DETECTION OF RECENT POSITIVE SELECTION IN COMMON RISK ALLELES ASSOCIATED TO AUTISM SPECTRUM DISORDER USING SINGLETON DENSITIES

Experimental Bachelor's Project in Molecular Medicine

Department of Biomedicine

Aarhus University

06.15.2021

Cæcilia Lind Skov-Jensen

201806070

Supervisors:

Thomas Damm Als

Jakob Grove

ABSTRACT

Autism spectrum disorder (ASD) is a highly genetic and heterogenous neurodevelopment disorder affecting approximately one in 160 children worldwide despite having negative fitness effects. This creates an evolutionary paradox, but to date no evidence has been found for positive selection in ASD-associated genes. However, loci associated to other psychiatric disorders such as schizophrenia are known to have been subjects of positive selection.

One method to detect recent positive selection is the Singleton Density Score (SDS), which uses whole-genome sequence (WGS) data to detect very recent allele frequency change on preexisting variants across several SNPs.

Applying this method to Faroese WGS data, signals of selection at ASD-associated alleles were identified, suggesting that alleles associated with ASD have been under recent positive selection.

ACKNOWLEDGEMENT

The work presented in this Bachelor’s Project was carried out between February and June 2021 under the supervision of Associate Professor Jakob Grove and Associate Professor Thomas Damm Als at the Department of Biomedicine at Aarhus University.

A genuine thank you to all the people that were part of this project and helped me during my work.
First of all, I would like to thank my supervisors Jakob Grove and Thomas Damm Als for guidance and enthusiasm throughout the entire project as well as for investing so many hours into this project. Thank you for giving me free rein to design my experiment as I would like, and especially, thank you for your patience through technical difficulties and several hundreds of emails back and forth during this strange situation with COVID-19.
Finally, I would like to thank my family and friends for being supportive and understanding during this intense time of working on this project.

TABLE OF CONTENTS

ABSTRACT 2

ACKNOWLEDGEMENT..3

INTRODUCTION 5

AUTISM SPECTRUM DISORDER (ASD) 5

EVIDENCE OF SELECTION FOR PSYCHIATRIC DISORDERS 5

METHODS FOR DETECTING SELECTION 6

METHODS...7

DATA PREPROCESSING 7

OBTAINING INPUTS FOR SDS ANALYSIS 8

COMPUTING RAW SDS SCORES (RSDS) 8

FURTHER FILTRATION 9

CALCULATING STANDARDIZED SDS SCORES 10

RESULTS 11

ANALYSIS OF PROBABILITY DISTRIBUTION 11

VISUALIZATION OF STRONG SDS SIGNALS 12

ANALYSIS OF SDS ENRICHMENT FOR ASD VARIANTS 13

DISCUSSION...16

SUMMARY OF RESULTS 16

WEAKNESSES OF THE STUDY 17

REFERENCES 18

APPENDIX A...20

TABLE A1: LONG-RANGE LD REGIONS 20

TABLE A2: MAPPING OF SIGNIFICANT GENES TO THE CLOSEST GENOMIC FEATURES 21

TABLE A3: TOP 20 RANKED ASD-GENES 22

INTRODUCTION

Autism spectrum disorder (ASD)

Autism spectrum disorder (ASD) is a heterogenous neurodevelopment disorder that is characterized by impairment in social interaction and communication skills, as well as atypical behaviors such as difficulty with transitioning from one activity to another and difficulty to focus on details[1].

It is a complex condition, meaning that both genetic factors and environmental factors play a role in developing the disorder. The environmental factors are thought to play a minor role, however ASD is known to be highly heritable with a heritability estimate in Scandinavia of 0.5-0.6[2], indicating that genetic factors play a major role.

Various ASD-associated genes have been identified using genome-wide association studies (GWAS), the latest being Satterstrom et al. (2020)[3], that identified 102 genes implicated in risk for ASD.

There is a lot of overlap of risk variants between different psychiatric disorders, such as schizophrenia, major depressive disorder (MDD) and ASD. Several genes with known effects on schizophrenia and MDD are also considered risk genes for ASD. For example, variants in the genes NEGR1 (Chr 1: 72729142), PTBP2 (Chr 1:96561801) and KMT2E (Chr 7:104744219) have been found to be significant in ASD as well as in depression and schizophrenia, respectively[4].

Evidence of selection for psychiatric disorders

Despite having negative fitness effects, ASD still persists in society affecting approximately 1 in 160 children worldwide[1]. Though, it seems that no clear evidence of selection has yet been discovered for ASD-associated genes.

However, evidence of positive selection has been found for loci associated to other psychiatric disorders such as schizophrenia. According to Li et al. (2016)[5] a schizophrenia risk variant (r13107325) in the SLC39A8 gene has experienced positive selection. This allele does however not immediately seem to have known associations with ASD. Another recent study conducted by Fujito et al. (2018)[6] detected positive selection on the schizophrenia-associated ST8SIA2 gene in post-glacial Asia and according to Sato et al. (2016)[7] a SNP (rs3784730) of the ST8SIA2 gene has known effects on ASD.

Methods for detecting selection

Former methods for detecting signs of selection, like the extended haplotype homozygosity (EHH)[8] and the integrated haplotype score (iHS)[9], focused on searching for sweeps where de novo mutations under strong positive selection sweep through a population towards fixation[10]. However, these methods are not powerful and does not have specificity for recent selection[10].

