Bachelor's Degree in Bioinformatics (UPF-UPC-UB)
Final Grade Project

Characterization of the full site frequency spectrum of GWAS risk alleles in psychiatric disorders

Dmytro Pravdyvets

Scientific directors: Hafid Laayouni[1], Oscal Lao Grueso[2]

[1]Bioinformatics Studies, ESCI-UPF, Pg. Pujades 1, 08003 Barcelona, Spain, [2]Population Genomics, Centre Nacional d'Anàlisi Genòmica (CNAG-CRG), Centre de Regulació Genòmica, Parc Científic de Barcelona - Torre I, Baldiri Reixac, 4, 08028 Barcelona

Abstract

Motivation:
A Single Nucleotide Variant, (SNV) is a substitution of a nucleotide at a specific position in a genome that occurs due to a unique mutation that may have a phenotypic impact. Looking at this from an evolutionary perspective, the original allele is classified as the ancestral one and the substitution as the derived one. The assignation of each allele is based on other lineages, and in this study, it is based on the allele present in chimp. Moreover, in a population, each SNV can have a different proportion of ancestral and derived alleles, which can be quantified as its Derived Allele Frequency (DAF). By estimating DAFs over all the SNVs of a population we obtain the Site Frequency Spectrum distribution, also (SFS). It has been shown that the distribution is not random, and it is shaped by demographic and selective factors [3]. Therefore, given that the selective factors depend on the phenotypical effect, by analyzing the SFS of the SNVs that are associated to a phenotype we can obtain hints about the evolution of the studied phenotype. The main tool for this study are Genome-wide association studies, (GWAS), that give us information about SNVs frequencies associated to the selected complex phenotype [6]. In our study we analyzed the SFS pattern of different psychiatric disorders: Attention Deficit Hyperactivity Disorder (ADHD) [13], Autism Spectrum Disorder (ASD) [14], Bipolar Disorder (BIP)[15], Major Depressive Disorder (MDD) [18], Cannabis Addiction (CANNABIS) [19], Obsessive-Compulsive Disorder (OCD) [17] and Schizophrenia (SCZ) [16] using the data available at Psychiatric Genomics Consortium (PGC).
.
Results:
In this study we obtained and analyzed the SFS distribution of risk alleles (SFS-risk) of the selected phenotypes of interest. By using Multidimensional Scaling, also known as classical MDS or Coordinates Analysis, we looked for similar SFS-risk patterns. and we have seen that not all the disorders have the same SFS-risk pattern, and this indicates that the selective factors that were and are acting on these disorders are different. More specifically ADHD, MDD and SCZ have an evidence of positive selection in the past, which may be due to the fact that the alleles related to those diseases may have been beneficial for the fitness of more ancient human populations, while how the environment and the society has evolved now, the mentioned traits are harmful. This statement does not seem to be true for other disorders, like BIP, where we did not observe the same pattern of the SFS-risk, meaning that this disorder evolved under different evolutive pressures.
Supplementary material: Internship Project, GitHub repository https://github.com/DimaPravdyvets/FGP.git

1. Introduction

Early Anatomically Modern Humans, that are also considered first modern human populations, evolved in a specific African continent environment. Nevertheless, humans were able to spread all over the planet Earth, conquering a wide range of drastically different environments in a relatively short time. From a biological point of view, this has been possible because of two events [1]: incorporating genetic variants from archaic species by archaic introgression [2] and/or by incorporating genetic variants that increased the fitness of the carriers in the new environment. We know that many human traits are polygenic [8], which means that they are controlled by a high number of genetic variants in the genome (loci). Hundreds and sometimes thousands of loci may be controlling a specific trait, so incorporating variation at different positions may end up affecting the same phenotype. In this type of traits, under normal conditions, a stabilizing selection is controlling the genetic variation that is holding the population close to the optimal phenotype. Nevertheless, in case that there are changes in the optimal phenotype, the population can adapt to it through small changes in frequencies through the loci that are affecting the phenotype. There

are several examples of selection acting on these traits, the easiest one to understand is height, where we can see a range of different heights across the population and not two groups of tall and short people only, other examples are skin color, eye color and other complex phenotypes.

Until recently, it was very difficult to detect polygenic adaptation as the changes in the frequencies are very small and the classical methods for detecting selection are unable to detect them [4], [5]. A possible solution to this problem would be having functional information about the SNVs associated to the phenotype of interest and, based on that, perform different types of analyses. Genome-Wide Association Studies (GWAS) is an approach that can identify the variants associated to a specific trait while reporting the effect size and statistical significance of the association. It is done by analyzing genetic variants in a large number of individuals where the complex phenotype of interest is known, and we have a set of cases and controls. The analysis is done by using genome-wide SNP arrays. By taking two groups, control and cases, it is expected to find variants at different frequencies in the two groups. The problem with commercial SNP arrays is that they identify regions with casual variants, therefore GWAS identify regions of common variants, showing regions of linkage disequilibrium that contain casual variants. The resulting data from the GWAS is imputed because genotype arrays in GWAS are based on tagging SNPs and do not genotype all of variants that are found. Hence, the genotype is inferred from a reference panel, for example 1000 Genome project. Then, for each marker a statistical test is applied to quantify the association between the phenotype and the genetic variant. The summary statistics of the GWAS analysis are interesting because they give a hint on the biological effect and the statistical relevance of such association, which is defined by its p-value.

For this study we used GWAS summary statistics for psychiatric disorders. There are few reasons for that. First, it is known that genes that are expressed in the human brain has been submitted to selective pressures [34], [35]. Second, the available data for these disorders is of high quality, usually a high sample size for the GWAS analysis, no extreme beta values for the associated variants and a derived allele frequency (DAF) fitting the expected distribution, among other factors. The latest is quite important for the evolutionary analyses carried out in this study. Also, previous studies on some of the considered disorders suggest that these traits can be under selective pressures in the past [23], [24], [25]. Another reason for doing this type of analysis in this area is because the evolutionary model underlying the disorders is unclear, as it does not fit any of the known ones. This study is aimed to help to define the evolutionary patterns of different psychiatric traits, that can later be used with the information from other studies to understand the evolutionary models psychiatric disorders are following. It is known that they do not fit the classical model, where selection should remove genetic variation that reduces the fitness of an individual [32]. This may be caused by a set of factors, starting from other genes affecting the associated to the trait of interest variants, environmental factors and different forms of balancing selection like pleiotropic antagonism and sexual antagonism among others. Is worth considering that the common variation associated with psychiatric disorders was affected by processes related to local adaptation in European populations [28]. The most interesting relationships that were found are of geo-climate relationships with disorders like minimum winter temperature with schizophrenia, precipitation rate with major depression, among others. This information indicates that even on specific population level of European population, due to different local adaptations, the acting selective pressures may vary, which can be a subject for further analysis.
The selected disorders for our study are: Attention Deficit Hyperactivity Disorder (ADHD), Autism Spectrum Disorder (ASD), Bipolar Disorder (BIP), Major Depressive Disorder (MDD), Cannabis Addiction, Obsessive-Compulsive Disorder (OCD) and Schizophrenia (SCZ). The GWAS summary statistics data was obtained from Psychiatric Genomics Consortium or PGC and were transformed previously in our lab to get the ancestral status of the variants (see Internship project).

The objectives of this project are i) description of the SFS of risk alleles of the selected disorders from the available data, ii) Identification of evolutionary pattern of the risk allele of the analyzed phenotypes based on the SFS-risk and iii) Identification of clusters of similar phenotypes based on the similarity of the patterns present in the risk allele.

