a lot of pathways, such as defective EXT2 causes exostoses 2, defective EXT1 causes exostoses 1, TRPS2 and CHDS, defective B4GALT7 causes EDS, progeroid type, defective B3GAT3 causes JDSSDHD, defective B3GALT6 causes EDSP2 and SEMDJL1, have a negative effect on organisms by causing different disorders and cancers. Retinoid metabolism and transport are involved in the processing of vitamin A which is important for growth, development and maintenance of the immune system and good vision.

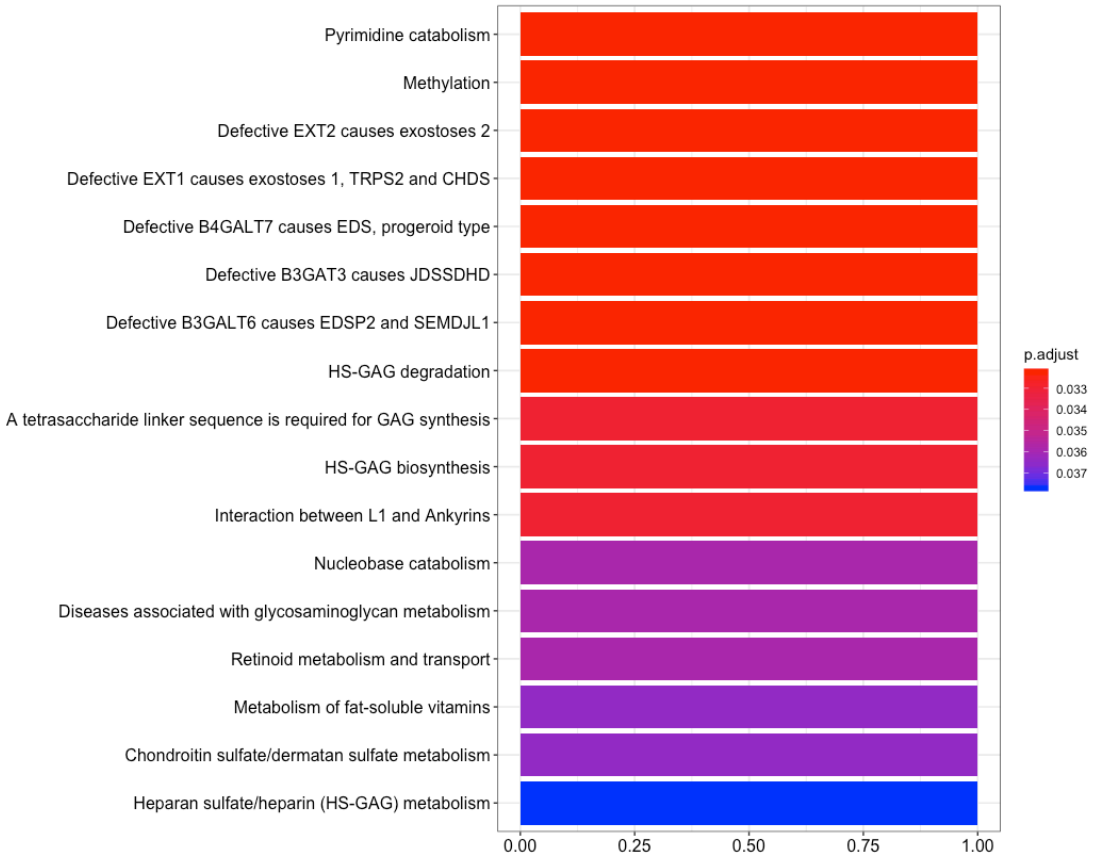

Figure 14b: Pathways for African populations in ADHD

5. DISCUSSION

In this study, we found evidence for positive selection acting on genes associated with different psychiatric disorders in all studied human populations (African, European, South Asian, and East Asian). Some of the risk loci under positive selection are shared between the populations, but the majority are not which implies that there are differences in the selection pressures in different environments, as expected. The results also seem to imply that recent positive selection has influenced the genetic variations underlying the disorders we analysed. However, it is still a challenge to quantify the influence these risk loci have on the disorders as their contribution to disorder's susceptibility is not yet fully understood.

5.1 Loci under positive selection

Our results show signals of positive selection on *HLA-B* in major depressive disorder and bipolar disorder in all continental groups. This gene belongs to the human leucocyte antigens (HLA) which have a key role in the recognition and presentation of antigens to the immune system. HLA genes belong to the most variable genes in the human genome and are known to be under some form of pathogen-driven balancing selection, with *HLA-B* potentially being under higher selective pressure from pathogens compared to *HLA-A* and *HLA-C* [66]. The HLA genes have been associated with psychiatric disorders such as schizophrenia and autism spectrum disorder, and a connection between these disorders to inflammation and the immune system has been established [67]. On one hand, there have been studies that showed that prenatal exposure to some infectious diseases can be a risk factor for the development of psychiatric disorders and that some neuroinflammation and other immune reactions can trigger the symptoms [67, 68, 69]. On the other hand, HLA have non-immune effect as they modulate synaptic pruning by microglia and other mechanisms which means that they can directly affect this physiological process [68]. Bipolar disorder has been associated with immune dysfunctions, inflammation, and maternal or fetal infections, and major depressive disorder to inflammations,

however further research is needed to explain this connection [69, 70, 71, 72]. Several studies suggest that there might be an evolutionary link between the immune and nervous system as they share important similarities such as sensing and reacting to the internal and external factors, intercellular communication by forming synapses, and plasticity [73, 74]. Another evolutionary rationale that connects these two systems is related to the so-called "sickness behaviour" when a person who is sick withdraws from social interactions to prevent the spreading of the sickness. It is believed that this behaviour was important in the past to prevent spreading of infections as then it was common to die from them [70, 73].

We found the gene *RBFOX1* to be under positive selection in the major depressive disorder for all continental groups. This gene has previously been implicated as a candidate gene for autism spectrum disorder because it has a significant role in neuronal migration and synapse network formation during corticogenesis [75, 76]. *PDE10A*, a gene that regulates dopaminergic neurotransmission, was found to be under positive selection in all continental groups for bipolar disorder, and this gene was previously associated with schizophrenia [77].

A recent GWAS of educational attainment (number of years of schooling) in 1.1 million European-descent individuals identified 1 271 independent genome-wide-significant SNPs [78], which implicate genes connected to brain-development processes and neuron-to-neuron communication. Additionally, this study investigated three phenotypes that seem to be correlated with educational attainment: cognitive (test) performance, self-reported math ability, and hardest math class completed. Several of the genes that we found to be under positive selection and associated with the different psychiatric disorders we investigated, have been found to also be connected to self-reported educational attainment (*CSMD1*, *DENND1A*, *SYNE2*, *CACNA1C*, *RBFOX1*, *FHIT*) and mathematical ability (*CSMD1*, *DENND1A*, *CACNA1C*, *RBFOX1*). *RBFOX1* was connected to cognitive function measurement in the same study. Other studies have demonstrated that intelligence has been connected to the genes: *KCNQ5* [79], *RBFOX1* [79, 80, 81], *PDE10A* [82].

5.2 Biological pathways

We have identified pathways shared by some of the disorders but none common for all three disorders. However, all of the disorders are associated with biological pathways that are related to the immune system, and, unsurprisingly, the nervous system, as reported in detail in section 4.2. Interestingly, all pathways but one found to be significant in ADD for European populations are also found in BPD of European populations, and as described in section 2.4 these two disorders have several symptoms in common and can often co-occur. MDD in East Asian populations and BPD in European, South Asian, and East Asian populations show overlap in five pathways: endosomal/vacuolar pathway, antigen presentation folding assembly and peptide loading of class I MHC, interferon alpha/beta signalling, ER phagosome, interferon gamma signalling, antigen processing-cross presentation. Other evidence contributing to the significance of the ER phagosome pathway were pointed out in a different study which identified X-box binding protein 1 (*XBP1)* as a risk gene in bipolar disorder. This gene has been shown to be a very important gene in the endoplasmic reticulum (ER) stress response [96].

5.3 Evolutionary perspective

The different evolutionary forces that acted on the human genomes in the ancestral human environment have selected a suite of traits that were necessary for survival and reproduction, and that persists in contemporary human populations. However, the selected traits, in many cases, may no longer reflect the needs of most contemporary humans. There could be two explanations for this: either the environment is changing much faster than the genome can adapt to it, or these traits emerge as the genome interacts with the new environments, much different than the ones before [84]. Hence, a mismatch between genes and the environment occurs and it can lead to (novel) diseases and disorders [85, 86, 87]. In order to explain this mismatch, it is important to understand how evolution has shaped our phenotypic traits and which genes are important contributors to the development of diseases and disorders. We found in our study that several of the genes under

positive selection (and associated with some of the studied psychiatric disorders) are also linked to educational attainment and mathematical ability. We suggest that educational attainment itself could act as a selective force driving the variability of these genes. Furthermore, we speculate that these positively selected genes become detrimental in individuals who are exposed to certain circumstances such as highly stressful environments, different life events, exposure to certain infectious diseases or drugs, leading subsequently to the susceptibility of developing certain psychiatric disorders.