A more recent method for detecting signatures of recent selection is the Singleton Density Score (SDS)[10], which uses whole-genome sequence (WGS) data to detect very recent change in allele frequency on preexisting variants across several SNPs. If a locus was subject to recent selection, it would result in shorter tip-branch length for the favored allele and therefore a reduction in the number of singletons (i.e., variants that only occur in a single individual and only one copy) around the favored allele. Therefore, haplotypes with the preferred allele tend to carry fewer singleton mutations[10]. By measuring the distances from the test SNP of interest to the nearest singletons on either side, the amount of singleton mutations can be estimated. The larger the distances to the nearest singletons are, the more it suggests that the SNP is subject to selection[10].

Hereafter, a distribution of distances for the three different genotypes at each of the test SNPs are made to calculate a maximum likelihood estimate of the log-ratio of mean tip-branch lengths (i.e., a raw SDS score) for the derived allele versus the ancestral allele[10]. This score is then standardized to mean = 0 and variance = 1 within intervals of derived allele frequency. If SDS > 0 it suggests an increased frequency of the derived allele and thereby signs of positive selection for that allele[10].

METHODS

Data preprocessing

Whole genome sequencing (WGS) data from 268 individuals (106 cases with schizophrenia and 162 controls) from the Faroe Islands was prepared and processed by my supervisor as described below: Using an Illumina Sequencing Platform, the individuals were whole genome sequenced six times and hereafter alignment, variant call and recalibration was conducted using the Genome Analysis Toolkit (GATK, https://gatk.broadinstitute.org/hc/en-us)[11]. In total approximately 10 million single nucleotide variants were identified based on multi-sample calling of the aligned sequences. Variant, sample and genotype quality-control was completed using Hail[12] with the following thresholds: Variant call-rate > 0.90 and sample call-rate > 0.90. The genotypes were filtered for allelic balance/allele depth. If variants did not pass the GATK[11] variant quality score recalibration (VQSR), the variants were removed. Hereafter, the following genotype filters were used to remove calls of low quality: (1) Homozygous reference calls with < 90% reads supporting the reference allele, (2) Homozygous variant calls with < 90% reads supporting the alternate allele, (3) Heterozygote calls with < 25% reads supporting the alternate allele. After using these genotype filters, samples that had a call-rate < 0.90 and variants with a call-rate < 0.99 were excluded. In addition, variants with a deviation from the Hardy-Weinberg proportions with P-value < $1x10^{-6}$ were removed.

After the exclusion of regions with high linkage disequilibrium (LD)[13], the genotypes were then pruned down to a set of approximately 20,000 markers. This was done by LD pruning in a sliding window approach as well as filtering for MAF > 0.01 using PLINK 1.9[14, 15] with an R^2 limit of 0.1, window size of 5000 and step size of 300. Pairs of individuals were identified with $\hat{\pi} < 0.2$ using PLINK's identity by state analysis and one individual from each pair was excluded at random. A Principal Component Analysis (PCA)[16, 17] was conducted using smartPCA implemented in EIGENSOFT version 6.1.4[18] on the relatedness-pruned set of individuals, subsequently projecting all individuals onto those eigenvectors based on their genotypes and removing any clear PCA-outliers. In total 189 individuals and 4,120,372 SNPs remained in the WGS data after above preparations.

The ASD genome-wide association study summary statistics were based on results from a meta-analysis of 18,381 ASD cases and 27,969 controls of European ancestry from the iPSYCH autism sample[19]. A quality control by my supervisor was performed on the data to exclude samples with an extremely high number of singleton variants that would otherwise be considered ethnic outliers[20]. After the quality control the summary statistics consisted of 9,112,386 SNPs.

Obtaining inputs for SDS analysis

The SDS method required the following set of input files that were prepared by my supervisor for each chromosome from the Faroese WGS data. Following Field et al. (2016)[10], test SNPs for the test SNP file were defined as focal SNPs with a derived allele frequency between 0.05-0.95 and a Hardy-Weinberg equilibrium (HWE) P-value > 1x10^{-6}. The test SNP file also contained information about the SNP-genotypes of each individual. For every test SNP in each individual the nearest upstream and downstream singletons and their locations were identified creating a singleton file. The SDS method furthermore required a singleton observability file, which contained the probabilities that singletons were observed in the data. In Field et al. (2016)[20] this was obtained by taking the reported sequencing depths, while in this study the average sequencing depth per individual was used. The file with gamma shapes contained an estimate of the ratio between the mean squared and variance of the ideal tip-branch length distribution[20]. The gamma shapes were estimated by using simulations from a demographic model for a central/northern European population presented by Tennessen et al. (2012)[21]. Furthermore, it was required to set the genetic boundaries between which the analyses were carried out. These were set to the starting point of 1 and end points equal to the size of each chromosome[20].

Computing raw SDS scores (rSDS)

Raw SDS scores were computed using the scripts provided by Field[22]. The raw SDS scores were calculated using the distances between the nearest singletons upstream and downstream from each test SNP. The distribution of singleton distances as a function of genotype at each SNP were used to compute a maximum likelihood estimate for the log-ratio of mean tip-branch lengths of the derived versus ancestral alleles[10], which corresponded to the raw SDS score.

After obtaining the raw SDS scores, the phenotype associated allele for each SNP was defined from the GWAS summary statistics based on the odds ratio (OR). If OR was above or equal to 1 the A1 allele was defined as the positive phenotype associated allele (and thereby the derived allele) and the A2 allele was defined as the negative associated allele. If OR was below 1 the A2 allele was defined as the positive phenotype associated allele (and thereby the derived allele) and the A1 allele was defined as the negative associated allele.