2. Methods

1. GWAS Databases

GWAS data was obtained from PGC consortium (PGC webpage). The PGC consortium mainly gathers data from psychiatric GWAS studies published during the last 10 years. In this study we have focused on the psychiatric disorders described in **Tab 1** after conducting a data screening. The subset of disorders we selected was based on different criteria. Firstly, we looked at the SFS pattern of the available disorders and discarded the ones with a pattern too extreme to be explained by demography or evolutionary processes. For example, anorexia only had Risk-Derived variants and Alcohol dependence only had variants at frequencies 0.0, 0.1, 0.2, 0.8 and 0.9, having no variants in between. Additionally, we were interested in studying disorders that are linked and related. After exploring the bibliography [11], [12], we came to a conclusion that the subset of SCZ, MDD, ADHD, BIP, ASD and OCD is the one with highest linkage between them and CANNABIS was added to see if a different type of psychiatric disorder that previously has not been associated to other disorders will have a similar evolutionary pattern or not. The data used for the GWAS analysis of the datasets is of European ancestry, which adds homogeneity to our study.

The way to obtain information for a GWAS analysis is by using genome wide SNP arrays. By taking two groups, controls and cases we expect to observe different variants at different frequencies, normally the one that is associated to the trait of interest is going to be found at higher frequencies in the case group than in a control group. Once we have the frequencies, a statistical analysis is performed to indicate the likelihood of a variant being associated with a trait. The commercial SNP arrays do not generally identify causal variants, for that GWAS identify common variants, which show a region of linkage disequilibrium that contains the causal variant(s).

The data we are working with has been imputed, because genotyping arrays in GWAS are based on tagging SNPs and do not genotype all existing variants. It is done by using known genotype from a reference panel, which in case of the data was 1000 Genome project reference panel. This approach boosts the coverage of genomic variation explained by the analysis, this lets a GWAS study report the effect of more SNVs than the ones covered by the original micro-array. Another positive aspect is that this approach helps to narrow down the location of the causal variants in the linkage disequilibrium area. The reason to use this strategy is because sequencing the whole genome for each individual of the study can be too costly, so only a subset of genome is measured. In this study we are only working with SNVs, 1 to 1 substitutions, not all the possible variations like deletion, insertion or copy number variants (CNV).

2. Identification of the ancestral allele

Most of GWAS studies report the risk in relation to the allele at low frequency (MAF) or the reference allele (RAF). To be able to determine possible pattern between the risk allele and the allele state (ancestral or derived), we first assigned the ancestral state of each SNV based on the allele present in chimp. Once the ancestral state was defined for the SNVs, the frequency of the alleles was calculated and always reported towards the derived allele. To compare different studies, the Odds Ratio were converted into Beta values by applying a log transformation. In case the reported risk allele is not the same than the derived allele, we had to invert the association value, switch positions of the reference and derived allele and invert the frequency, so that the information we have is always referring to the derived allele. Besides that, for homogeneity, we transformed al the association values into Beta, changing Odds Ratio into log (Odds Ratio), which is equal to Beta and maintaining Beta in the datasets that already had it.

3. Setting the threshold for the associated SNVs for each disorder

Alleles reported as associated to a given psychiatric disease were defined using the associated p-value from the GWAS summary statistic. For this type of studies, the classical threshold for assuming that a SNV is truly associated is a p-value being smaller than 10e-8. So, our first approach to identify and quantify Risk-Ancestral and Risk-Derived SNVs was done using that threshold. Statistical power is function of sample size, allele frequency and effect size (defined as OR or beta) and p-values are highly affected by the size of the study. Due to big differences in the sample sizes across the studies, we relaxed the threshold of statistical significance in some particular phenotypes, based on the sample size (see **Tab 1**) of the related study. For SCZ, MDD and CANNABIS the threshold was maintained at 10e-8, while for ADHD, ASD, BIP and OCD the threshold was lowered to 10e-6.

Disorder	Case	Control	Total
ADHD	20183	35191	55374
ASD	18381	27969	46350
BIP	29764	31358	61122
CANNABIS	-	-	184765
MDD	246363	561190	807553
SCZ	36989	113075	150064
OCD	2688	7037	9725

Tab. 1. Sample size of the studies done for each disorder. This table represents the sample size each GWAS study that was done for the disorders had. It can be seen that OCD has less than 10000 individuals, which later implies problems with the data as we do not have enough statistical power. CANNABIS study did not describe the exact amount of cases and controls.

4. DAF rounding

To facilitate the analysis and data visualization in our study, the DAF values were rounded to 1 decimal, making a total of 10 DAF bins, starting at 0.0, and going by step of 0.1, up to 0.9. This change facilitates the representation of SFS-risk distributions, interpreting their patterns using different methodologies. Besides that, by grouping SNVs of similar frequency into 1 group, we are making sure that the results are going to be more homogeneous and with less missing data which is better for the types of evolutionary analysis that are performed in this study.

5. Multidimensional Scaling

Classical Multidimensional Scaling (MDS) [9] is a technique to reduce the dimensionality of the data from a distance matrix between objects. The map may be one, two, three and more dimensions. In our study we are using the Classical MDS, also known as Principal Coordinates Analysis (PCoA). MDS takes a similarity matrix between the

objects computed using k features (i.e. coordinates), as an input, and outputs a coordinate matrix of uncorrelated variables sorted by their contribution in the distance matrix that minimizes the loss function called strain. The strain function in classical MDS looks like this:

$$\mathrm{Strain}_D(x_1, x_2, \ldots, x_N) = \left(\frac{\sum_{i,j} \left(b_{ij} - \langle x_i, x_j \rangle \right)^2}{\sum_{i,j} b_{ij}^2} \right)^{1/2}$$

The main steps of the classical MDS are:

1. Set up the matrix of squared proximities
2. Apply the double centering: $\mathbf{B} = -1/2 * \mathbf{J} * \mathbf{P}^{(2)} * \mathbf{J}$ using the matrix $\mathbf{J} = \mathbf{I} - \mathbf{n}^{-1} * \mathbf{1}\mathbf{1}'$, where **n** is the number of objects.
3. Extract the **m** largest positive eigenvalues $\lambda_1 \ldots \lambda_m$ of **B** and the corresponding m eigenvectors $\mathbf{e}_1 \ldots \mathbf{e}_m$.
4. An m-dimensional spatial configuration of the **n** objects is derived from the coordinate matrix $\mathbf{X} = \mathbf{E}_m \mathbf{\Lambda}^{1/2}$, where $\mathbf{E}_m$ is the matrix of m eigenvectors and $\mathbf{\Lambda}_m$ is the diagonal matrix of **m** eigenvalues of B, respectively.

These new dimensions can be used to represent in a lower dimensional space the relationships of the different objects.

We used two different approaches with MDS analysis in our study. The first one is based on a chi-squared computed distances computed with a two-way table of every possible combination of disorders per each DAF bin that has a number of Risk-Derived and Risk-Ancestral SNVs. This approach gives us 10 different matrices, that are then summed into one matrix that is used as the distance matrix. Tab 2 shows an example of the two-way table used to compute the squared chi for the values of 0.0 DAF bin, between ADHD and ASD.

Disorder	Risk-Ancestral 0.0	Risk-Derived 0.0
ADHD	85	102
ASD	35	59

Tab. 2. Example two-way table ADHD and ASD. An example table that was used in the computation of distances based on chi squared, same tables were done for each frequency bin and all the possible combination of disorders.

In case of two-way tables that had 0 values in it, the resulting chi-squared did not have a value, so when obtaining the whole value by summing chi-squared of different bins, those NaN values were ignored.

The second approach was done by using Euclidean distances from the ratios of Risk-Derived / Risk-Ancestral alleles per DAF bin computed from **Tab 3.** The computed distances were later used in the classical MDS to obtain the map.