Other studies proposed different evolutionary hypotheses. In the case of ADHD, it has been considered that the traits that are seen as disadvantageous at present, were advantageous in ancestral settings. For example, hunters had to be aware of all the signals from the environment during hunting, react to them quickly, and in a creative way [88]. Additionally, ADHD has also been connected to language problems [89] and specifically, inner speech is thought to be the cause of a person not being able to self-regulate their attention [88]. On the other hand, taking risks, and novelty-seeking behaviour, that are part of ADHD, can be seen as positive even in contemporary societies [90]. BPD has been connected to some of the characteristics of urban life such as noise and light pollution which are causing decupling between our habits and adaptive demands. From the evolutionary perspective, light and noise could cause a stress reaction that triggers the need to use mental and physical resources [91]. As already mentioned, depression can be organism's response to inflammations [70, 73] and different stressors (both psychological and physiological) can trigger inflammations or prolong them [92]. Stress has also been implicated in the development of mood disorders [2, p676], hence inflammation and depression seem to be highly intertwined and difficult to analyse.

Knowledge of genes and pathways underlying these disorders could be useful not just for our understanding of the disorders and their origin, but also for the development of novel and personalized treatments that would be more effective [93]. For example, the effectiveness of antidepressants was found to be heritable [93], and targeting inhibitors of the gene *PDE10A* is being tested as a treatment for schizophrenia [77, 94]. The gene CACNA1C, which belongs to a family of genes that provide instructions for making calcium channels, encodes the α1C subunit of the L-type voltage-gated Ca21 channel and this channel is viewed as a potential target for therapeutic drugs [93]. However, positive selection is acting mainly in the introns, the regulative acting DNA sequences of this gene. Thus, positive selection is influencing the expression of this gene and may thereby

modulate neuropsychiatric disease risk in modern human populations. Consequently, future therapeutic approaches should focus specifically on this evolutionary relevant gene and order to control the expression of this gene.

5.4 Limitations

The results that we have here are only as good as the data and tools that we used, the initial assumptions made by the researchers of the GWAS studies, as well as the understanding of the disorders in question as described in the DSM-V [93]. As disorders are now seen on a spectrum rather than a binary state, there appears to be a broader possibility of overlapping. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) provides us with guides for the diagnosis of mental disorders. However, it does not represent the ultimate, unchangeable definition of the disorders, as it has gone through many revisions since its first appearance in 1952, and is still the subject of many debates [1]. Even before the existence of the manual, views of disorders were changed to fit new theories and discoveries. For example, in the 19th century, Kraepelin made the distinction between schizophrenia and manic-depressive disorder, and only in the late 1950s, the distinction between unipolar (major) depression and bipolar disorder was made. Until the 1970s, autism was considered to be childhood schizophrenia. Some of the symptoms such as psychosis, mood dysregulation, and cognitive impairments were beyond the diagnostic categories at the time. Having overlapping symptoms and foundations, or different variants of one underlying symptom sparks debate over the actual differences between disorders [21]. Therefore, the genetic basis of the disorders is only as good as our definition of them as well as the validity of the diagnosis of individuals whose data is being used in the studies.

5.5 Conclusion

Currently, there is not enough information that could give us a clear evolutionary perspective on the psychiatric disorders that we investigated. This can be contributed to the complexity, and heterogeneity of the disorders, as each of them has a range of symptoms and is connected to numerous genes, as well as to different environmental factors. However, studies like these can be useful to guide further research by singling out the most relevant genes that would then be investigated with other methods. Moreover, since the disorders we investigated share both genes and phenotypic characteristics, and can sometimes be co-morbid, the question can be posed about the way these disorders are diagnosed, and whether we need to rethink and reclassify them.

One thing that is important to have in mind is that, from an evolutionary standpoint, there is no such thing as a 'normal' genome. As the human society is growing and developing very fast, a possibility for a mismatch between the genome and the environment emerges, making certain human traits unfavourable. Therefore, it is important to understand how evolution has shaped our genome and how our recently developed lifestyles affect its functions. Further interdisciplinary research can lead us to better explanations and novel treatments for different disorders, however, it is equally important to think more on how we are creating this new world and what things we need to change in order for it to be favourable for everyone.

5.6 Ethical consideration

Every discovery in science has the potential to be misinterpreted and misused by certain individuals, in a way that can cause harm to others. Discoveries from evolutionary biology and anthropology have a history of being misused by the ones in power so it is necessary to ask the question whether it is all worth it, even though it might not seem like there is a clear-cut answer to it. Some believe that potential intellectual and medical benefits are indeed making the discoveries very well worth it, but they also advocate for taking "responsibility for the accurate popularization and public dissemination of this research, including active opposition to misinterpretation." Much of the work in science is

still done by “an unrepresentative subset of our species, namely men in developed countries”, which undoubtedly influences the way the studies are conducted [2, p14].

With studying human diversity, the problem of racism always appears. However, genetic discoveries decline the existence of discrete human groups and therefore prove that race is completely a social construct. Information from these studies can be of great importance for everyone as it may have important medical and forensic applications [2, p319].

REFERENCES

[1] Jiménez, J. P., Bottom, A., Herrera, L., Leighton, C., Rossi, J. L., Quevedo, Y., … Luyten, P. (2018). Psychotherapy and Genetic Neuroscience: An Emerging Dialog. *Frontiers in genetics*, *9*, 257. doi:10.3389/fgene.2018.00257

[2] Jobling, M., Hollox, E., & Hurles, M. (2014). Human evolutionary genetics. New York: Garland Science, Taylor & Francis Group

[3] Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., … Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419. doi.org/10.1126/science.1260419

[4] Bae, B. I., Jayaraman, D., & Walsh, C. A. (2015). Genetic changes shaping the human brain. *Developmental cell*, *32*(4), 423–434. doi:10.1016/j.devcel.2015.01.035

[5] Harrison, G. F., Sanz, J., Boulais, J., Mina, M. J., Grenier, J.-C., Leng, Y., … Barreiro, L. B. (2019). Natural selection contributed to immunological differences between hunter-gatherers and agriculturalists. *Nature Ecology & Evolution*, 3(8), 1253–1264. https://doi.org/10.1038/s41559-019-0947-6

[6] Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLOS Biology*, *4*(3), e72. doi:10.1371/journal.pbio.0040072

[7] Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., … Pritchard, J. K. (2009). The role of geography in human adaptation. *PLOS Genetics*, *5*(6), e1000500. doi:10.1371/journal.pgen.1000500

[8] N., L. K., Tobias, U., W., F. M., Kim, S., B., M. G., Armin, M., … John, O.-S. (2015). The extended evolutionary synthesis: its structure, assumptions and predictions. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1813), 20151019. doi:10.1098/rspb.2015.1019

[9] Ho, W.-C., & Zhang, J. (2018). Evolutionary adaptations to new environments generally reverse plastic phenotypic changes. *Nature Communications*, *9*(1), 350. https://doi.org/10.1038/s41467-017-02724-5

[10] Daly, M. J., Robinson, E. B., & Neale, B. M. (2016). Chapter 3 - Natural Selection and Neuropsychiatric Disease: Theory, Observation, and Emerging Genetic Findings. In T. Lehner, B. L. Miller, & M. W. B. T.-G. State Circuits, and Pathways in Clinical Neuropsychiatry (Eds.) (pp. 51–61). San Diego: Academic Press. doi.org:10.1016/B978-0-12-800105-9.00003-2

[11] Pritchard, J. K., Pickrell, J. K., & Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current biology: CB*, *20*(4), R208–R215. doi:10.1016/j.cub.2009.11.055