The GWAS-markers were then matched to markers from the Faroese WGS data with raw SDS scores by merging the two datasets based on chromosome number and SNP position (in total 3,980,972 SNPs were merged). This was done to polarize the raw SDS scores such that the positive associated alleles from the GWAS summary statistics represented the derived allele in the Faroese WGS data and the negative associated alleles represented the ancestral allele as they do in Field et al.[10]. Until now the reference allele in the WGS data was the ancestral allele and the alternative allele was the derived allele. If the positive associated allele from the GWAS corresponded to the ancestral allele in the Faroese WGS data, the raw SDS score was inverted.

Further filtration

After the computation of raw SDS scores, extreme signals (i.e., signals with raw SDS scores up to +49 and down to -67) were observed in the SDS dataset. The SNPs with extreme raw SDS showed a pattern of genotype counts of 0 for one or both homozygous genotypes (i.e., nG0 = 0 and/or nG2 = 0), while there was a relatively high amount of heterozygous and therefore a relatively large deviation from the Hardy-Weinberg proportions. In the quality control the data was filtered for HWE P-value < $1x10^{-6}$, but that did not seem to be enough to avoid extreme values. Therefore, further filtration for SNPs with HWE P-value < $1x10^{-5}$ was necessary, which further removed 95,641 SNPs. Afterwards the SNPs with genotype counts under 2 (i.e., nG0 < 2 and nG1 < 2 and nG2 < 2) were excluded from the SDS dataset to avoid the extreme raw SDS scores as these could influence the distribution of raw SDS. This further removed 282,751 SNPs from the SDS dataset resulting in a total of 3,602,580 SNPs after the filtrations.

After this filtration long-range LD regions[13] (Table A1 in Appendix A) were excluded from the calculation of standardized SDS scores (in total 110,448 SNPs were excluded from the calculation). This was done to avoid that the regions with long-range LD would stand out from the rest of the genome because of bias from extreme signals[20]. If they remained a part of the standardization, it could affect all markers (due to the standardization). It was therefore preferable to exclude them so that they would not affect the rest of the genome.

Calculating standardized SDS scores

The raw SDS scores were then standardized within derived allele frequency (DAF) bins of 10% (e.g., 0.1-0.2). The standardization was done by subtracting the mean DAF bin score and dividing by the DAF bin standard deviation[20].

Subsequently two-sided SDS P-values for each standardized SDS score were obtained from a standard normal distribution using the absolute value of the standardized scores.
However, it seemed that three extreme signals were still observed with SDS P-value = 0, so in order to avoid bias from these extreme signals, the three P-values were manually set to the lowest value possible in R (P-value = $2.22\text{x}10^{-16}$). In table 1 below an overview of the SDS data is shown.

Table 1: SDS data

Chr	Position	SNP	A1	A2	Ancestral	Derived	Ass. allele	P-value	rSDS	SDS	SDS P-value
1	100000012	rs10875231	T	G	G	T	T	0.8475	-0.3266	-0.9352	0.3497
1	100000827	rs6678176	T	C	C	T	C	0.9566	0.3683	1.1342	0.2567
1	100000843	rs78286437	T	C	T	C	T	0.7694	0.1226	0.2204	0.8255
1	100001201	rs76909621	T	G	G	T	T	0.0714	-0.1053	-0.2531	0.8002
1	100099170	rs7529598	A	C	C	A	A	0.7036	0.6279	2.0624	0.0392
1	100002490	rs78642210	T	C	C	T	C	0.8240	0.4205	0.7886	0.4303
1	100002713	rs77140576	T	C	C	T	T	0.0222	-0.1216	-0.2920	0.7703
1	100002714	rs113470118	A	G	A	G	A	0.7826	0.0945	0.1667	0.8676
1	100002882	rs7545818	T	G	T	G	T	0.9886	0.5344	1.6524	0.0985
1	100083263	rs12722988	A	G	A	G	A	0.8362	1.026	2.4455	0.0145

Table 1 | Standardized SDS scores. Section of the SDS data table showing chromosome number, SNP position, SNP name, allele 1, allele 2, ancestral allele, derived allele, ASD associated allele, ASD GWAS P-value, raw SDS score, standardized SDS score and SDS P-value.

RESULTS

Analysis of probability distribution

Because the raw SDS scores were standardized, it was expected that the SDS scores and the associated SDS P-values were normally distributed. To test whether the distribution of SDS P-values was normally distributed a quantile-quantile (Q-Q) plot was made (Fig. 1).