3. Results and Discussion

In this study we have analyzed the SFS of genetic variants associated to particular psychiatric disorders such as Attention Deficit Hyperactivity Disorder (ADHD), Autism Spectrum Disorder (ASD), Bipolar Disorder (BIP), Major Depressive Disorder (MDD), Cannabis Addiction, Obsessive-Compulsive Disorder (OCD) and Schizophrenia (SCZ). We have taken advantage of the summary statistics from GWAS made available at the PGC to generate the SFS using the evolutionary status (ancestral/derived) of the risk allele.

1. SFS independently of the statistical association

First, we computed the SFS independently of the statistical association. The shape of the DAF distributions we have in the datasets look like an exponential one, where we have a lot of SNVs at low frequencies and very few at high frequencies. This result is a classical theoretical prediction in population genetics [33]. From a time perspective, this result can be intuitively explained by the lifespan of a newly raised mutation. Due to forces like genetic drift and selective pressures such as negative selection, it is unlikely for a mutation to stay in a population for such a long time to reach a high frequency (right tail of the distribution), while de-novo mutations are a common event in human genomes (left tail of the distribution).

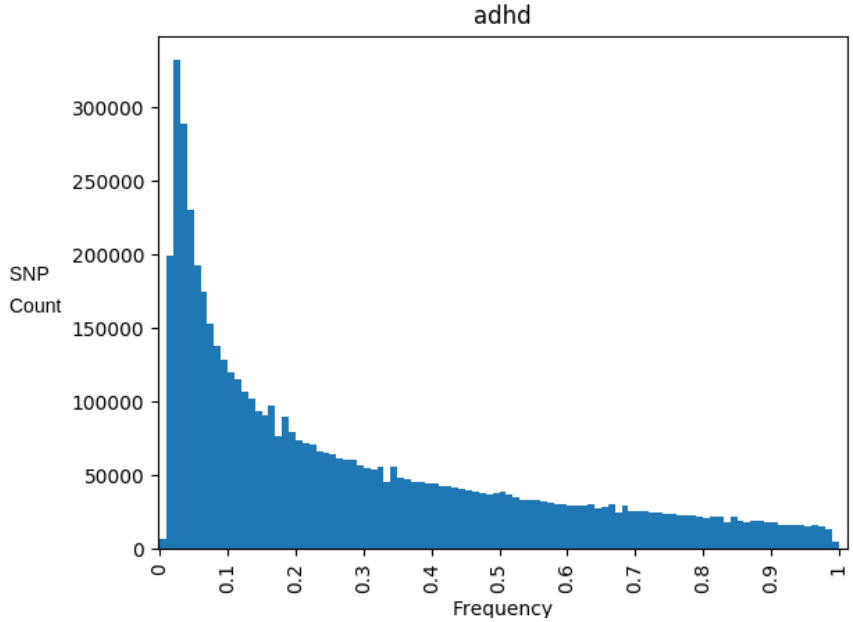

Fig. 1. Attention deficit hyperactivity disorder total SNV count. This table represents the total amount of SNVs that are found per each frequency of the derived alleles in the population. We can see that there are a lot more SNVs for low DAF frequency, and very few for high frequency. The shape of DAF distribution looks like this because of different genetic forces acting on the SNVs, making it hard for them to get to high frequencies, while because of de-novo mutations and other events we get a lot of SNVs at low frequencies.

2. Total number of associated SNVs after the filtering

Once the thresholds of statistical association between the phenotype and the genetic variation were established, we used them to discard the non-significant SNVs. The subset of data we were left with was then separated based on the ancestral state of the SNV. **Tab 3** shows the total counts of SNVs per different bins of DAF.

Disorders	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Total
ADHD	187	115	235	524	423	195	254	107	11	4	2055
ASD	94	83	1822	124	94	118	204	19	16	1	2575
BIP	308	836	610	535	681	409	162	102	197	8	3848
CANNABIS	4	50	115	213	13	24	230	11	30	0	690
MDD	157	562	783	478	992	708	268	311	117	7	4383
SCZ	1374	1914	2099	1520	1309	832	839	722	215	25	10849
OCD	11	2	26	22	23	10	4	7	0	0	105

Tab. 3. Number of SNVs per DAF bin after the filtering. This table represent the number of SNVs each disorder has after filtering based on the p-value associated to each SNV. BIP, OCD, ADHD and ASD had a p-value threshold of $< 10e^{-6}$, while MDD, SCZ and CANNABIS were filtered using the classical threshold of $<10e^{-8}$. The difference of the threshold is due to big differences in the sample sizes of the studies.

We can see that SNVs reported as significant at high derived allele frequencies in the population are rare, or, in case of some disorders, absent. Besides that, we can see that frequencies between 0.3 and 0.7 are the ones that have the greatest number of variants in general. This is something that we expect to observe, considering that we are looking at polygenic traits and that GWAS have more power to detect statistically significant associations in variants at intermediate frequencies [6].

3. Number of Risk-Ancestral and Risk-Derived SNVs in the selected disorders and SFS-risk distribution

After the initial screening, we quantified the amount of SNVs of each type (Risk-Derived and Risk-Ancestral) that are found in each bin of DAF, the bins were defined using 0.1 step, meaning that there is a total of 10 bins, starting at 0.0 and ending at 0.9. The obtained table is the primary tool used in the rest of our analysis in this study. Some disorders do not have SNVs at certain bins or only have one type of alleles at that bin, making difficult to consider these bins in following analyses. **Tab 4** represents this information.

DAF	0.0		0.1		0.2		0.3		0.4		0.5		0.6		0.7		0.8		0.9	
Disorder	R-A	R-D	R-A	R-D	R-A	R-D	R-A	R-D	R-A	R-D	R-A	R-D	R-A	R-D	R-A	R-D	R-A	R-D	R-A	R-D
ADHD	85	102	47	68	90	145	241	283	200	223	96	99	103	151	33	74	8	3	3	1
ASD	35	59	38	45	631	1191	56	68	39	55	59	59	91	114	11	8	6	10	0	1
BIP	152	156	325	511	246	364	218	317	308	373	167	242	57	105	36	66	66	131	5	3
MDD	69	88	246	316	359	424	239	239	483	509	372	336	138	130	135	176	62	55	7	0
SCZ	586	788	786	1128	868	1231	623	897	519	790	360	472	387	452	311	411	98	117	7	18
CANN	1	3	29	21	54	61	100	113	9	4	9	15	105	125	7	4	14	16	0	0
OCD	3	8	0	2	12	14	14	8	16	7	8	2	2	2	5	2	0	0	0	0

Tab. 4. Total number of Risk-Derived and Risk-Ancestral alleles per DAF bin. Main table of the analysis where the amount of Risk-Derived (R-D) and Risk-Ancestral (R-A) alleles found in each DAF bin, CANN represents CANNABIS in this table.

Substantial differences were observed between the SFS distributions for each disease when only the statistically significant SNVs where considered. **Fig 2**, **Fig 3**, **Fig 4** and **Fig 5** show the SFS distribution for each of the considered disorders of the study.