[12] Berg, J. J., & Coop, G. (2014). A population genetic signal of polygenic adaptation. *PLoS Genetics*, *10*(8). doi:10.1371/journal.pgen.1004412
[13] Stephan, W. (2016). Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular Ecology*, *25*(1), 79–88. doi:10.1111/mec.13288
[14] Polimanti, R., Yang, B. Z., Zhao, H., & Gelernter, J. (2016). Evidence of polygenic adaptation in the systems genetics of anthropometric traits. *PLOS ONE*, *11*(8), e0160654. doi:10.1371/journal.pone.0160654
[15] White, L., Romagné, F., Müller, E., Erlebach, E., Weihmann, A., Parra, G., … Castellano, S. (2015). Genetic adaptation to levels of dietary selenium in recent human history. *Molecular Biology and Evolution*, *32*(6), 1507–1518. doi:10.1093/molbev/msv043
[16] Polimanti, R., & Gelernter, J. (2017). Widespread signatures of positive selection in common risk alleles associated to autism spectrum disorder. *PLoS genetics*, *13*(2), e1006618. doi:10.1371/journal.pgen.1006618
[17] - Global Health Estimates 2016: Disease burden by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.
[18] Crespi, B., Summers, K., & Dorus, S. (2007). Adaptive evolution of genes underlying schizophrenia. *Proceedings. Biological sciences*, *274*(1627), 2801–2810. doi:10.1098/rspb.2007.0876
[19] Banerjee, N., Polushina, T., Bettella, F., Steen, V. M., Andreassen, O. A., & Le Hellard, S. (2019). Analysis of differentially methylated regions in great apes and extinct hominids provides support for the evolutionary hypothesis of schizophrenia. *Schizophrenia Research*, *206*, 209–216. https://doi.org/https://doi.org/10.1016/j.schres.2018.11.025
[20] Vasseur, E., & Quintana-Murci, L. (2013). The impact of natural selection on health and disease: uses of the population genetics approach in humans. *Evolutionary applications*, *6*(4), 596–607. doi:10.1111/eva.12045
[21] Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet (London, England)*, *381*(9875), 1371–1379. doi:10.1016/S0140-6736(12)62129-1
[22] van Hulzen, K. J. E., Scholz, C. J., Franke, B., Ripke, S., Klein, M., McQuillin, A., … Reif, A. (2017). Genetic overlap between attention-deficit/hyperactivity disorder and bipolar disorder: evidence from genome-wide association study meta-analysis. *Biological Psychiatry*, 82(9), 634–641. doi:10.1016/j.biopsych.2016.08.040
[23] Liu, Y., Blackwood, D. H., Caesar, S., de Geus, E. J., Farmer, A., Ferreira, M. A., … Welcome Trust Case-Control Consortium (2011). Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder. *Molecular psychiatry*,*16*(1), 2–4. doi:10.1038/mp.2009.107

[24] Pettersson, E., Lichtenstein, P., Larsson, H., Song, J., Attention Deficit/Hyperactivity Disorder Working Group of the iPSYCH-Broad-PGC Consortium, Autism Spectrum Disorder Working Group of the iPSYCH-Broad-PGC Consortium, Bipolar Disorder Working Group of the PGC, Eating Disorder Working Group of the PGC, Major Depressive Disorder Working Group of the PGC, Obsessive Compulsive Disorders and Tourette Syndrome Working Group of the PGC, Schizophrenia CLOZUK, Substance Use Disorder Working Group of the PGC, Agrawal, A., … Polderman, T. (2019). Genetic influences on eight psychiatric disorders based on family data of 4 408 646 full and half-siblings, and genetic data of 333 748 cases and controls. *Psychological medicine*, *49*(7), 1166–1173. doi:10.1017/S0033291718002039
[25] Freeman, S., & Herron, J. C. (2007). *Evolutionary Analysis*. Pearson Prentice Hall.
[26] Relethford, J. (2015). Human population genetics.
[27] Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*, 195. Retrieved from https://doi.org/10.1038/nrg2526
[28] Stoneking, M. (2017). An introduction to molecular anthropology.
[29] Hedrick, P. (2004). Population Genetics and Ecology.
[30] Wright S. (1931). Evolution in Mendelian Populations. Genetics, 16(2), 97–159.
[31] Osada N. Genetic diversity in humans and non-human primates and its evolutionary consequences. *Genes Genet Syst* (2015) 90(3):133–45.10.1266/ggs.90.133
[32] Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLOS Genetics*, *2*(4), e64. Retrieved from doi.org/10.1371/journal.pgen.0020064
[33] Haplotype/haplotypes, retrieved from https://www.nature.com/scitable/definition/haplotype-haplotypes-142/
[34] Booker, T. R., Jackson, B. C., & Keightley, P. D. (2017). Detecting positive selection in the genome. *BMC Biology*, *15*(1), 98. https://doi.org/10.1186/s12915-017-0434-y
[35] Norrgard, K. (2008) Genetic variation and disease: GWAS. *Nature Education* 1(1):87
[36] MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., … Parkinson, H. (2016). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, *45*(D1), D896–D901. doi.org:10.1093/nar/gkw1133
[37] Ikeda, M. , Saito, T. , Kondo, K. and Iwata, N. (2018), Genome-wide association studies of bipolar disorder: A systematic review of recent findings and their clinical implications. *Psychiatry Clin. Neurosci*., *72*: 52-63. doi:10.1111/pcn.12611
[38] Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(August). doi.org/10.1038/s41576-019-0127-1
[39] Dedic, N., Pöhlmann, M. L., Richter, J. S., Mehta, D., Czamara, D., Metzger, M. W., … Jurik, A. (2017). Cross-disorder risk gene CACNA1C differentially modulates susceptibility to

psychiatric disorders during development and adulthood. *Nature Publishing Group*, *23*(3), 533–543. doi:10.1038/mp.2017.133

[40] Johnson, K. E., & Voight, B. F. (2018). Patterns of shared signatures of recent positive selection across human populations. *Nature Ecology & Evolution*, *2*(4), 713–720. doi:10.1038/s41559-018-0478-6

[41] American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: Author

[42] Glahn, D. C., Knowles, E. E. M., Mathias, S. R., Almasy, L., Hodgson, K., Yao, N., ... Blangero, J. (2016). Conceptualizing Major Depression: From Genes to Neuroanatomy to Epidemiology. *In Genomics, Circuits, and Pathways in Clinical Neuropsychiatry* (pp. 487-501). Elsevier Inc.. doi.org:10.1016/B978-0-12-800105-9.00031-7

[43] Bear, M. F., Connors, B. W., & Paradiso, M. A. (2007). Neuroscience: Exploring the brain. Philadelphia, PA: Lippincott Williams & Wilkins.

[44] Kendler, K. S., Gatz, M., Gardner, C. O., & Pedersen, N. L. (2006). A swedish national twin study of lifetime major depression. *American Journal of Psychiatry*, *163*(1), 109–114. doi.org/10.1176/appi.ajp.163.1.109

[45] Bearden, C. E., Zandi, P. P., & Freimer, N. B. (2016). Molecular Architecture and Neurobiology of Bipolar Disorder. In *Genomics, Circuits, and Pathways in Clinical Neuropsychiatry* (pp. 467-486). Elsevier Inc.. doi.org:10.1016/B978-0-12-800105-9.00030-5

[46] Merikangas, K. R., Cui, L., Heaton, L., Nakamura, E., Roca, C., Ding, J., … Angst, J. (2013). Independence of familial transmission of mania and depression: results of the NIMH family study of affective spectrum disorders. *Molecular Psychiatry*, *19*, 214. Retrieved from doi.org/10.1038/mp.2013.116

[47] Cuellar, A. K., Johnson, S. L., & Winters, R. (2005). Distinctions between bipolar and unipolar depression. *Clinical psychology review*, *25*(3), 307–339. doi:10.1016/j.cpr.2004.12.002

[48] Faraone, S. V, & Larsson, H. (2018). Genetics of attention deficit hyperactivity disorder. *Molecular Psychiatry*, *24*(4), 562–575. doi:10.1038/s41380-018-0070-0

[49] Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., … Neale, B. M. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature genetics*, *51*(1), 63–75. doi:10.1038/s41588-018-0269-7

[50] Moon, A. L., Haan, N., Wilkinson, L. S., Thomas, K. L., & Hall, J. (2018). CACNA1C: Association With Psychiatric Disorders, Behavior, and Neurogenesis. *Schizophrenia bulletin*, *44*(5), 958–965. doi:10.1093/schbul/sby096

[51] Huang, J., Perlis, R. H., Lee, P. H., Rush, A. J., Fava, M., Sachs, G. S., … Smoller, J. W. (2010). Cross-disorder genomewide analysis of schizophrenia, bipolar disorder, and depression. *The American journal of psychiatry*, *167*(10), 1254–1263. doi:10.1176/appi.ajp.2010.09091335

[52] Consortium, T. 1000 G. P., Auton, A., Abecasis, G. R., Altshuler (Co-Chair), D. M., Durbin (Co-Chair), R. M., Abecasis, G. R., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*, 68. doi.org:10.1038/nature15393

[53] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[54] Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics, 47*(1), 97–120. doi:10.1146/annurev-genet-111212-133526

[55] Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., … Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837. doi:10.1038/nature01140

[56] Szpiech, Z. A., & Hernandez, R. D. (2014). selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular biology and evolution*, *31*(10), 2824–2827. doi:10.1093/molbev/msu211

[57] Lotan, A., Fenckova, M., Bralten, J., Alttoa, A., Dixson, L., Williams, R. W., & van der Voet, M. (2014). Neuroinformatic analyses of common and distinct genetic components associated with major neuropsychiatric disorders. *Frontiers in neuroscience*, 8, 331. doi:10.3389/fnins.2014.00331

[58] Nesse, R. M., & Stearns, S. C. (2008). The great opportunity: Evolutionary applications to medicine and public health. *Evolutionary applications*, *1*(1), 28–48. doi:10.1111/j.1752-4571.2007.00006.x

[59] Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4, 7. doi:10.1186/s13742-015-0047-8

[60] Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27(15)*, 2156–2158. doi:10.1093/bioinformatics/btr330

[61] Ensembl Variant Effect Predictor web interface. http://www.ensembl.org/vep.