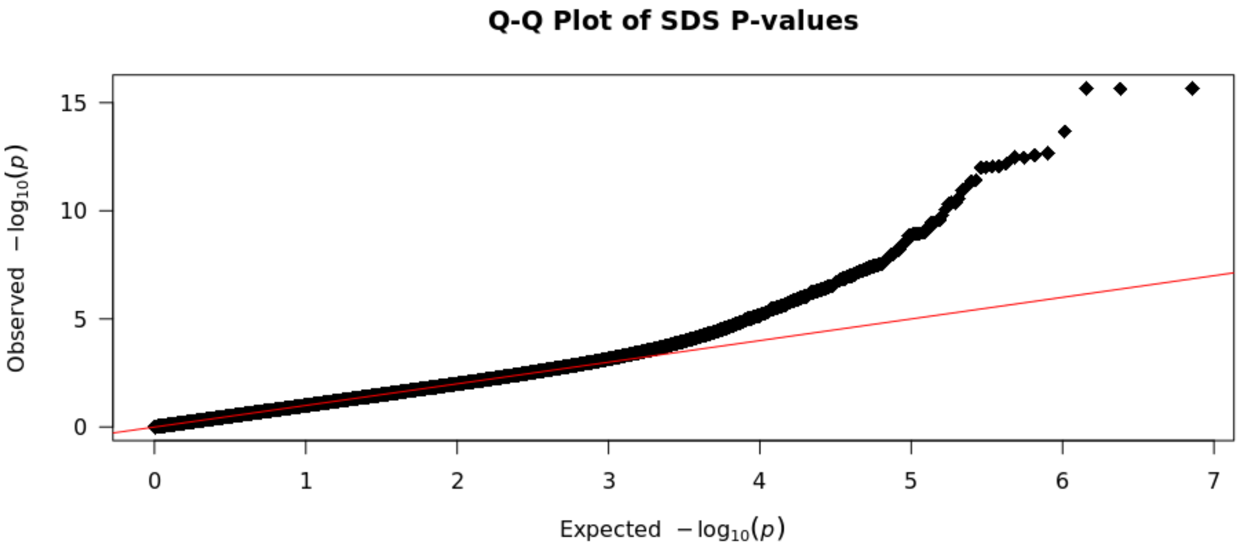

Fig. 1 | Quantile-quantile plot of SDS P-values. The observed SDS P-values (y-axis) were plotted against the expected SDS P-values under the null hypothesis (x-axis). The red diagonal line indicates the pattern under the null hypothesis.

The Q-Q plot indicated that the distribution of SDS P-values did not perfectly match with the pattern under the null hypothesis that the SDS P-values were normally distributed. Some observed SDS P-values were clearly more significant than expected under the null hypothesis as seen by the points moving towards the y-axis. This deviation suggested a sizable inflation that could be due to LD.

LD-score regression, which handles LD, was subsequently performed on the data by my supervisor, which actually showed that the data was deflated, i.e., that the signals were weaker than it was expected. This was consistent with the small sample size and that we did not expect to observe strong signals, and therefore LD seemed to be a good explanation for the inflation shown in Fig. 1.

Visualization of strong SDS signals

To test whether any SNPs showed strong SDS signals and therefore signals of selection, a Manhattan plot was made. The distribution of SDS P-values is shown in Fig. 2 below.

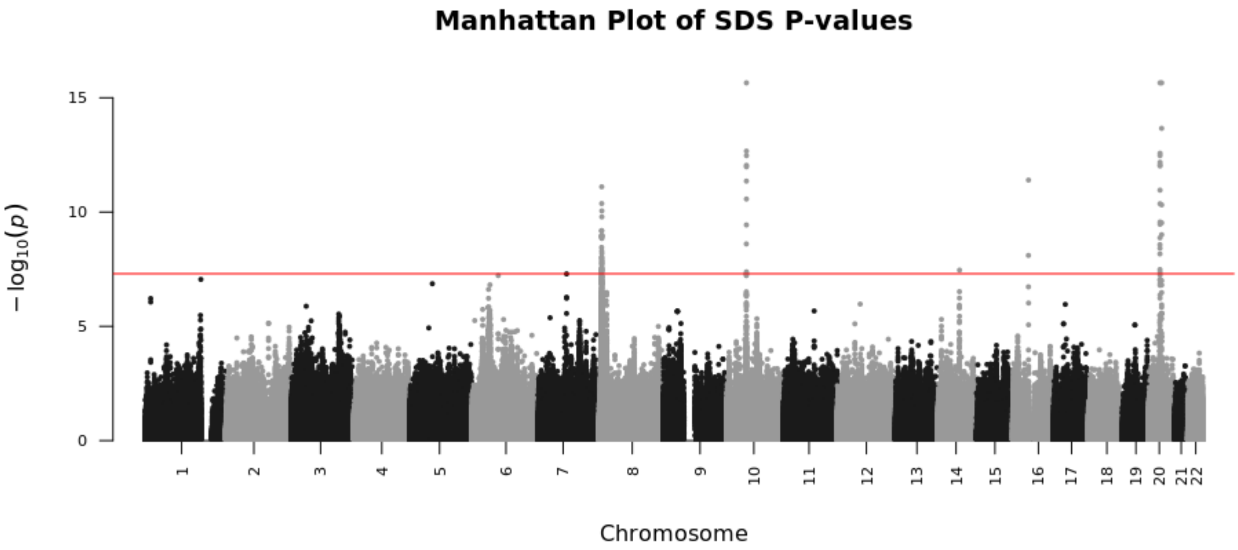

Fig. 2 | Manhattan plot of SDS P-values. The x-axis shows the genomic position (chromosome 1-22), and the y-axis shows the statistical significance as -log10(SDS P-value) of the SDS. Genome-wide significance line was set to -log10(5x10^{-8}). SDS P-values are two-sided tail probabilities of standard normal distribution.

The plot indicated five regions of genome-wide significance (SDS P-value < 5x10^{-8}), and 72 SNPs of significance was observed (SNPs with SDS P-value < 5x10^{-8}).

A list of 19,427 genes from Human Genome Resources at NCBI[23] was obtained and used to look for surrounding genes for the significant SNPs. This was done by mapping the significant SNPs to the gene in closest proximity based on the NCBI gene list using the Map2NCBI package. This resulted in six different genes: MCPH1, LOC101929373, GPHN, LOC101929605, ZNF337 and DEFB115. However, neither of those genes seemed to have known effects on ASD or any other psychiatric disorder. A section of the mapping results is shown in Table A2 in Appendix A.

Analysis of SDS enrichment for ASD variants

To test whether alleles with known phenotypic effects tended to be targets of selection, a list of 102 risk genes associated to autism spectrum disorder was obtained from Satterstrom et al. (2020)[3].