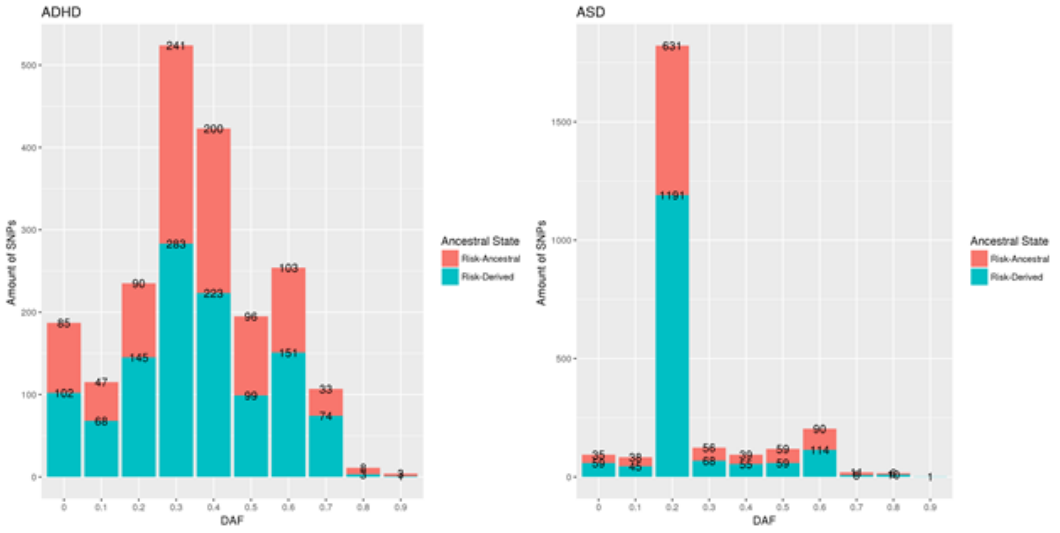

Fig. 2. Attention deficit hyperactivity disorder and Autism Spectrum Disorder Ancestral State barplot. The barplots show us the amount of Risk-Ancestral and Risk-Derived alleles that are found in each 0.1 bin of DAF frequency. ADHD shows and enrichment of SNVs in medium frequencies while ASD has a huge enrichment at frequency of 0.2, which is strange and can be caused by different genetic events like linkage disequilibrium, among others.

ADHD plot shows that the frequency of SNVs at high DAF is small. SNVs at low DAF are a bit more frequently identified as statistically significant, and the most at medium frequencies. This result suggests that the distribution of SNVs statistically associated to a particular phenotype is a mixture between the genetic architecture of the disease, and the statistical power for detecting the associations. For SNVs at low and high DAF the statistical power for detecting an association given an effect size is smaller than for SNVs at intermediate frequencies. Since most of the SNVs in the population are at low frequency due to the evolutionary processes described previously, the expected proportion of statistically significant SNVs at high frequencies should be smaller compared to the fraction of SNVs at low frequencies, and both smaller than the proportion of SNVs at intermediate frequencies due to GWAS power. However, this expected pattern is not observed in ASD. This phenotype shows a very different and unique pattern that is not observed in other disorders, where at DAF=0.2, we see a huge enrichment of significant variants. This may be caused by allelic association or linkage disequilibrium where we have dependencies of frequencies at more than 2 loci.

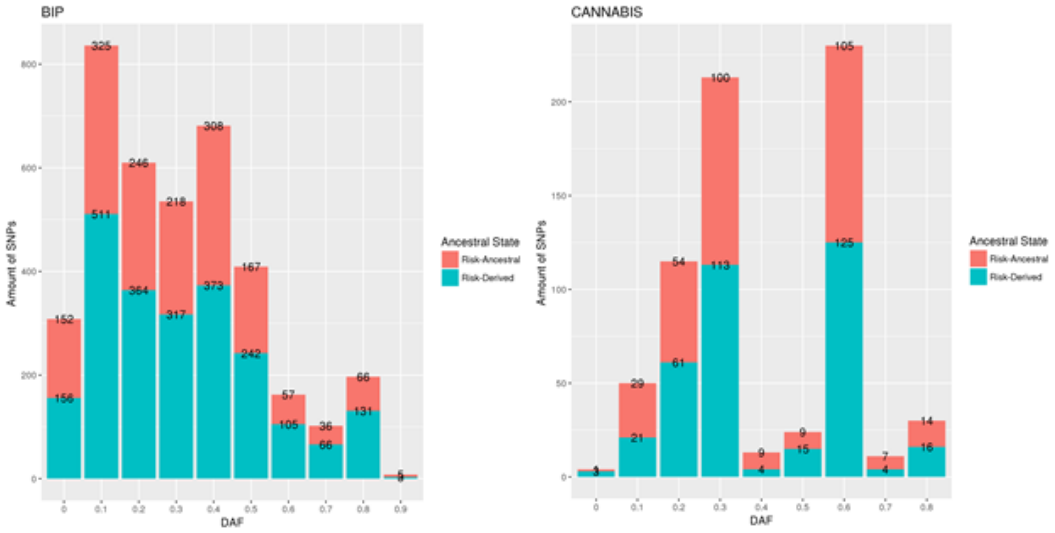

Fig. 3. Bipolar Disorder and Cannabis use disorder Ancestral State barplot. This barplots show us the amount of Risk-Ancestral and Risk-Derived alleles that are found in each 0.1 bin of DAF frequency. BIP shows a similar pattern to ADHD, where we have more SNVs at low-medium frequencies and less at high frequencies. CANNABIS has strange gaps at 0.4 and 0.5 frequency, which is not expected to happen and is quite strange.

BIP shows a similar distribution as ADHD, while in CANNABIS there is a gap with few SNVs at 0.4 and 0.5. This result is unexpected, and it probably relates to some kind of ascertainment bias of the study when defining the SNVs. In fact, the data for CANNABIS study is a combination of data from 3 different resources, ICC (International Cannabis Consortium), UK-Biobank and 23andMe, because of that the genotypes were imputed using 1000 Genomes project phase 1 release reference set.

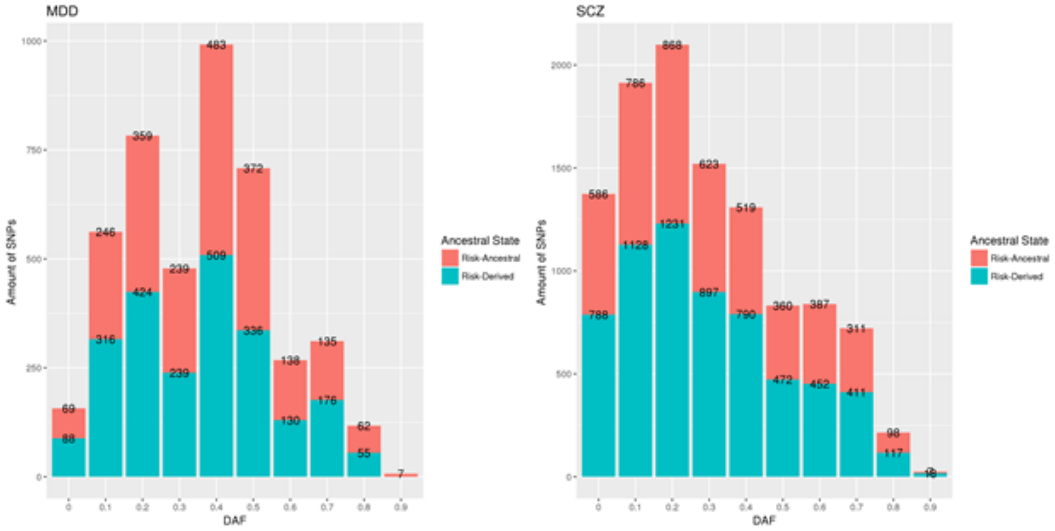

Fig. 4. Major Depressive Disorder and Schizophrenia Ancestral State barplot. This barplots show us the amount of Risk-Ancestral and Risk-Derived alleles that are found in each 0.1 bin of DAF frequency. MDD and SCZ show the best plots in terms of no missing data, proper distributions and interesting patters, where in case of SCZ we can see that the amount of Risk-Ancestral alleles is decreasing for higher DAF.