[62] Yu, G., & He, Q. Y. (2016). ReactomePA: An R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, *12*(2), 477–479. doi:10.1039/c5mb00663e

[63] Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *BioRxiv*, 5165. doi.org/10.1101/005165

[64] Meyer, D., C Aguiar, V. R., Bitarello, B. D., C Brandt, D. Y., & Nunes, K. (2018). A genomic perspective on HLA evolution. *Immunogenetics*, *70*(1), 5–27. doi:10.1007/s00251-017-1017-3

[65] Denninger, J. W., & Marletta, M. A. (1999). Guanylate cyclase and the NO/cGMP signaling pathway, *Biochimica et Biophysica Acta (BBA) – Bioenergetics*, *1411*(3), 334–350. doi.org/10.1016/S0005-2728(99)00024-9

[66] Prugnolle, F., Manica, A., Charpentier, M., Guégan, J. F., Guernier, V., Balloux, F., … Breton, J. F. (2005). Pathogen-Driven Selection and Worldwide HLA Class I Diversity, *Current Biology, 15(11)*, 1022–1027. doi.org/10.1016/j.cub.2005.04.050

[67] Liu, X., Li, Z., Fan, C., Zhang, D., & Chen, J. (2017). Genetics implicate common mechanisms in autism and schizophrenia : synaptic activity and immunity, Journal of Medical Genetics 2017;54:511-520. https://doi.org/10.1136/jmedgenet-2016-104487

[68] Mokhtari, R., & Lachman, H. M. (2016). The Major Histocompatibility Complex (MHC) in Schizophrenia: A Review. *Journal of clinical & cellular immunology*, *7*(6), 479. doi:10.4172/2155-9899.1000479

[69] Avramopoulos, D., Pearce, B. D., McGrath, J., Wolyniec, P., Wang, R., Eckart, N., … Pulver, A. E. (2015). Infection and inflammation in schizophrenia and bipolar disorder: a genome wide study for interactions with genetic variation. *PloS one*, *10*(3), e0116696. doi:10.1371/journal.pone.0116696

[70] Dantzer, R., O'Connor, J. C., Freund, G. G., Johnson, R. W., & Kelley, K. W. (2008). From inflammation to sickness and depression: when the immune system subjugates the brain. *Nature reviews. Neuroscience*, *9*(1), 46–56. doi:10.1038/nrn2297

[71] Figueiredo, T. C. (2012). Reconsidering the Association Between the Major Histocompatibility Complex and Bipolar Disorder, *Journal of molecular neuroscience,* 47(1), 26–30. https://doi.org/10.1007/s12031-011-9656-6

[72] Wohleb, E. S., Franklin, T., Iwata, M., & Duman, R. S. (2016). Integrating neuroimmune systems in the neurobiology of depression. *Nature reviews neuroscience*, 17, 497-511. Retrieved from https://doi.org/10.1038/nrn.2016.69

[73] Kipnis J. (2016). Multifaceted interactions between adaptive immunity and the central nervous system. *Science (New York, N.Y.)*, *353*(6301), 766–771. doi:10.1126/science.aag2638

[74] Kioussis, D., & Pachnis, V. (2009). Essay immune and nervous systems: more than just a superficial similarity? *Immunity*, 31(5), 705–710. https://doi.org/10.1016/j.immuni.2009.09.009

[75] Lee, J. A., Damianov, A., Lin, C. H., Fontes, M., Parikshak, N. N., Anderson, E. S., … Martin, K. C. (2016). Cytoplasmic rbfox1 regulates the expression of synaptic and autism-related genes. *Neuron*, *89*(1), 113–128. doi:10.1016/j.neuron.2015.11.025

[76] Hamada, N., Ito, H., Nishijo, T., Iwamoto, I., Morishita, R., Tabata, H., … Nagata, K. (2016). Essential role of the nuclear isoform of RBFOX1, a candidate gene for autism spectrum disorders, in the brain development. *Scientific reports*, *6*, 30805. doi:10.1038/srep30805

[77] Kehler, J., & Nielsen, J. (2011). PDE10A Inhibitors : Novel Therapeutic Drugs for Schizophrenia, Current Pharmaceutical Design, 17, 137-150, doi.org/10.2174/138161211795049624

[78] Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., … Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics*, *50*(8), 1112–1121. doi:10.1038/s41588-018-0147-3

[79] Davies, G., Lam, M., Harris, S. E., Trampush, J. W., Luciano, M., Hill, W. D., … Deary, I. J. (2018). Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nature communications*, *9*(1), 2098. doi:10.1038/s41467-018-04362-x

[80] Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C. A., … Posthuma, D. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature genetics*, *50*(7), 912–919. doi:10.1038/s41588-018-0152-6

[81] Hill, W. D., Marioni, R. E., Maghzian, O., Ritchie, S. J., Hagenaars, S. P., McIntosh, A. M., … Deary, I. J. (2019). A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Molecular psychiatry*, *24*(2), 169–181. doi:10.1038/s41380-017-0001-5

[82] Fan, Q., Verhoeven, V. J., Wojciechowski, R., Barathi, V. A., Hysi, P. G., Guggenheim, J. A., … Mäkelä, K. M. (2016). Meta-analysis of gene-environment-wide association scans accounting for education level identifies additional loci for refractive error. *Nature communications*, *7*, 11008. doi:10.1038/ncomms11008

[83] Kakiuchi, C., Iwamoto, K., Ishiwata, M., Bundo, M., Kasahara, T., Kusumi, I., … Kato, T. (2003). Impaired feedback regulation of XBP1 as a genetic risk factor for bipolar disorder. *Nature Genetics*, 35(2), 171–175. https://doi.org/10.1038/ng1235

[84] Corbett, S., Courtiol, A., Lummaa, V., Moorad, J., & Stearns, S. (2018). The transition to modernity and chronic disease: Mismatch and natural selection. *Nature Reviews Genetics*, *19*(7), 419–430. https://doi.org/10.1038/s41576-018-0012-3

[85] Simonti, C. N., & Capra, J. A. (2015). The evolution of the human genome. *Current opinion in genetics & development*, *35*, 9–15. doi:10.1016/j.gde.2015.08.005

[86] Rodrıguez, J. A., Marigorta, U. M., & Navarro, A. (2014). Integrating genomics into evolutionary medicine, *Current Opinion in Genetics & Development, 29*, doi.org/10.1016/j.gde.2014.08.009

[87] Li, N. P., van Vugt, M., & Colarelli, S. M. (2018). The Evolutionary Mismatch Hypothesis: Implications for Psychological Science. *Current Directions in Psychological Science*, *27*(1), 38–44. https://doi.org/10.1177/0963721417731378

[88] Baird, J., Stevenson, J., & Williams, D. (2000). The Evolution of ADHD: A Disorder of Communication? *The Quarterly Review of Biology, 75*(1), 17-35. Retrieved from http://www-jstor-org.uaccess.univie.ac.at/stable/2664498

[89] Hawkins, E., Gathercole, S., Astle, D., The Calm Team, & Holmes, J. (2016). Language Problems and ADHD Symptoms: How Specific Are the Links?. *Brain sciences*, *6*(4), 50. doi:10.3390/brainsci6040050

[90] Williams, J., & Taylor, E. (2006). The evolution of hyperactivity, impulsivity and cognitive diversity. *Journal of the Royal Society, Interface*, *3*(8), 399–413. doi:10.1098/rsif.2005.0102

[91] Carta, M. G., Preti, A., & Akiskal, H. S. (2018). Coping with the New Era: Noise and Light Pollution, Hperactivity and Steroid Hormones. Towards an Evolutionary View of Bipolar Disorders. *Clinical practice and epidemiology in mental health : CP & EMH*, *14*, 33–36. doi:10.2174/1745017901814010033

[92] Kiecolt-Glaser, J. K., Derry, H. M., & Fagundes, C. P. (2015). Inflammation: depression fans the flames and feasts on the heat. *The American journal of psychiatry*,*172*(11), 1075–1091. doi:10.1176/appi.ajp.2015.15020152