The 3,602,580 SNPs from the SDS dataset were then merged with the 19,427 genes from the NCBI gene list based on chromosome number and SNP position resulting in variants found in 13,725 different genes. The mean SDS score was then calculated for each of the genes as well as the two-sided SDS P-value for each mean SDS score obtained from a standard normal distribution. The SDS P-values were calculated as only one SNP per gene (sd = 1), because if the normal distribution was scaled with sd = 1/sqrt(*n*) it would be assumed that there was *n* number of independent SNPs in each gene, which was most unlikely due to LD within the genes.

The 13,725 genes were then merged with the 102 ASD-associated genes from Satterstrom et al. (2020)[3] resulting in 80 gene matches. Principally, all the 102 ASD genes should have been found in the data, but the explanation for why variation was only identified in 80 ASD genes in the Faroese data could be because of all the filtering that was conducted on the SDS dataset or because of the small sample size and the MAF > 0.01 filtering.

The top 20 ranked genes out of the 80 ASD genes were selected based on the 20 genes with the lowest Q-values from Satterstrom et al. (2020)[3] (Table A3 in Appendix A).

In addition, the mean SDS value for the entire ASD gene set consisting of 80 genes as well as the SDS P-value was calculated to investigate whether the entire ASD-associated gene set was target of selection. The SDS P-value was calculated using a normal distribution with mean = 0 and sd = 1/sqrt(80). Here it was assumed that there was no LD in between the genes and that the SNPs therefore were independent in each gene. In Fig. 3 the SDS signals for the selected set of genes with known effects on ASD are shown.

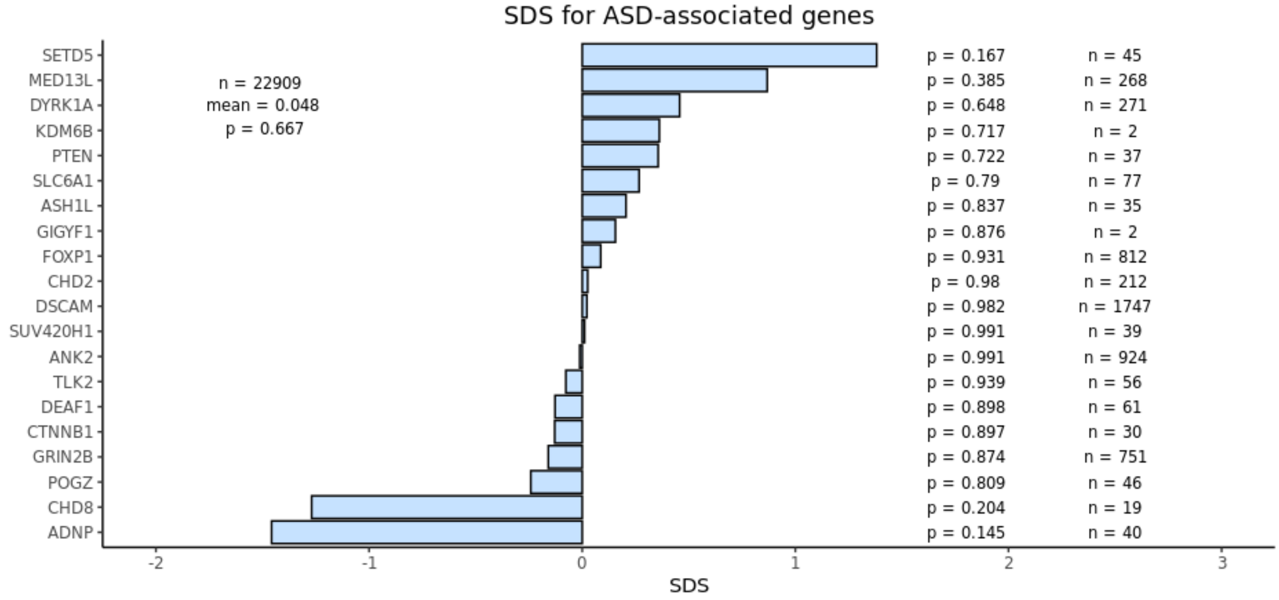

Fig. 3 | Bar plot of SDS signals for the top 20 ranked ASD genes. The x-axis shows the SDS score, while the y-axis shows the ASD-associated gene name. SDS P-value and number of SNPs for each gene are shown to the right. Mean SDS value as well as the associated SDS P-value of the whole 80 ASD gene set is shown in upper left corner.

The bar plot indicated that 12 of the 20 genes (SETD5, MED13L, DYRK1A, KDM6B, PTEN, SLCA1, ASH1L, GIGYF1, FOXP1, CHD2, DSCAM and SUV420H1) had positive mean SDS values (which indicated positive selection) and 8 of the 20 genes (ANK2, TLK2, DEAF1, CTNNB1, GRIN2B, POGZ, CHD8 and ADNP) had negative mean SDS scores, which suggested negative selection. However, none of the SDS P-values for the genes seemed to be significant (SDS P-value < 0.05), which implied that the probability of the results being random was very high.

The mean SDS value for the entire ASD gene set of 0.048 indicated that the ASD-associated genes overall had experienced positive selection, however the SDS P-value of 0.667 implied that there was a high probability that the mean SDS value occurred by chance.

Finally, a distribution of trait-SDS scores (tSDS) was made for all the ASD-associated SNPs and compared to the distribution of all SNPs in the SDS dataset (Fig. 4). The ASD-associated SNPs (22,909 SNPs) were filtered for significance with ASD GWAS P-value < 0.05 and only these SNPs were used for the distribution. In total 1,462 SNPs were used in the distribution of trait-SDS scores, and 3,602,580 SNPs were used in the null distribution (i.e., the distribution of all SNPs in the SDS dataset).