MDD and SCZ have very good proportions of SNVs per DAF bin. This can be due to the fact that those two disorders are the ones with the biggest sample sizes. Consequently, more significant SNVs have been identified. SCZ plot shows a very clear decrease in the number of Risk-Ancestral SNVs for higher frequency.

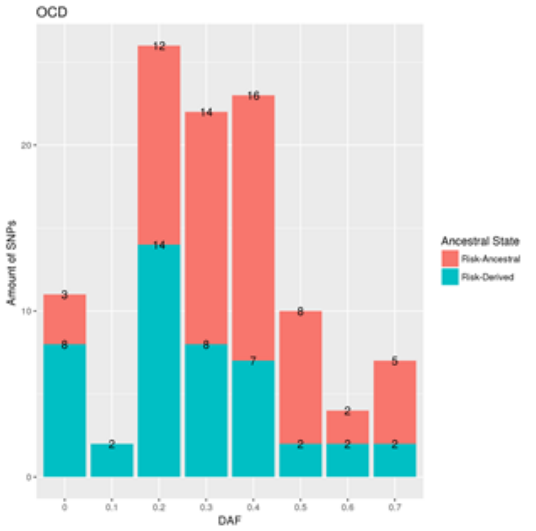

Fig. 5. Obsessive Compulsive Disorder Ancestral State barplot. This barplots show us the amount of Risk-Ancestral and Risk-Derived alleles that are found in each 0.1 bin of DAF frequency. As can be seen, there are frequencies where we have no SNVs like 0.8 and 0.9 and at 0.1, we only have Risk-Derived SNVs which makes it more difficult for us to use this data for other analysis. The most likely explanation of this is the small sample size for the OCD GWAS study as there are less than 10000 individuals in total, and even with a lowered to $< 10e^{-6}$ p-value threshold, we only get 105 SNVs in total.

As can be seen, due to the very small sample size of the OCD GWAS, we have only 105 significant SNVs. Moreover, some bins like 0.8 or 0.9 do not have significant SNVs and 0.1 only have 2 Risk-Derived SNVs. Overall, these results precluded meaningful results for OCD in some analyses.

Given that the number of SNVs statistically associated depends on both biological and statistical factors, in order to control by the statistical power of each study to compare the diseases from an evolutionary point of view, we computed the ratio of statistically significant SNVs where the risk allele is the derived against the ones where the risk allele is the ancestral for each DAF bin, the resulting ratios can be seen in **Tab 5**. **Fig 6** shows the plot of the logarithm of the ratios.

Disorder	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ADHD	1.2	1.44	1.61	1.17	1.12	1.03	1.46	2.24	0.375	0.33
ASD	1.68	1.18	1.88	1.21	1.41	1	1.26	0.72	1.6667	NA
BIP	1.02	1.57	1.47	1.45	1.21	1.44	1.84	1.83	1.985	NA
MDD	1.27	1.28	1.18	1	1.05	0.90	0.94	1.30	0.8871	NA
SCZ	1.34	1.46	1.41	1.43	1.52	1.31	1.16	1.32	1.1939	NA
OCD	2.66	NA	1.16	0.57	0.43	0.25	1	0.4	NA	NA
CANNABIS	NA	0.72	1.12	1.13	0.44	1.67	1.19	0.57	1.1429	NA

Tab. 5. Ratio between Risk-Derived and Risk-Ancestral alleles per DAF bin. This table shows us the data we obtained by computing the ratio between Risk-Derived and Risk-Ancestral alleles of different disorders per each DAF bin. The NA values occur because some bins i) do not have associated variants, ii) only have alleles of one type, either Risk-Derived or Risk-Ancestral, so the ratio cannot be computed or iii) due to very small amount of associated SNVs found in this bin, the obtained ratio is not representative and was not included.

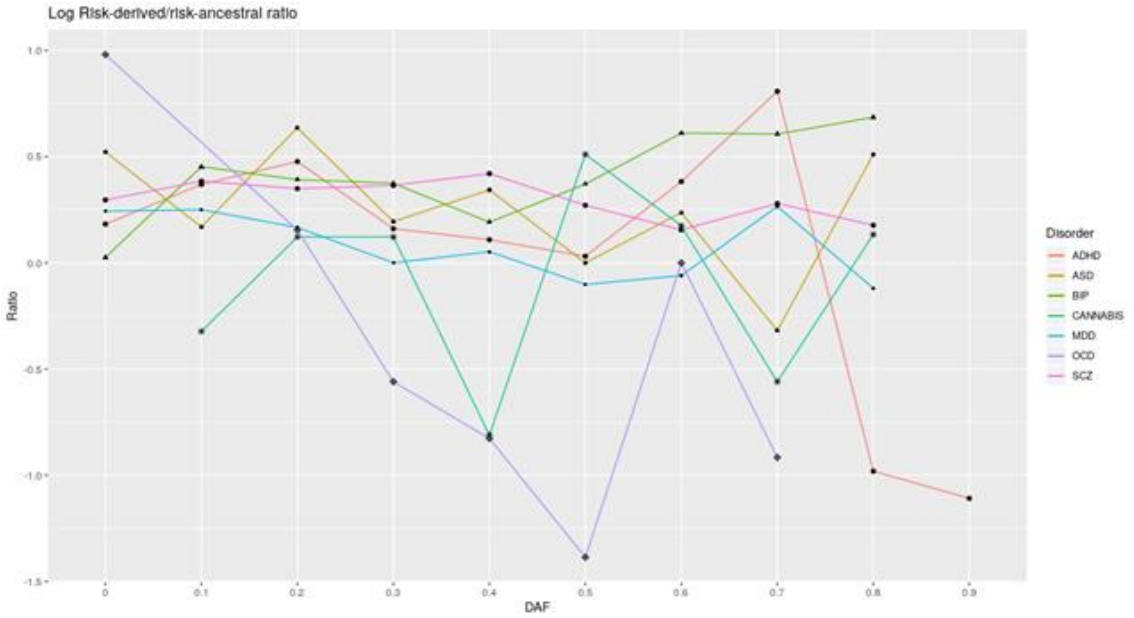

Fig. 6. Ratio between Risk-Derived and Risk-Ancestral frequencies. The plot shows the logarithm of the ratio between Risk-Derived and Rik-Ancestral alleles found at each bin of frequency we have for each disorder. Values above 0 indicate that for a given DAF category there are more SNVs where the risk allele is the derived allele whereas negative values indicate that there are more SNVs where the ancestral allele is the risk allele. Notice that for some diseases and DAF there were not enough observations of one of the categories and therefore they are not represented. It can be seen that some disorders show a similar pattern. For example, MDD and SCZ show a clear decrease in Risk-Derived and increase in Risk-Ancestral SNVs for higher frequencies Similar statement can be done for ASD, ADHD and CANNABIS, even though their patterns are different. BIP on the other hand show a completely opposite pattern, where there are more Risk-Derived alleles for higher frequencies than for lower ones. This may imply different selective forces acting on different disorders. OCD is very high to discuss as it has very little SNVs and some of the ratios may not be representative. Besides that, most of the disorder do not have ratio at 0.9 bin, due to lack of data and the resulting ratios being not representative.