[93] Wendland, J. R., & Ehlers, M. D. (2016). Translating Neurogenomics Into New Medicines. *Biological Psychiatry*, *79*(8), 650–656. https://doi.org/10.1016/j.biopsych.2015.04.027

[94] Suzuki, K., & Kimura, H. (2018). TAK-063, a novel PDE10A inhibitor with balanced activation of direct and indirect pathways, provides a unique opportunity for the treatment of schizophrenia. *CNS Neuroscience & Therapeutics*, *24*(7), 604–614. https://doi.org/10.1111/cns.12798

[95] Howard, D. M., Adams, M. J., Clarke, T.-K., Hafferty, J. D., Gibson, J., Shirali, M., … Consortium, M. D. D. W. G. of the P. G. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature Neuroscience*, *22*(3), 343–352. doi:10.1038/s41593-018-0326-7

[96] Keller, M. C., & Visscher, P. M. (2015). Genetic variation links creativity to psychiatric disorders. *Nature neuroscience*, *18*(7), 928–929. doi:10.1038/nn.4047

APPENDIX

Abstract

Despite intensive research, there are still major gaps in our understanding of the evolution of the human brain with its different structures and functions. We want to understand how different psychiatric disorders, which can have adverse effects on people's lives, fit into the evolutionary narrative since their risk loci should be under strong negative selection. It has been suggested that some genes under positive selection are linked to increased susceptibility to certain psychiatric disorders, and that these disorders emerged as humans developed higher cognitive abilities, such as language, intelligence, and creativity.

In this study, we focused on three psychiatric disorders Major Depressive Disorder, Bipolar Disorder, and Attention-Deficit/Hyperactivity Disorder to test the hypothesis that genetic loci associated with these disorders are under some form of positive selection in different human populations, and that these selected loci assemble into biochemical pathways that have a correlated evolutionary history. In order to do so, we integrated genomic data from different human populations (from four continental groups: African, European, South Asian and East Asian), data from genome-wide association studies (GWAS), and biochemical pathway analysis, and performed a genome-wide scan for recent positive selection using haplotype and single nucleotide polymorphism (SNP) based selection models.

We found evidence for positive selection acting on genes associated with different psychiatric disorders in all studied human populations. Majority of the risk loci are not shared between populations, which implies that there are differences in the selection pressures in different environments. Most of the risk loci are connected directly to the brain functions such as *CACNA1C, RBFOX1, PDE10A,* but some are not, like *HLA-B* which is involved in the immune system and was previously identified to be under some form of pathogen-driven selection. Additionally, several genes found in our study are also linked to educational attainment and mathematical ability.

We suggest that there is a certain mismatch between the genome and environmental/lifestyle conditions in contemporary society, leading to these disorders. The mismatch is caused either by the environment changing much faster than the genome can adapt to it, or these traits emerge as the genome interacts with the new environments, much different than the ones before. However, further research is needed to determine which of these hypotheses might be correct.

Keywords: natural selection, evolution, major depressive disorder, bipolar disorder, attention-deficit/hyperactivity disorder, population genetics, GWAS

Zusammenfassung

Trotz intensiver Forschung bestehen immer noch große Lücken in unserem Verständnis über die Evolution des menschlichen Gehirns und seinen verschiedenen Strukturen und Funktionen. Wir wollen insbesondere verstehen, wie verschiedene psychiatrische Erkrankungen, die sich negativ auf das Leben von Menschen auswirken können, in die evolutionäre Erzählung passen, da zu erwarten ist, dass die genetischen Risiko-Loci einer starken negativen Selektion unterliegen sollten. Es wird vermutet, dass einige Gene, die positiver Selektion unterliegen, mit einer erhöhten Anfälligkeit für bestimmte psychiatrische Störungen im Zusammenhang stehen und diese Erkrankungen auftraten, als der Mensch höhere kognitive Fähigkeiten wie Sprache, Intelligenz und Kreativität entwickelte.

In dieser Studie fokussierten wir uns auf die psychischen Störungen Depression, Bipolare Störung und Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung, um die Hypothese zu testen, dass bestimmte genetische Loci, die mit diesen Erkrankungen in Verbindung stehen, in verschiedenen menschlichen Populationen, einer bestimmten Form der positiven Selektion unterliegen sowie in spezifischen biochemischen Netzwerken eingebettet sind, die über eine zusammenhängende evolutionäre Geschichte verfügen. Zu diesem Zweck haben wir humane Genomdaten aus verschiedenen Populationen (aus vier kontinentalen Gruppen: Afrika, Europa, Südasien und Ostasien), Daten aus genomweiten Assoziationsstudien (GWAS) und biochemische Netzwerkanalysen integriert. Für die genomweite Analyse der positiven Selektion haben wir Daten von sowohl Haplotypen wie auch Einzelnukleotid-Polymorphismen (SNP) verwendet.

Für alle untersuchten Populationsgruppen fanden wir Belege für das Wirken von positiver Selektion auf Gene, die mit den von uns betrachteten psychischen Erkrankungen assoziiert sind. Die Mehrheit der Risiko-Loci sind allerdings populationsspezifisch, was darauf hindeutet, dass sich der Selektionsdruck in unterschiedlichen geographischen Umweltbedingungen unterschiedlich auf die Gene auswirkt. Die meisten Risiko-Loci unter positiver Selektion stehen in direktem Zusammenhang mit Gehirnfunktionen, wie die Gene *CACNA1C*, *RBFOX1* und *PDE10A*. Einige andere jedoch nicht, wie beispielsweise das Gen *HLA-B*, welches am Immunsystem beteiligt ist und zuvor als Gen unter pathogen-geleiteter Selektion identifiziert wurde. Darüber hinaus fanden wir in unserer Studie einige unter positiver Selektion liegende Gene, die mit Bildungsabschlüssen und mathematischen Fähigkeiten assoziiert sind.

Wir gehen daher davon aus, dass eine gewisse (evolutionäre) Fehlanpassung zwischen dem Genom und den Umwelten/Lebensbedingungen in der gegenwärtigen Gesellschaft zu diesen Krankheiten führt. Die Fehlanpassung wird entweder dadurch verursacht, dass sich die Umgebung viel schneller verändert, als sich das Genom anpassen kann, oder dass diese Erkrankungen auftreten, wenn das Genom ganz anders als zuvor mit den neuen Umgebungen interagiert. Es sind jedoch weitere Untersuchungen erforderlich, um festzustellen, welche dieser Hypothesen richtig sein könnten.

Schlagwörter: Natürliche Selektion, Evolution, Depression, Bipolare Störung, Aumerksamkeitsdefizit-/Hyperaktivitätsstörung, Populationsgenetik, GWAS

SUPPLEMENTARY MATERIAL

Table 5a: MDD variants that localize to regions with evidence of positive selection in East Asians. SNP column indicates the variant with the highest absolute normalized iHS value. F_{st} is calculated between JPT and GBR. Summaries for the genes were retrieved from genecards.org.

Risk region	SNP	Position	Risk allele	Summary for the gene	Risk loci	F_{st}	iHS
10q24.32	rs73176378	3995719	G	Protein coding gene	CSMD1	0.142966	4.85725
6p22.1	rs1611205	29759823	C	An RNA Gene, and is affiliated with the antisense RNA class.	HCG4	0	2.83792
6p21.33	rs1050564	31324562	A	Plays a role in the immune system by presenting peptides derived from the endoplasmic reticulum lumen and is expressed in nearly all cells	HLA-B	0.0155693	-5.16901
9q33.3	rs10818851	126395781	A	Member of the connecdenn family, and as such functions as a guanine nucleotide exchange factor (GEFs) for the early endosomal small GTPase RAB35 (MIM 604199) and binds to clathrin and clathrin adaptor protein-2 (AP2; see MIM	DENND1A	0.0538601	2.66146

				601024). Thus, connec-denns link RAB35 activation with the clathrin machinery			
3p14.2	rs6762122	60781899	C	The protein encoded by this gene is a P1-P3-bis(5'-adenosyl) triphosphate hydrolase involved in purine metabolism. This gene encompasses the common fragile site FRA3B on chromosome 3, where carcinogen-induced damage can lead to translocations and aberrant transcripts. In fact, aberrant transcripts from this gene have been found in about half of all esophageal, stomach, and colon carcinomas	FHIT	0.202794	2.27716
6p22.1	rs73428122	29951049	G	This gene lies within the MHC class I region. It is believed to be non-coding, but its function has not been determined.	HCG9	0.144896	3.03686
6q13	rs9360636	73809809	C	This gene is a member of	KCNQ5	0.0978858	-2.48488

				the KCNQ potassium channel gene family that is differentially expressed in subregions of the brain and in skeletal muscle.			
16p13.3	rs7196083	7334387	A	Evolutionarily conserved, and regulates tissue-specific alternative splicing in metazoa	RBFOX1	0.324292	2.48642
8q11.21	rs35583064	51190587	A	The protein encoded by this gene is a member of the syntrophin family.	SNTG1	0.0682518	3.03303
10q25.1	rs7071229	106879156	A	This gene encodes a type-I receptor transmembrane protein that is a member of the vacuolar protein sorting 10 receptor family.	SORCS3	0.0474965	2.14786

Table 5b: MDD variants that localize to regions with evidence of positive selection in South Asians. SNP column indicates the variant with the highest absolute normalized iHS value. F_{st} is calculated between PJL and GBR. Summaries for the genes were retrieved from genecards.org.