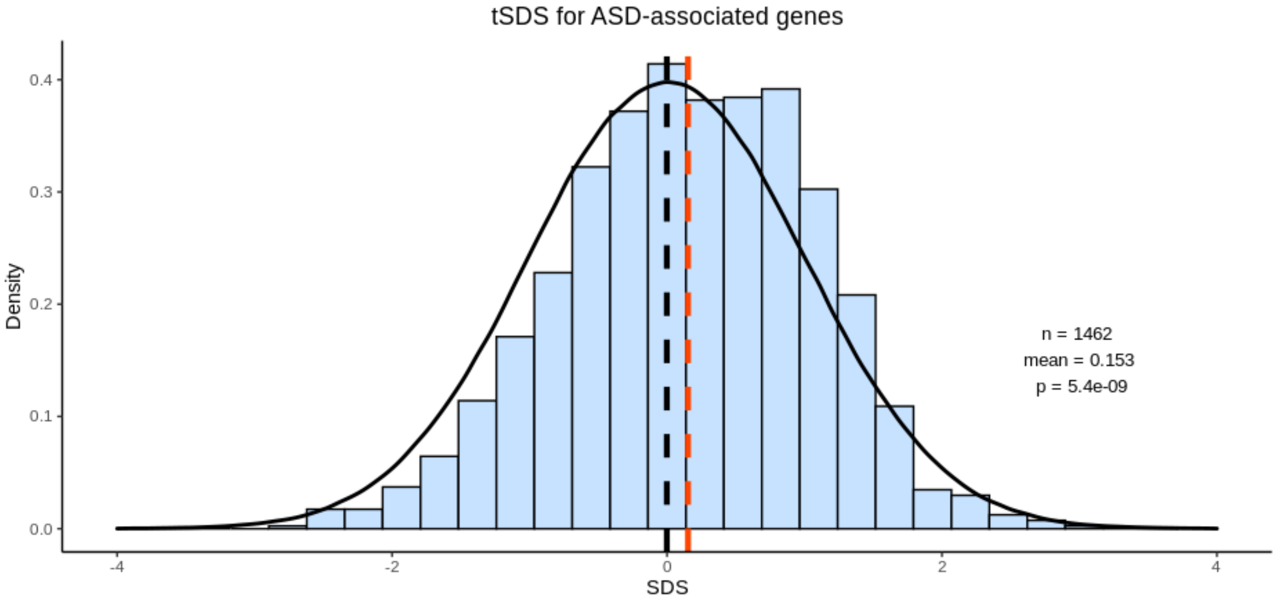

Fig. 4 | Normal distribution of trait-SDS scores at 1,462 ASD-associated SNPs. The x-axis shows the SDS score, and the y-axis shows the density. The black full line represents the null distribution, and the blue bars represent the distribution of the ASD-associated SNPs. The black dashed line represents the mean SDS for the null distribution and the red dashed line represents the mean SDS score for the 1,462 ASD-associated SNPs.

The mean trait-SDS score of 0.153 indicated that the ASD-associated alleles have been subjects to positive selection and since the SDS P-value of 5.4x10^{-9} was significant (SDS P-value < 0.05), it indicated that the probability of obtaining the mean trait-SDS score at random was not high.

DISCUSSION

Summary of results

The Faroese genomic dataset was analyzed to identify signatures of recent positive selection, and to determine whether the data contained a signal of polygenic selection acting on autism spectrum disorder (ASD). The results from the Manhattan plot indicated that 5 regions were targets of selection as these regions contained significant SDS P-values < $5x10^{-8}$ (Fig. 2). However, none of these SNPs under selection had known effects on ASD.

The bar plot (Fig. 3) indicated that some of the ASD-associated genes were subjects to positive selection and that some ASD-associated genes seemed to be subjects to negative selection. However, the SDS P-value = 0.667 implied that these results were most likely due to chance.

The distribution of trait-SDS scores for all the ASD-associated SNPs revealed that the ASD-associated alleles have been subjects to positive selection overall and the associated SDS P-value = $5.4x10^{-9}$ implied that it was most unlikely that the result occurred by chance.

The mean SDS value (mean = 0.153) was relatively high, indicating that the ASD-associated alleles have experienced recent positive selection. This is an interesting result because of the negative influence ASD is known to have on the survival and the fecundity of individuals with ASD, with a fertility ratio of respectively 0.25 and 0.48 for men and women with autism[24]. One should therefore think that the ASD-associated alleles work early in life and that there would be stronger selection against those alleles (i.e., negative selection on them). The result could therefore indicate that the positive selection detected is caused by other things. It is for example known that there is positive genetic correlation between ASD and other traits such as IQ and educational attainment[4], which means that many of the alleles that increase the risk of ASD broadly overlap with alleles for high IQ and alleles that make one more likely to take a long education. Thereby the positive selection detected for ASD-associated alleles is not necessarily because these alleles occur in individuals with ASD but could be because these ASD-associated alleles occur in individuals that do not have ASD but have for example high IQ and are carriers of the ASD-associated allele. In light of this, it is not unreasonable to find positive selection for these alleles.

Weaknesses of the study

One of the obvious disadvantages of this study is its very small population size (189 individuals) compared to that of Field et al. (2016) (3195 individuals)[10], which has larger statistic power. The small sample size influences the chance to observe differences between SDS and zero. Because of the small sample size, the difference must be greater in order for us to be able to detect it, while with a larger sample size, it would be possible to detect smaller differences. However, the SDS P-value should reflect whether one can believe the difference or not regardless of sample size.