If there were no selective pressures, one can expect that within each DAF category the probability of being the risk allele should not depend on its evolutionary status (that is, the log of the ratio should be 0 or close to 0). However, the fact that we observe discrepancies in the ratios among diseases and DAF suggests that there are different selective pressures between diseases and between DAF categories. If we think in the classical birth-death-fixation process of a derived allele, one can expect that derived alleles at low frequencies in the population (i.e. DAF closer to 0) are relatively young in the population or are under negative selective pressures. In contrast, derived alleles at high frequency (i.e. DAF closer to 1) imply that either they have survived for long in the population or that there have been positive selective pressures for increasing its frequency in the population. Therefore, for a given disease the differences of the ratio along the DAF profile should reflect the evolutionary history of the disease.
We can see that most of the considered disorders deviate from the ratio we would expect with no selective pressures. More specifically, SCZ and MDD show a very similar pattern, where the ratio tends to decrease for higher DAF categories. Taking into account that an event of positive selection would tend to increase the frequency of derived allele and that an event of negative selection would tend to decrease the frequency and increase the number of SNVs at low frequency, this result is in agreement with the presence of purifying selection ongoing in these phenotypes for long time. That is, new (derived) alleles that cause or are associated to an increase in the risk of suffering the disease are not tolerated and cleaned from the population over time. Interestingly, a previous study analyzing SCZ by means of classical tests of selection suggested that this trait was under positive selection [27] and same conclusions were obtained for ASD [23]. Our results would point to positive selective pressures towards the protective alleles against SCZ, in agreement with a recent study found evidence supporting a recent negative selection for SCZ risk alleles using another type of statistics [30]. Another study of selective pressures for psychiatric disorders detected evidences of negative selection for SCZ. A similar statement can be made about ASD and CANNABIS, even though the pattern of these disorders is different compared to SCZ and MDD, but a similar one between them. An interesting pattern is seen in ADHD. The frequency of SNVs where the risk is the derived allele increases with the DAF up to 0.7. This implies that genetic variants associated to the phenotype were more tolerated in the past that nowadays, in agreement with a previous study [24] and [31]. However, caution must be taken because the ratios > 0.7 seem to be not as representative as the rest. For BIP we see exactly the opposite pattern. For higher DAF we have more Risk-Derived SNVs than for low DAFs.

Our results can also be interpreted from a biological point of view. Most of the analyzed phenotypes are associated to a lower fitness. It has been seen that patients affected by schizophrenia, autism, and anorexia nervosa had significantly less children [32]. Therefore, our results would agree with this empirical observation. However, ADHD patients show a mixture of phenotypes which, on average, should also decrease their fitness [36]. In our case, we observed a mixture of effects. For DAF close to 0 (I.e. present to current times) our results would agree with a lower fitness of ADHD. However, for older times this pattern is reverted. Overall, caution is advised when trying to interpret the evolutionary history of a disease based on the fitness that is estimated in the current particular environment. BIP did not show strong negative selection in the analysis, while in our study we can even see positive selection for BIP ancestral alleles. Lastly

vulnerability to depression and substance use disorders may be preserved by balancing selection [32], suggesting that common variants depend on other genes or environmental factors and that the causal genes seem to be beneficial in the siblings of the affected individuals, which may help to explain the fact for positive selection for the associated SNVs in the past. Lastly OCD shows a unique pattern, but this is explained by small number of associated SNVs, which gives lack if ratio at different DAF bins like 0.1, 0.8, 0.9 most likely due to the very small sample size of the related study.

In summary, these results suggest that the different phenotypes are evolving by different evolutionary and selective forces. Besides that, we can also see that not all the patterns are similar. The next step in our analysis is to find similarities between the patterns and see what disorders have evolved in a similar way.

4. MDS analysis

In order to reduce the dimensionality to classify the diseases according to their evolutionary pattern as defined by the frequency of risk alleles associated to either the ancestral state or the derived state at each DAF bin, we used MDS. First, we analyzed the results obtained from the squared chi distances from the **Tab 6**, the resulting map can be seen in the **Fig 7**.

Disorder	ADHD	ASD	BIP	MDD	SCZ	CANNABIS	OCD
ADHD	0	13.32	15.9	20.38	28.26	14.72	12.43
ASD	13.32	0	21.82	44.48	24.14	11.80	10.36
BIP	15.9	21.82	0	58.89	29.08	20.49	16.55
MDD	20.38	44.48	58.89	0	64	9.564	8.23
SCZ	28.26	24.14	29.08	64	0	13.60	17.02
CANNABIS	14.72	11.8	20.49	9.56	13.60	0	5.94
OCD	12.43	10.36	16.55	8.23	17.02	5.94	0

Tab. 6. Chi-Squared distances computed from original table containing all SNVS. This is the distance matrix that was computed using chi-squared on all the possible combination of disorders using a two-way table as shown in **Tab 4**. The diagonal distances are equal to 0, because we are comparing the disorder to itself.

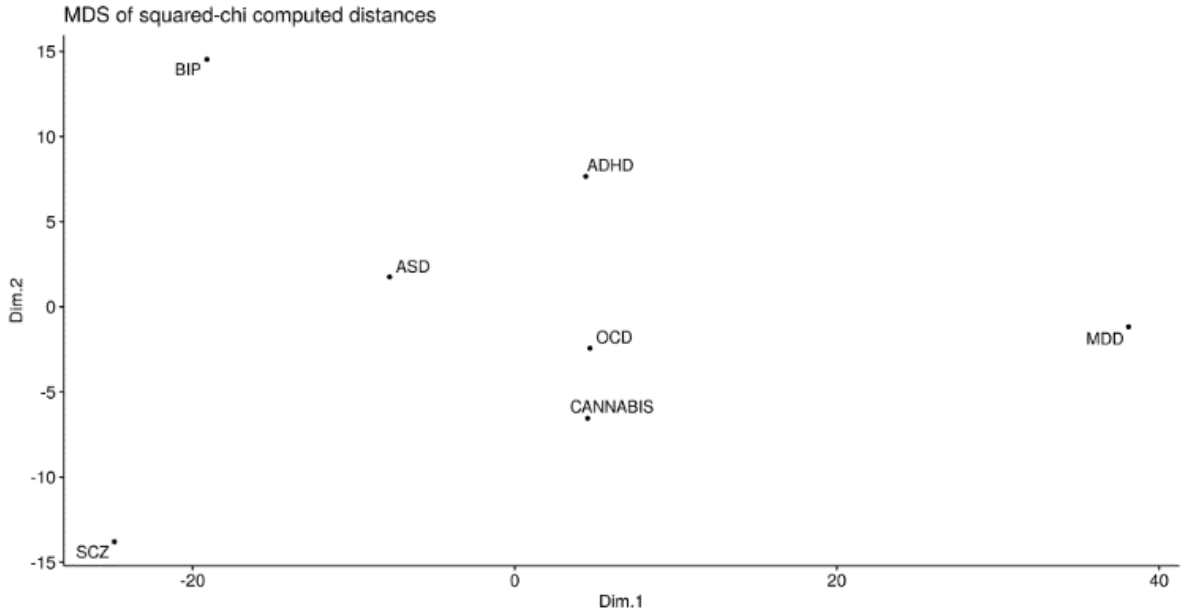

Fig. 7. MDS of chi-squared based distances. MDS map computed based on the chi-squared distances of two-way tables done across all the possible combination of disorders using data from **Tab 6**. Based on this map we can see that ASD, ADHD, OCD and CANNABIS are similar, while BIP, MDD and SCZ are different from the other disorders.

The first dimension (41.04% of variance explained) places MDD at one side and SCZ at the other. The second dimension (20.71%) separates SCZ and BIP. Considering both, we can see that OCD is close to CANNABIS, ADHD and ASD. This means that these four disorders show a similar risk allele profile over all the DAF bins. SCZ, BIP and MDD are far from all the points nor they are together if we look at them in Dimension 1. However, looking at MDD in Dimension 2 we can see that it is close to the original cluster of OCD, CANNABIS, ADS and ADHD.