Risk region	SNP	Position	Risk allele	Summary for the gene	Risk loci	F_{st}	iHS
12p13.33	rs7311147	2707821	A	Encodes an alpha-1 subunit of a voltage-dependent calcium channel	CACNA1C	0.00377911	-2.45722
8p23.2	rs13282148	3993284	T	Protein coding gene	CSMD1	0.00157088	-3.41053
6p21.33	rs2507989	31324415	C	Plays a role in the immune system by presenting peptides derived from the endoplasmic reticulum lumen and is expressed in nearly all cells	HLA-B	0.000924243	-2.7346
9q33.3	rs7028815	126435369	G	Member of the connecdenn family, and as such functions as a guanine nucleotide exchange factor (GEFs) for the early endosomal small GTPase RAB35 (MIM 604199) and binds to clathrin and clathrin adaptor protein-2 (AP2; see MIM 601024). Thus, connecdenns link RAB35 activation with the	DENND1A	126435369	2.11395

				clathrin machinery			
11q14.3	rs1499028	88437355	A	This gene encodes a member of the G-protein coupled receptor 3 protein family.	GRM5	0.0121886	-2.44104
6q13	rs2350366	73802914	A	This gene is a member of the KCNQ potassium channel gene family that is differentially expressed in subregions of the brain and in skeletal muscle.	KCNQ5	0.0115585	-2.6384
16p13.3	rs11077035	6423281	A	Evolutionarily conserved, and regulates tissue-specific alternative splicing in metazoa	RBFOX1	0.0228812	2.16616
9p24.2	rs7021039	3418251	C	A member of the regulatory factor X gene family, which encodes transcription factors that contain a highly-conserved winged helix DNA binding domain	RFX3	0.00109562	3.52017
4p13	rs2581430	42014262	A	Protein coding gene	SLC30A9	0.0179387	-2.70579
7p21.3	rs6956916	11446902	G	The protein encoded by this gene is found almost exclusively in endothelial cells from placenta	THSD7A	0.0935088	3.12866

				and umbilical cord.			
11q14.3	rs7116589	88946742	C	The enzyme encoded by this gene catalyzes the first 2 steps, and at least 1 subsequent step, in the conversion of tyrosine to melanin.	TYR	0.00825201	2.22721
8p23.1	rs10107658	10815146	A	Protein coding gene	XKR6	0.0019486	2.92558
6p22.1	rs9393886	28050039	G	This gene encodes a member of the Kruppel family of zinc finger proteins.	ZNF165	0.0111974	2.77715

Table 5c: MDD variants that localize to regions with evidence of positive selection in Africans. SNP column indicates the variant with the highest absolute normalized iHS value. F_{st} is calculated between LWK and GBR. Summaries for the genes were retrieved from genecards.org.

Risk region	SNP	Position	Risk allele	Summary for the gene	Risk loci	F_{st}	iHS
10q21.2	rs7902632	62403569	T	Linked to the integral membrane proteins to the underlying spectrin-actin cytoskeleton and play key roles in activities such as cell motility, activation, proliferation, contact, and the maintenance of specialized membrane domains.	ANK3	0.00466626	2.58194
10q24.32	rs56946876	104653545	C	Catalyzes the transfer of a methyl group from S-adenosyl-L-methionine (AdoMet) to trivalent arsenical and may play a role in arsenic metabolism	AS3MT	0.0155494	3.20831
10q24.32	rs62504975	3099448	T	Protein coding gene	CSMD1	0.105924	3.07925
1p21.3	rs10875050	97623853	T	The protein encoded by this gene is a pyrimidine catabolic enzyme and the initial and rate-limiting	DPYD	0.0314899	-2.77173

				factor in the pathway of uracil and thymidine catabolism.			
6p21.33	rs17881210	31324448	A	Plays a role in the immune system by presenting peptides derived from the endoplasmic reticulum lumen and is expressed in nearly all cells	HLA-B	0.051136	-3.0066
5q14.3	rs571315134	87826947	G	This is an evolutionarily conserved gene that produces alternatively spliced long noncoding RNAs that may be expressed predominantly in the brain and visual cortex.	LINC00461	0.00343137	4.16421
9q33.1	rs62576078	119325576	T	This gene encodes a protein that is expressed in the brain and may function in neuronal migration, based on functional studies of the related astrotactin 1 gene in human and mouse.	ASTN2	0.0486771	-3.6675
18q21.2	rs73459710	50284023	G	This gene encodes a	DCC	0.52259	-2.42561

				netrin 1 receptor. The transmembrane protein is a member of the immunoglobulin superfamily of cell adhesion molecules, and mediates axon guidance of neuronal growth cones towards sources of netrin 1 ligand.			
3p14.2	rs2118811	61170006	A	The protein encoded by this gene is a P1-P3-bis(5'-adenosyl) triphosphate hydrolase involved in purine metabolism. This gene encompasses the common fragile site FRA3B on chromosome 3, where carcinogen-induced damage can lead to translocations and aberrant transcripts. In fact, aberrant transcripts from this gene have been found in about half of all	FHIT	0.0248776	3.58267

				esophageal, stomach, and colon carcinomas			
12q24.23	rs147349185	118061143	A	Protein coding gene. Among its related pathways are RET signaling and innate immune system	KSR	0.0987192	-4.30578
8q21.3	rs78056007	89152324	T	Proteins of the matrix metalloproteinase (MMP) family are involved in the breakdown of extracellular matrix in normal physiological processes, such as embryonic development, reproduction, and tissue remodeling, as well as in disease processes, such as arthritis and metastasis.	MMP16	0.0398003	-3.47014
2p24.3	rs6704702	15599850	G	This gene encodes a protein with two leucine zipper domains, a ribosomal protein S14 signature domain and a Sec39 like domain.	NBAS	0.0681849	5.15007

13q21.32	rs9529179	67634808	T	This gene encodes a member of the proto-cadherin family, and cadherin superfam-ily, of transmem-brane pro-teins con-taining cadherin domains.	PCDH9	0.0481356	-3.38846
16p13.3	rs11640567	7072309	T	Evolution-arily con-served, and regulates tissue-spe-cific alter-native splicing in metazoa	RBFOX1	0.0140535	-2.95395
16p13.12	rs75247378	13180441	G	Protein coding gene	SHISA9	0.0332116	-2.30228

Table 6a: BPD variants that localize to regions with evidence of positive selection in East Asians. SNP column indicates the variant with the highest absolute normalized iHS value. F_{st} is calculated between JPT and GBR. Summaries for the genes were retrieved from genecards.org.

Risk region	SNP	Position	Risk allele	Summary for the gene	Risk loci	F_{st}	iHS
10q24.32	rs73176378	3995719	G	Protein coding gene	CSMD1	0.142966	4.85725
6q22.31	rs9489446	118898675	A	The protein encoded by this gene was identified as a breast cancer antigen. Nothing more is known of its function at this time.	CEP85L	0.303354	2.62417
6p22.1	rs1611205	29759823	C	An RNA Gene, and is affiliated with the antisense RNA class.	HCG4	0	2.83792
6p21.33	rs1050564	31324562	A	Plays a role in the immune system by presenting peptides derived from the endoplasmic reticulum lumen and is expressed in nearly all cells	HLA-B	0.0155693	-5.16901
12p11.21	rs10734790	30799230	C	The protein encoded by this gene is a member of a class of approximately 20 potential Ran targets that share a sequence motif related to the Ran-binding site of importin-beta.	IPO8	0	2.11547
12q24.11	rs2075433	109833920	G	Protein coding gene	MYO1H	0.0103955	2.68807
6q27	rs524122	165736425	C	The protein encoded by this gene belongs to the cyclic nucleotide phosphodiesterase family. It plays a role in signal transduction by regulating the intracellular concentration of cyclic nucleotides.	PDE10A	0.00085939	2.85294

Table 6b: BPD variants that localize to regions with evidence of positive selection in South Asians. SNP column indicates the variant with the highest absolute normalized iHS value. F_{st} is calculated between PJL and GBR. Summaries for the genes were retrieved from genecards.org.