Furthermore, it is not certain that singletons are really singletons in such a small sample as this one. It is very likely that other individuals in the rest of the Faroese population (that was not included in this sample) possess the variant. Thereby, it does not take much to observe a singleton in this sample, which could affect the results.

However, it speaks to the advantage of the method that it is able to detect selection in such a small sample like this one. But of course, we do not know for sure if the result we have detected is true until we have replicated it in a larger study.

For further investigation of the reliability of the positive selection detected for ASD-associated alleles, it could be interesting to test it on a larger sample and see if it gives the same results. This could be done by either simulating data or by using the UK10K data as they did in Field et al. (2016)[10] and then test the data with the alleles turned after ASD to see if it replicated the results.

REFERENCES

[1] WHO. "Autism spectrum disorders." https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders (accessed 06/01, 2021).

[2] G. Ramaswami and D. H. Geschwind, "Genetics of autism spectrum disorder," (in eng), *Handb Clin Neurol,* vol. 147, pp. 321-329, 2018, doi: 10.1016/b978-0-444-63233-3.00021-x.

[3] F. K. Satterstrom *et al.*, "Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism," (in eng), *Cell,* vol. 180, no. 3, pp. 568-584.e23, Feb 6 2020, doi: 10.1016/j.cell.2019.12.036.

[4] J. Grove *et al.*, "Identification of common genetic risk variants for autism spectrum disorder," (in eng), *Nat Genet,* vol. 51, no. 3, pp. 431-444, Mar 2019, doi: 10.1038/s41588-019-0344-8.

[5] M. Li *et al.*, "Recent Positive Selection Drives the Expansion of a Schizophrenia Risk Nonsynonymous Variant at SLC39A8 in Europeans," (in eng), *Schizophr Bull,* vol. 42, no. 1, pp. 178-90, Jan 2016, doi: 10.1093/schbul/sbv070.

[6] N. T. Fujito *et al.*, "Positive selection on schizophrenia-associated ST8SIA2 gene in post-glacial Asia," (in eng), *PLoS One,* vol. 13, no. 7, p. e0200278, 2018, doi: 10.1371/journal.pone.0200278.

[7] C. Sato, M. Hane, and K. Kitajima, "Relationship between ST8SIA2, polysialic acid and its binding molecules, and psychiatric disorders," (in eng), *Biochim Biophys Acta,* vol. 1860, no. 8, pp. 1739-52, Aug 2016, doi: 10.1016/j.bbagen.2016.04.015.

[8] P. C. Sabeti *et al.*, "Detecting recent positive selection in the human genome from haplotype structure," (in eng), *Nature,* vol. 419, no. 6909, pp. 832-7, Oct 24 2002, doi: 10.1038/nature01140.

[9] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard, "A map of recent positive selection in the human genome," (in eng), *PLoS Biol,* vol. 4, no. 3, p. e72, Mar 2006, doi: 10.1371/journal.pbio.0040072.

[10] Y. Field *et al.*, "Detection of human adaptation during the past 2000 years," (in eng), *Science,* vol. 354, no. 6313, pp. 760-764, Nov 11 2016, doi: 10.1126/science.aag0776.

[11] M. A. DePristo *et al.*, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," (in eng), *Nat Genet,* vol. 43, no. 5, pp. 491-8, May 2011, doi: 10.1038/ng.806.

[12] Hail-Team. "Hail 0.2.13-81ab564db2b4." https://github.com/hail-is/hail (accessed 06/01, 2021).

[13] A. L. Price *et al.*, "Long-range LD can confound genome scans in admixed populations," (in eng), *Am J Hum Genet,* vol. 83, no. 1, pp. 132-5; author reply 135-9, Jul 2008, doi: 10.1016/j.ajhg.2008.06.005.

[14] S. Purcell *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses," (in eng), *Am J Hum Genet,* vol. 81, no. 3, pp. 559-75, Sep 2007, doi: 10.1086/519795.

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

[16] D. Reich, A. L. Price, and N. Patterson, "Principal component analysis of genetic data," (in eng), *Nat Genet,* vol. 40, no. 5, pp. 491-2, May 2008, doi: 10.1038/ng0508-491.

[17] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," (in eng), *Nat Genet,* vol. 38, no. 8, pp. 904-9, Aug 2006, doi: 10.1038/ng1847.

[18] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," (in eng), *PLoS Genet,* vol. 2, no. 12, p. e190, Dec 2006, doi: 10.1371/journal.pgen.0020190.

[19] C. B. Pedersen *et al.*, "The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders," (in eng), *Mol Psychiatry,* vol. 23, no. 1, pp. 6-14, Jan 2017, doi: 10.1038/mp.2017.196.

[20] Y. Field *et al.*, "Supplementary Materials for Detection of human adaptation during the past 2000 years," *Science,* 2016, doi: 10.1126/science.aag0776.

[21] J. A. Tennessen *et al.*, "Evolution and functional impact of rare coding variation from deep sequencing of human exomes," (in eng), *Science,* vol. 337, no. 6090, pp. 64-9, Jul 6 2012, doi: 10.1126/science.1219240.

[22] Y. Field. "R code to compute the Singleton Density Score (SDS)." https://github.com/yairf/SDS (accessed 05/27, 2021).

[23] NCBI. "Human Genome Resources at NCBI." https://www.ncbi.nlm.nih.gov/genome/guide/human/?fbclid=IwAR3vLEoYx8MAwkqRALFYzvw1RVWB9JGd5VNJu2lNmBS-xBIBQ8hP6ORS1eM#download (accessed 05/29, 2021).

[24] R. A. Power *et al.*, "Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings," (in eng), *JAMA Psychiatry,* vol. 70, no. 1, pp. 22-30, Jan 2013, doi: 10.1001/jamapsychiatry.2013.268.