The second approach was done by using the ratios that were computed in the previous part of the study using **Tab 3**,

those ratios can be seen in **Tab 4**. First, we computed the Euclidean distances between the ratios. This distance was later used in the Classical MDS. **Fig 8** shows the projection of the different diseases in the first two dimensions.

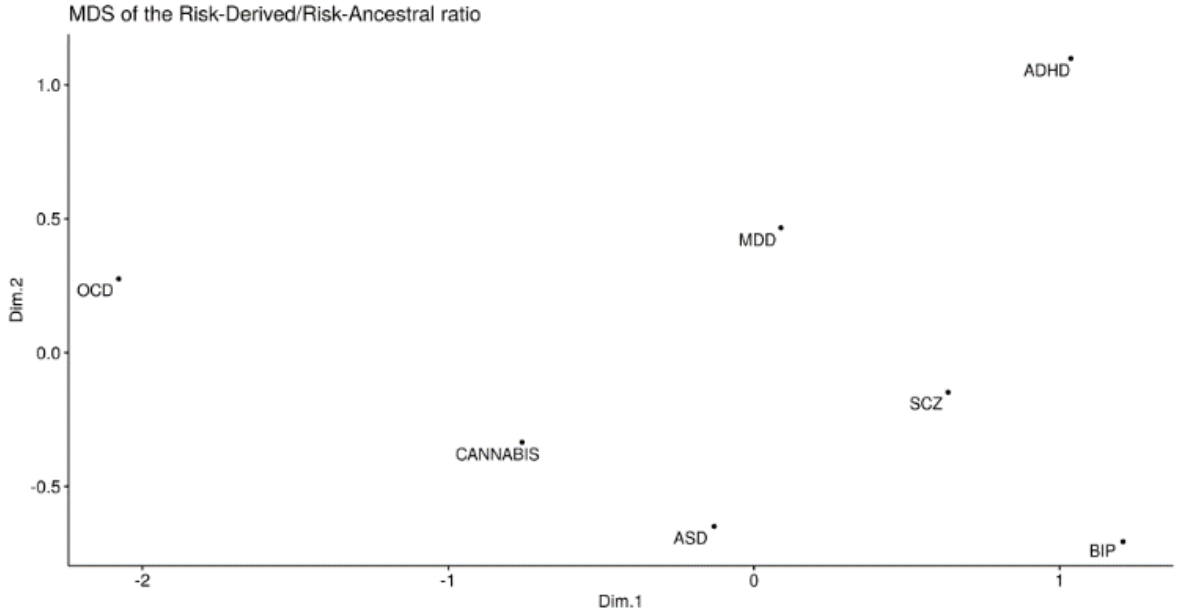

Fig. 8. MDS of the computed ratios. MDS map computed based on the Risk-Derived/Risk-Ancestral ratios from **Tab 5.** This map indicates that CANNABIS is similar to ASD, MDD is similar to SCZ. ADHD, BIP and OCD are the most different ones.

This map is extremely interesting, as we can see the results we were getting from observing the original ratio plot can be also seen here. MDD, SCZ appear close on this map in both dimension 1 (34.49 % of variance explained) and 2 (18.76% of variance explained), this is interesting considering, that the risk allele ratio profiles over all the DAF bins in the original graph are also quite similar. CANNABIS and ASD also seem to be close and their risk allele ratios profile across the DAF bins show multiple similarities, it is interesting considering that it has been tested that they have strong association on genetic correlation level [19]. OCD is the disorder with less data, which ends up giving not complete ratios across the DAF bins, this is most likely the cause for it to be an outlier with no neighbors around. BIP and ADHD seem to be close to MDD and SCZ on dimension 1, this can be explained by the fact that BIP has a very similar pattern with MDD up to DAF of 0.6, then it is unique. ADHD is also quite similar to MDD and SCZ, but because of less data, there is a big drop-off in ratio at high DAFs which may be the reason for ADHD to be higher than MDD and SCZ on the map.

5. Putting everything into perspective

From a biological point of view and considering other studies in the field [11], [12], [20] that suggest a strong relationship between SCZ with ADHD, MDD, OCD and BIP are also confirmed by our study, besides BIP, where we can clearly see, that in both MDS approaches and the original **Fig 6,** the present SFS-risk pattern is not alike the other disorders. So even though similar genetic variants [11], [22] affect those disorders, the evolutionary forces that are acting upon them are different. Another interesting fact is that based on other studies ASD has very little similarity in genetic variants that underly different disorders besides ADHD [11], but our study shows that the selective forces acting on ASD are also very similar to the ones acting on CANNABIS which reinforces the results for the found genetic correlation between the disorders [19]. ASD and ADHD do not show similarities in our study, while other studies suggest that these two disorders are related and have comorbidities on different levels [37]. Also, from the point of view of substance use disorders and in our case CANNABIS, other studies [20], [21], suggest that individuals that already have a set of other disorders like MDD, ADHD, BIP among others tend to develop addictions to different substances, which may explain the similar evolutionary pattern present in them, as they could be co-evolving.
Also, our results can be compared with the paper published by Cross-Disorder Group of the Psychiatric Genomics Consortium [22], where their group studied eight psychiatric disorders with data also from PGC. The result of their meta-analysis across eight disorders (anorexia nervosa, attention-deficit/hyper-activity disorder, autism spectrum disorder, bipolar disorder, major depression, obsessive-compulsive disorder, schizophrenia, and Tourette syndrome) detected 109 loci associated with at least 2 psychiatric disorders, where 23 loci have pleiotropic effect on four or more disorders and 11 loci with antagonistic effects on multiple disorders. Besides that, the SNV-based genetic correlation indicates that high correlation between SCZ and BIP, MDD with ASD and ADHD [22], which if compared to our study MDD, ASD and ADHD show similar SFS-risk pattern together with SCZ and CANNABIS, while BIP shows a completely different behavior. This paper also shows cases the functionality of the associated SNVs that are mainly expressed in brain and pituitary, more specifically pleiotropic ones, while non-pleiotropic ones are enriched for occipital cortex. Considering this information and other results from Cross-Disorder Group of the Psychiatric Genomics Consortium together with the results of our study may be very useful and interesting. It could be possible to

explain why specific SNVs are present and significant in multiple disorders using the evolutionary information obtained for each disorder. This can help to direct the research towards identifying specific causal pathways [32].

4. Conclusions

This study has shown us that psychiatric disorders are the results of different evolutionary trajectories and selective pressures. Some of the traits may have been beneficial in the past, but signs of purifying selection also seem to appear when looking at the SFSs. Still this statement is not true for all of them, as BIP has a very different pattern if compared to the rest, while MDS has shown us that MDD, SCZ are very alike, same is true for ASD and CANNABIS that maintained their similarities through different analysis, ADHD is also close to them, and OCD has lack of data, which makes it hard to compare it to the rest of disorders. The similarities of evolutionary patterns that we find can also be reinforced by other studies of comorbidities [20], temporal relationships [21] and genetic similarities lying under the disorders of interest [11].

This study can be repeated using other psychiatric or related disorders and other populations, not only European because we know that DAF frequencies of the same SNVs tend to be different in a lot of traits in Asian, African and other populations, which can be incredibly interesting and help us even more to understand the underlying selective pressures acting on these complex phenotypes in different parts of the world, to see if they are similar or not and possibly to understand the reason behind that. Understanding the evolution of those disorders can be later applied in the field of evolutionary medicine by looking at the functionality of alleles that are co-evolving or evolving with similar patterns in different disorders, which may help with the treatment of the disorders [10].