Risk region	SNP	Position	Risk allele	Summary for the gene	Risk loci	F_{st}	iHS
12p13.33	rs7311147	2707821	A	Encodes an alpha-1 subunit of a voltage-dependent calcium channel	CACNA1C	0.00377911	-2.45722
8p23.2	rs13282148	3993284	T	Protein coding gene	CSMD1	0.00157088	-3.41053
6q22.31	rs485416	118916395	C	The protein encoded by this gene was identified as a breast cancer antigen. Nothing more is known of its function at this time.	CEP85L	0.0115138	-3.619
6p21.33	rs2507989	31324415	C	Plays a role in the immune system by presenting peptides derived from the endoplasmic reticulum lumen and is expressed in nearly all cells	HLA-B	0.000924243	-2.7346
6q27	rs480275	165776052	T	Encodes a protein that belongs to the cyclic nucleotide phosphodiesterase family. It plays a role in signal transduction by regulating the intracellular concentration of cyclic nucleotides.	PDE10A	0.0118256	3.91601

Table 6c: BPD variants that localize to regions with evidence of positive selection in Africans. SNP column indicates the variant with the highest absolute normalized iHS value. F_{st} is calculated between LWK and GBR. Summaries for the genes were retrieved from genecards.org.

Risk region	SNP	Position	Risk allele	Summary for the gene	Risk loci	F_{st}	iHS
10q21.2	rs7902632	62403569	T	Linked to the integral membrane proteins to the underlying spectrin-actin cytoskeleton and play key roles in activities such as cell motility, activation, proliferation, contact, and the maintenance of specialized membrane domains.	ANK3	0.00466626	2.58194
10q24.32	rs56946876	104653545	C	Catalyzes the transfer of a methyl group from S-adenosyl-L-methionine (AdoMet) to trivalent arsenical and may play a role in arsenic metabolism	AS3MT	0.0155494	3.20831
10q24.32	rs62504975	3099448	T	Protein coding gene	CSMD1	0.105924	3.07925
1p21.3	rs10875050	97623853	T	The protein encoded by this gene is a pyrimidine catabolic enzyme and the initial and rate-limiting factor in the pathway of uracil and thymidine catabolism.	DPYD	0.0314899	-2.77173
6p21.33	rs17881210	31324448	A	Plays a role in the immune system by presenting peptides derived from	HLA-B	0.051136	-3.0066

				the endoplasmic reticulum lumen and is expressed in nearly all cells			
5q14.3	rs571315134	87826947	G	This is an evolutionarily conserved gene that produces alternatively spliced long non-coding RNAs that may be expressed predominantly in the brain and visual cortex.	LINC00461	0.00343137	4.16421
6q27	rs536125857	166032650	A	The protein encoded by this gene belongs to the cyclic nucleotide phosphodiesterase family. It plays a role in signal transduction by regulating the intracellular concentration of cyclic nucleotides.	PDE10A	0.284871	-3.72269
16p13.12	rs75247378	13180441	G	Protein coding gene	SHISA9	0.0332116	-2.30228
10q25.3	rs11196191	114780633	C	This gene encodes a high mobility group (HMG) box-containing transcription factor that plays a key role in the Wnt signaling pathway.	TCF7L2	0.265437	2.68933

Table 7a: ADD variants that localize to regions with evidence of positive selection in East Asians. SNP column indicates the variant with the highest absolute normalized iHS value. F_{st} is calculated between JPT and GBR. Summaries for the genes were retrieved from genecards.org.

Risk region	SNP	Position	Risk allele	Summary for the gene	Risk loci	F_{st}	iHS
10q24.32	rs73176378	3995719	G	Protein coding gene	CSMD1	0.142966	4.85725
6q22.31	rs9489446	118898675	A	The protein encoded by this gene was identified as a breast cancer antigen. Nothing more is known of its function at this time.	CEP85L	0.303354	2.62417

Table 7b: ADD variants that localize to regions with evidence of positive selection in South Asians. SNP column indicates the variant with the highest absolute normalized iHS value. F_{st} is calculated between PJL and GBR. Summaries for the genes were retrieved from genecards.org.

Risk region	SNP	Position	Risk allele	Summary for the gene	Risk loci	F_{st}	iHS
12p13.33	rs7311147	2707821	A	Encodes an alpha-1 subunit of a voltage-dependent calcium channel	CACNA1C	0.00377911	-2.45722
8p23.2	rs13282148	3993284	T	Protein coding gene	CSMD1	0.00157088	-3.41053
6q22.31	rs485416	118916395	C	The protein encoded by this gene was identified as a breast cancer antigen. Nothing more is known of its function at this time.	CEP85L	0.0115138	-3.619

Table 7c: ADD variants that localize to regions with evidence of positive selection in Africans. SNP column indicates the variant with the highest absolute normalized iHS value. F_{st} is calculated between LWK and GBR. Summaries for the genes were retrieved from genecards.org.

Risk region	SNP	Position	Risk allele	Summary for the gene	Risk loci	F_{st}	iHS
10q21.2	rs7902632	62403569	T	Linked to the integral membrane proteins to the underlying spectrin-actin cytoskeleton and play key roles in activities such as cell motility, activation, proliferation, contact, and the maintenance of specialized membrane domains.	ANK3	0.00466626	2.58194
10q24.32	rs56946876	104653545	C	Catalyzes the transfer of a methyl group from S-adenosyl-L-methionine (AdoMet) to trivalent arsenical and may play a role in arsenic metabolism	AS3MT	0.0155494	3.20831
10q24.32	rs62504975	3099448	T	Protein coding gene	CSMD1	0.105924	3.07925
1p21.3	rs10875050	97623853	T	The protein encoded by this gene is a pyrimidine catabolic enzyme and the initial and rate-limiting factor in the pathway of uracil and thymidine catabolism.	DPYD	0.0314899	-2.77173

13q31.3	rs2209880	93875303	T	The glypicans comprise a family of glycosylphosphatidylinositol-anchored heparan sulfate proteoglycans, and they have been implicated in the control of cell growth and cell division.	GPC6	0.0679644	-3.39081