APPENDIX A

Table A1: Long-range LD regions

Chromosome	Start	End
1	48227413	52227412
2	86146489	101133568
2	134783530	138283530
2	183291755	190291755
3	47524996	50024996
3	83417310	86917310
3	88917310	96017310
5	44464243	50464243
5	97972100	100472101
5	128972101	131972101
5	135472101	138472101
6	25392021	33392022
6	56892041	63942041
6	139958307	142458307
7	55032506	66362565
8	7962590	11962591
8	42880843	49837447
8	111930824	114930824
10	36959994	43679994
11	46043424	57243424
11	87860352	90860352
12	33108733	41713733
12	111015617	113515617
20	32536339	35066586

Table A1 | Long-range LD regions. A list of the long-range LD regions[13] with chromosome number, start and end positions that were excluded from the calculation of standardized SDS.

Table A2: Mapping of significant genes to the closest genomic features

Chr	Position	Gene	Start	End	Distance	Inside or outside gene?
8	5672791	MCPH1	6264113	6501140	591322	Nearest feature is > 25,000 bp Before Feature
8	6254028	MCPH1	6264113	6501140	10085	Marker is > 5000 bp <=25000 bp Before Feature
8	6260430	MCPH1	6264113	6501140	3683	Marker is > 2500 bp <=5000 bp Before Feature
8	6510150	MCPH1	6264113	6501140	9010	Marker is > 5000 bp <=25000 bp After Feature
10	42418884	LOC101929373	42737971	42772290	319087	Nearest feature is > 25,000 bp Before Feature
14	67378055	GPHN	66974125	67648525	0	Yes, Inside Gene
16	34186994	LOC101929605	33298270	33318787	868207	Nearest feature is > 25,000 bp After Feature
16	34319104	LOC101929605	33298270	33318787	1000317	Nearest feature is > 1 Mb After Feature
20	25858431	ZNF337	25654744	25677515	180916	Nearest feature is > 25,000 bp After Feature
20	29567314	DEFB115	29845467	29847435	278153	Nearest feature is > 25,000 bp Before Feature
20	29590783	DEFB115	29845467	29847435	254684	Nearest feature is > 25,000 bp Before Feature

Table A2 | Section of the mapping of significant genes to the closest genomic features.

Nearest feature is > 25000 bp Before Feature means that the closest feature is located before the marker position and is more than 25,000 bp from the marker.

Marker is > 5000 bp <=25000 bp Before Feature means that the closest feature is located before the marker position and is between 5,000 bp and 25,000 bp from the marker.

Marker is > 2500 bp <=5000 bp Before Feature means that the closest feature is located before the marker position and is between 2,500 bp and 5,000 bp from the marker.

Marker is > 5000 bp <=25000 bp After Feature means that the closest feature is located after the marker position and is between 5,000 bp and 25,000 bp from the marker.

Nearest feature is > 25,000 bp After Feature means that the closest feature is located after the marker position and is more than 25,000 bp from the marker.

Nearest feature is > 1 Mb After Feature means that the closest feature is located after the marker position and is more than 1,000,000 bp (1 Mb) from the marker.

Yes, Inside Gene means that the marker is located in the closest feature.

Table A3: Top 20 ranked ASD-genes

Gene	SDS	SDS P-value	Q-value	P-value	Function	Number of SNPs
CHD8	-1.2695	0.2043	0.00e+00	2.86e-07	Gene expression regulation	19
ADNP	-1.4575	0.1450	8.52e-15	2.86e-07	Gene expression regulation	40
FOXP1	0.0868	0.9308	1.77e-12	2.86e-07	Gene expression regulation	812
POGZ	-0.2413	0.8093	1.09e-10	2.86e-07	Gene expression regulation	46
SUV420H1	0.0110	0.9912	5.48e-10	2.86e-07	Gene expression regulation	39
DYRK1A	0.4570	0.6477	8.22e-10	2.86e-07	Cytoskeleton	271
SLC6A1	0.2669	0.7896	1.92e-09	2.86e-07	Neuronal communication	77
GRIN2B	-0.1591	0.8736	2.04e-08	2.86e-07	Neuronal communication	751
PTEN	0.3563	0.7216	5.26e-08	2.86e-07	Neuronal communication	37
MED13L	0.8679	0.3854	1.84e-06	2.86e-07	Gene expression regulation	268
GIGYF1	0.1557	0.8763	3.51e-06	2.86e-07	Other	2
CHD2	0.0256	0.9796	5.47e-06	2.86e-07	Gene expression regulation	212
ANK2	-0.0117	0.9906	1.43e-05	2.86e-07	Neuronal communication	924
ASH1L	0.2052	0.8374	2.04e-05	2.86e-07	Gene expression regulation	35
TLK2	-0.0767	0.9389	2.76e-05	2.86e-07	Gene expression regulation	56
CTNNB1	-0.1291	0.8973	3.98e-05	5.72e-07	Gene expression regulation	30
DEAF1	-0.1276	0.8985	6.92e-05	2.86e-07	Gene expression regulation	61
KDM6B	0.3618	0.7175	1.01e-04	1.72e-06	Gene expression regulation	2
DSCAM	0.0220	0.9825	1.35e-04	1.72e-06	Neuronal communication	1747
SETD5	1.3815	0.1671	1.84e-04	4.00e-06	Gene expression regulation	45

Table A3 | Top 20 ranked genes out of the 80 ASD genes. A list of the top 20 ranked ASD genes selected for the bar plot out of the 80 ASD genes found in the dataset. The genes were selected based on lowest Q-value.