And finally, this study brings GWAS summary statistic data into a new field of research, where it can be used to shed light on the evolutionary history of human diseases. By simply looking for the ancestral status of the associated variants from specific GWAS studies of interest we can obtain new information which is extremely important and useful for evolutionary biology and medicine [29].

References

[1] Yair Field, Evan A Boyle, Natalie Telis, Ziyue Gao, Kyle J. Gaulton, David Golan, Loic Yengo, Ghislain Rocheleau, Philippe Froguel, Mark I. McCarthy, Jonathan K. Pritchard* Detection of human adaptation during the past 2000 years. Science First Release, VOL 354, ISSUE 6313, 13 October 2016

[2] Simonti CN, Vernot B, Bastarache L, The phenotypic legacy of admixture between modern humans and Neandertals. Science. 2016 Feb 12;351(6274):737-41

[3] Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, Carlos D. Bustamante Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLOS GENETICS October 23, 2009

[4] Angela M. Hancock and Anna Di Rienzo Detecting the Genetic Signature of Natural Selection in Human Populations: Models, Methods, and Data. Annu Rev Anthropol. 2008; 37: 197–217.

[5] Pritchard, J., Di Rienzo, A. Adaptation – not by sweeps alone. Nature Reviews Genetics volume 11, 665–667 (2010).

[6] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé and David Meyre, Guillaume Paré Benefits and limitations of genome-wide association studies. Nature Reviews Genetics Psychiatry volume 20, 467–484 (2019)

[7] Renato Polimanti, Joel Gelernter Widespread signatures of positive selection in common risk alleles associated to autism spectrum disorder. PLOS GENETICS, February 10, 2017

[8] Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169, June 15, 2017 a 2017 Elsevier Inc.

[9] Hout MC, Papesh MH, Goldinger SD. Multidimensional scaling. Wiley Interdiscip Rev Cogn Sci. 2013;4(1):93-103. doi:10.1002/wcs.1203

[10] Grinde B. Evolution-based Approach to Understand and Classify Mental Disorders. J Psychol Brain Stud. 2017, 1:1.

[11] Michael Marshall ROOTS OF MENTAL ILLNESS Researchers are beginning to untangle the common biology that links supposedly distinct psychiatric conditions, Nature, Vol 581, 7 May 2020.

[12] David Adam, Research suggests that mental illnesses lie along a spectrum— but the field's latest diagnostic manual still splits them apart, NATURE, VOL 496, 25 APRIL 2013

[13] Demontis, D., Walters, R.K., Martin, J Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nature Genetics. 28 September 2018.

[14] Jakob Grove, Stephan Ripke, Anders D. Børglum Identification of common genetic risk variants for autism spectrum disorder. Nature Genetics, volume 51, 431–444 (2019).

[15] Eli A. Stahl, Gerome Breen, the Bipolar Disorder Working Group of the Psychiatric Genomics Consortium Genome-wide association study identifies 30 Loci Associated with Bipolar Disorder. Nature Genetics volume 51, pages793–803(2019)

[16] Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature, 2014 Jul 24;511(7510):421-7

[17] International Obsessive-Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) and OCD Collaborative Genetics Association Studies (OCGAS). Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. Nature Molecular Psychiatry 2018 May;23(5):1181-1188.

[18] Howard, D.M., Adams, M.J., Clarke, T. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci* **22,** 343–352 (2019)

[19] Pasman, J.A., Verweij, K.J.H., Gerring, Z. *et al.* GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal effect of schizophrenia liability. *Nat Neurosci* **21,** 1161–1170 (2018).

[20] Plana-Ripoll O, Pedersen CB, Holtz Y, et al. Exploring Comorbidity Within Mental Disorders Among a Danish National Population. JAMA Psychiatry. 2019;76(3):259–270.

[21] Mu-Lin Chiu, Chi-Fung Cheng, Wen-Miin Liang, Pen-Tang Lin, Trong-Neng Wu, and Chiu-Ying Chen, The Temporal Relationship between Selected Mental Disorders and Substance-Related Disorders: A Nationwide Population-Based Cohort
Study, Hindawi Psychiatry Journal Volume 2018, Article ID 5697103

[22] Cross-Disorder Group of the Psychiatric Genomics Consortium, Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders, 2019, Cell179, 1469–1482

[23] Renato Polimanti, Joel Gelernter, Widespread signatures of positive selection in common risk alleles associated to autism spectrum disorder, Plos Genetics, February 10, 2017

[24] Esteller-Cucala, P., Maceda, I., Børglum, Lao O., A.D. et al. Genomic analysis of the natural history of attention-deficit/hyperactivity disorder using Neanderthal and ancient Homo sapiens samples. Sci Rep 10, 8622 (2020).

[25] Daiki X. Sato Masakado Kawata, Positive and balancing selection on SLC18A1 gene associated with psychiatric disorders and human-unique personality traits, Wiley Online Library, 21 August 2018

[26] Zachary Durisko, PhD1,2, Benoit H. Mulsant, MD, MS1,3, Kwame McKenzie, BM, MRCPsych1,3,4, and Paul W. Andrews, PhD, JD2, Using Evolutionary Theory to Guide Mental Health Research, 016, Vol. 61(3) 159-165

[27] Srinivasan S, Bettella F, Mattingsdal M, et al. Genetic Markers of Human Evolution Are Enriched in Schizophrenia. Biol Psychiatry. 2016;80(4):284-292.

[28] Polimanti R, Kayser MH, Gelernter J. Local adaptation in European populations affected the genetics of psychiatric disorders and behavioral traits. Genome Med. 2018;10(1):24. Published 2018 Mar 26.

[29] Durisko Z, Mulsant BH, McKenzie K, Andrews PW. Using Evolutionary Theory to Guide Mental Health Research. Can J Psychiatry. 2016;61(3):159-165.

[30] Chenxing Liu, Ian Everall, Christos Pantelis and Chad Bousman, Interrogating the Evolutionary Paradox of Schizophrenia: A Novel Framework and Evidence Supporting Recent Negative Selection of Schizophrenia Risk Alleles, Frontiers in Genetics, April 2019, Volume 10, Article 389.

[31] Evangelos Vassos, Paul O'Reilly, Cathryn Lewis, MEASURING EVOLUTIONARY PRESSURE IN A POLYGENIC FRAMEWORK FOR COMPLEX DISEASES, European Neuropsychopharmacology, Volume 29, Supplement 3, 2019

[32] Power RA, Kyaga S, Uher R, et al. Fecundity of Patients with Schizophrenia, Autism, Bipolar Disorder, Depression, Anorexia Nervosa, or Substance Abuse vs Their Unaffected Siblings. JAMA Psychiatry. 2013;70(1):22–30.

[33] John Wakely, Coalescent Theory: An Introduction

[34] M.A.Rosales-Reynosoa, C.I.Juárez-Vázqueza, P.Barros-Núñez, Evolution and genomics of the human brain, Neurología (English Edition), Volume 33, Issue 4, May 2018, Pages 254-265

[35] Philipp Khaitovich, Kun Tang, Henriette Franz, Janet Kelso, Ines Hellmann, Wolfgang Enard, Michael Lachmann, and Svante Pääbo, Positive selection on gene expression in the human brain, Current Biology Vol 16 No 10

[36] Faraone SV, Rostain AL, Blader J, et al. Practitioner Review: Emotional dysregulation in attention-deficit/hyperactivity disorder - implications for clinical recognition and intervention. J Child Psychol Psychiatry. 2019;60(2):133-150.

[37] Ghirardi L, Pettersson E, Taylor MJ, et al. Genetic and environmental contribution to the overlap between ADHD and ASD trait dimensions in young adults: a twin study. Psychol Med. 2019;49(10):1713-1721