Script 1:

```
#!/bin/bash

### download the 1000 Genomes files; chr-X will be downloaded sepa-
rately ###

FTP_SITE=ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502
wget $FTP_SITE/integrated_call_samples_v3.20130502.ALL.panel

for CHR in $(seq 1 22); do
   FILE=$FTP_SITE/ALL.chr$CHR.phase3_shapeit2_mvncall_inte-
grated_v5a.20130502.genotypes.vcf.gz
   wget $FILE $FILE.tbi
   sleep 60
done

FILE=$FTP_SITE/ALL.chrX.phase3_shapeit2_mvncall_inte-
grated_v1b.20130502.genotypes.vcf.gz
wget $FILE $FILE.tbi

# rename ***chrX version from v1b to v5a #
rename -v 's/v1b/v5a/' *


### prepare 1KG **autosomes** files with vcftools; minQ30, MAF 0.05,
bi-allelic, no indels ####

for CHR in $(seq 1 22); do

    ./vcftools --gzvcf ALL.chr$CHR.phase3_shapeit2_mvncall_inte-
grated_v5a.20130502.genotypes.vcf.gz --chr $CHR --minQ 30 --maf 0.05 -
-min-alleles 2 --max-alleles 2 --remove-indels --recode --out
chr$CHR.phase3_integrated.genotypes_maf5

done

# *** prepare 1KG **Chr-X** file with vcftools, MAF 0.05, bi-allelic,
no indels - rename file in the outputfile as chr23 #
```

```
./vcftools --gzvcf ALL.chrX.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz --chr X --minQ 30 --maf 0.05 --min-alleles 2 --max-alleles 2 --remove-indels --exclude-positions remove_these_pos_in_chrX.txt --recode --out chr23.phase3_integrated.genotypes_maf5


### use from the 26 populatins only 12 (from EUR, AFR, SAS, and EAS  - see http://www.internationalgenome.org/category/population/) ###


#!/bin/bash

for CHR in $(seq 1 23); do

	pops=(LWK GWD ESN GBR FIN TSI PJB BEB ITU JPT CHB KHV)

	for i in ${pops[@]};

	do

	./vcftools --vcf chr$CHR.phase3_integrated.genotypes_maf5.recode.vcf --keep pop${i}.txt --recode --out chr$CHR.phase3_integrated.genotypes_maf5_POPinfo_pop${i}

	./plink2 --vcf chr$CHR.phase3_integrated.genotypes_maf5_POPinfo_pop${i}.recode.vcf --export haps --out chr$CHR.phase3_integrated.genotypes_maf5_POPinfo_pop${i}

	./plink2 --vcf chr$CHR.phase3_integrated.genotypes_maf5_POPinfo_pop${i}.recode.vcf --make-bed --out chr$CHR.phase3_integrated.genotypes_maf5_POPinfo_pop${i}

	./plink --bim chr$CHR.phase3_integrated.genotypes_maf5_POPinfo_pop${i}.bim --cm-map genetic_map_chr$CHR.combined_b37.txt $CHR --make-just-bim --keep-allele-order --out chr$CHR.phase3_integrated.genotypes_maf5_POPinfo_pop${i}.map
```

```
awk '{print $1,$2,$3,$4}' chr$CHR.phase3_integrated.geno-
types_maf5_POPinfo_pop${i}.map.bim > chr$CHR.phase3_integrated.geno-
types_maf5_POPinfo_pop${i}.map

      done
done


# **** bashscript to transpose .haps files ****
#!/bin/bash

for CHR in $(seq 1 23); do

      pops=(LWK GWD ESN GBR FIN TSI PJB BEB ITU JPT CHB KHV

      for i in ${pops[@]};

      do

      awk '
      {
       for (i=1; i<=NF; i++)  {
            a[NR,i] = $i
      }
       maxRows = (NF > maxRows ? NF : maxRows)
       maxCols = NR
      }
      NF>p { p = NF }
      END {
          for(j=1; j<=p; j++) {
                  str=a[1,j]
                  for(i=2; i<=NR; i++){
                              str=str" "a[i,j];
                  }
                  print str
            }

      }' chr$CHR.phase3_integrated.genotypes_maf5_POPinfo_pop${i}.haps
> chr$CHR.phase3_integrated.genotypes_maf5_POPinfo_pop${i}.trans-
posed.haps
```

```
	done
done

# **** delete the first 5 rows (chr, rs, pos, snp) save then the clean transposed.haps as .thaps ****

#!/bin/bash

for CHR in $(seq 1 23); do

	pops=(LWK GWD ESN GBR FIN TSI PJB BEB ITU JPT CHB KHV)

	for i in ${pops[@]};

	do

	sed '1,5 d' chr$CHR.phase3_integrated.genotypes_maf5_POP-info_pop${i}.transposed.haps > chr$CHR.phase3_integrated.geno-types_maf5_POPinfo_pop${i}.thaps

	done
done

# ***** run SELSCAN *****

#!/bin/bash

pops=(LWK GWD ESN GBR FIN TSI PJB BEB ITU JPT CHB KHV)

for chr in $(seq 1 23); do

	for i in "${pops[@]}";

	do

	./selscan --ihs --hap chr$CHR.phase3_integrated.geno-types_maf5_POPinfo_pop${i}.thaps --map chr$CHR.phase3_integrated.geno-types_maf5_POPinfo_pop${i}.map --out chr$CHR.phase3_integrated.geno-types_maf5_POPinfo_pop${i} --threads 8

	done
```

```
done

./norm --ihs --files *.ihs.out
```

Script 2a:

```
#!/bin/bash

# this script indentifies test_genes/loci that are under positive se-
lection; the scripts counts the iHS values per loci and generates a
file with test_genes with >= 20 iHS-values per gene/loci; finaly the
script compares three populations per genetic ancestry (i.e., AFR,
EUR, SAS, EAS) for overlap of test_genes under selection, i.e.,
test_genes which have in all three population iHS-values >= 20 per
gene/loci; the script prints a final file for each genetic ancestry
with test_genes that are under positive selection in the respective
genetic ancestry (continental) population. Prepare a list (test_genes)
of genes/loci which you want to test for positive selection. The
test_gene list needs to be checked for correct 'gene_symbole' and EN-
TREZ name.

# 1.) find loci in the chromosomes which have significant (crit-value
=1 which is variable no 9) iHS values

for CHR in $(seq 1 23); do

        pops=(BEB CHB ESN FIN GBR GWD ITU JPT KHV LWK PJL TSI)

      for i in ${pops[@]};

      do
            awk '{ if ($9 == 1) { print } }' chr$CHR.phase3_inte-
grated.genotypes_maf5_POPinfo_pop${i}.ihs.out.100bins.norm.Variants >
chr$CHR.pop${i}_norm_variants_ihs2.temp
      done
done

####
```

```
# 2.) provide a gene list "test_genes" (list of genes which are of interest) and check which of these genes are under positive selection in the file generated above
####

for CHR in $(seq 1 23); do

      pops=(BEB CHB ESN FIN GBR GWD ITU JPT KHV LWK PJL TSI)

      for i in ${pops[@]};

      do
            for j in $(cat test_genes.txt); do
            if grep -wq "$j" chr$CHR.pop${i}_norm_variants_ihs2.temp
                  then echo "$j" >>
test_genes_ihs2.chr$CHR.pop${i}.txt
            fi
            done
            #this part creates an empty file if the file does not already exist
            if [ ! -e test_genes_ihs2.chr$CHR.pop${i}.txt ]
            then
            echo "nok" > test_genes_ihs2.chr$CHR.pop${i}.txt
            else
            fi
            echo "test_genes with iHS >2 finished for chr$CHR ${i}"
      done

done

#####
# 3.) count the number of significant iHS values per SNP position and gene/loci
#####

for CHR in $(seq 1 23); do

      pops=(BEB CHB ESN FIN GBR GWD ITU JPT KHV LWK PJL TSI)

      for i in ${pops[@]};
```

```
      do
            for j in $(cat test_genes_ihs2.chr$CHR.pop${i}.txt); do
            grep -o "$j" chr$CHR.pop${i}_norm_variants_ihs2.temp | wc
-l &>> counted_ih2_test_genes_chr$CHR.pop${i}.txt
            done
            echo "iHS >2 counts in test_genes finished for Chr$CHR
pop${i}"
      done

done


#####
# 4.) Rscript to merge indentified test_genes under selection and
count of iHS values per test_gene/loci
#####
echo "merging test_genes and counts for iHS values per loci"

R CMD BATCH merge_test_gene_names_count_noColnames.R

#####
# 5.) finalise the analysis: analyse the merged test_genes with counts
and get a list of test_genes wiht >= 20 iHS values
#####

for CHR in $(seq 1 23); do

        pops=(BEB CHB ESN FIN GBR GWD ITU JPT KHV LWK PJL TSI)

      for i in ${pops[@]};

      do
            awk '{ if ($2 >= 20) { print } }'
chr$CHR.pop${i}_test_genes_with_counted_ihs2_values_merged.txt >
chr$CHR.test_genes_under_selection_in_pop${i}.txt
      done
done


#####
```

```
# 6.) compare the 3 populations per genetic ancestry (i.e., AFR, EUR, SAS, EAS) for overlapp of test_genes/loci under positive selection; compare first BEB vs ITU and then BEB-ITU vs PJL; then compare  save outputfiles as .dat files
#####

for CHR in $(seq 1 23); do

      awk 'FNR==NR {a[$1]++;next} ($1 in a) {print $1}' chr$CHR.test_genes_under_selection_in_popGWD.txt chr$CHR.test_genes_under_selection_in_popESN.txt > test_genes_under_selection_in_GWD_and_ESN.temp

      awk 'FNR==NR {a[$1]++;next} ($1 in a) {print $1}' test_genes_under_selection_in_GWD_and_ESN.temp chr$CHR.test_genes_under_selection_in_popLWK.txt > chr$CHR.test.genes_under_selection_in_GWD_and_ESN_and_LWK.dat

done

# make temporary (.temp) file with genes under selection in South-East Asia; sort according to the chr numbering (chr1 to chr23)

for i in $(ls | grep ".dat" | sort --version-sort);do cat $i >> AFR_test_genes_under_positive_selection.txt;done

# delete generated temporary files
rm *.temp
```

Script 2b

```
library(data.table)
for (i in 1:23){

  for (pop in c("LWK", "GWD", "ESN", "GBR", "FIN", "TSI", "PJL",
"BEB", "ITU", "JPT", "CHB", "KHV")){

    gene_name <- paste0("test_genes_ihs2.chr",i,".pop",pop,".txt");
    print(gene_name)
    input <- read.table(gene_name, header = FALSE)

    counts <- paste0("counted_ih2_test_genes_chr",i,".pop",pop,".txt")
    print(counts)
    output <- read.table(counts, header = FALSE)

    merged_files_test_genes_counts <- merge(input, output)

    new_file <-
paste0("chr",i,".pop",pop,"_test_genes_with_counted_ihs2_val-
ues_merged.txt")
    print(new_file)

    write.table(merged_files_test_genes_counts, file= new_file,
col.names=T, row.names=F,sep=" ", quote=FALSE)
  }
}
```