Tampere University

Karoliina Salenius

INTEGRATIVE ANALYSIS OF MULTI-GENOMICS USING VARIABLE SELECTION FOR AUTISM DATA

Master's Thesis

Faculty of Medicine and Health Technology

April 2020

Abstract

Integrative analysis of multi-genomics using variable selection for autism data
Karoliina Salenius
Master's Thesis
Tampere University
Reviewers: Prof. Matti Nykter and PhD Jake Lin
Supervisors: Prof. Matti Nykter and PhD Reija Autio
Pages: 52
April 2020

Autism is a growing health issue and the last decade has seen its prevalence double. The prognosis of those affected is better with early intervention. This could be enabled by detecting the condition in younger children, but currently no medical test exists for the diagnosis of autism. Thus, the need for biomarkers that reliably detect autism is urgent. The etiology of autism is not well understood due to the heterogeneity and complexity of the condition. The aim of this thesis was to determine a method to study different datatypes measured from the same individuals to obtain a more holistic view of the genetic phenomena occurring in autism. A model was constructed with data from epigenomic, transcriptomic and genomic measurements by selecting the features that best correlate between the datasets and also contribute to the autism spectrum phenotype. Such feature selection from combined suitable data could also be used in biomarker research. In order to find the best fit for the model, estimates of the optimal number of components and features to select were determined using unsupervised approaches. The resultant model was validated with unseen data.

Keywords: multi-omics, genomics, transcriptomics, epigenomics, copy number variation, autism

Tiivistelmä

Usean omiikan yhdistäminen autismikirjon tutkimuksessa
Karoliina Salenius
Pro Gradu tutkielma
Tampereen yliopisto
Tarkastajat: Prof. Matti Nykter ja TkT Jake Lin
Ohjaajat: Prof. Matti Nykter ja TkT Reija Autio
Sivut: 52
Huhtikuu 2020

Autismikirjo on yleistyvä keskushermoston kehityshäiriö jonka esiintyvyys on kaksinkertaistunut viimeisten kymmenen vuoden aikana. Autismikirjon henkilön ennuste on sitä parempi, mitä aiemmin hoito aloitetaan. Tämä edellyttää varhaisempaa diagnoosia, mutta toistaiseksi ei ole olemassa lääketieteellistä testiä mikä mahdollistaisi taudinmäärityksen aikaisemmin. Autismikirjon etiologia on huonosti tunnettu mm. taudin heterogeenisyyden ja monimuotoisuuden vuoksi. Tämän tutkielman tavoitteena oli pyrkiä muodostamaan kokonaisvaltaisempi kuva autismikirjon geneettisestä taustasta tutkimalla samoilta henkilöiltä mitattuja eri datatyyppejä yhdessä. Datan integraatio tehtiin muodostamalla malli epigeneettistä, transkriptomista sekä genomista mittausdataa käyttäen ja valitsemalla näistä eniten toistensa kanssa korreloivat sekä fenotyyppiin vaikuttavat piirteet. Optimaali pääkomponenttien sekä valittavien piirteiden lukumäärä estimoitiin käyttäen datalähtöisiä menetelmiä. Malli validoitiin erillisillä näytteillä.

Avainsanat: multiomiikka, genomiikka, transkriptomi, epigenomi, kopioluvun variaatio, autismi

Preface

This thesis was made for the GEMMA (Genome, Environment, Microbiome and Metagenome in Autism) -project in the Computational Biology Group at Tampere University. I want to thank the group members for all the advice and support. Most of all I would like to thank my supervisors Matti Nykter and Reija Autio for giving me the chance and aiding me throughout this process.

In Tampere, 22nd April 2020

Karoliina Salenius

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

aCGH	Comparative Genomic Hybridization Array
ASD	Autism Spectrum Disorder
CNV	Copy Number Variation
CpG	Genomic site where cytosine is followed by guanine
cDNA	Complementary DNA, a single stranded DNA converted from RNA
FDR	False Discovery Rate
KEGG	Kyoto Encyclopedia of Genes and Genomes
LDA	Linear Discriminant Analysis
Lasso	Least Absolute Shrinkage and Selection Operator
mRNA	Messenger RNA
PCA	Principal Components Analysis
PLS	Partial Least Squares
SNP	Single Nucleotide Polymorphism
SSC	Simons Simplex Collection
SFARI	Simons Foundation Autism Research Initiative
sGCCA	sparse Generalized Canonical Correlation Analysis

1 Introduction

The increase of individuals diagnosed with autism has more than doubled in the past decade. There is no cure for autism and those affected often need a lifetime of support. The current trend in prevalence rates is alarming and distinguishing the causes behind this is imperative in order to reverse the present trajectory. Autism is a heterogenous disorder and its etiology is not well understood, although the presence of environmental factors accompanied by a genetic susceptibility seem to be evident. Studies have mostly been conducted on one or two genomic layers and have revealed a plethora of genes and genomic loci associated with autism. Several environmental triggers have also been suggested to contribute to the manifestation of autism. A reasonable deduction from the current state of research would be to combine data from several biological layers to better understand the complex molecular phenomena that underlies the pathophysiology of autism.

Recent developments in the technologies have yielded better availability of high-throughput measurements from different omics data. The aim of this thesis was to utilize publicly available data from different omics measurements in an integrative approach for a better perspective on autism pathophysiology. The structure of the thesis is as follows: the next part is dedicated to explaining the biological and methodological background essential in the study. The main objectives of the study are further described in the chapter 3. Chapter 4 provides the information on the materials along with the detailed demonstration of the workflow of the experimental processes conducted in the study. In chapter 5 the results obtained from the individual data analysis along with the integration model are presented. Chapter 6 resolves to discuss these further with some insight as to why some of the challenges in the analyses may have arisen. Chapter 7 concludes this thesis and addresses the future prospects of integrative analysis in the field of autism research.

2 Literature review

2.1 Autism

Autism Spectrum Disorder (ASD) is an umbrella term for a range of early onset developmental disorders, all of which are defined by varying levels of impairment in communication and social interactions and are often accompanied by repetitive behavior and restricted areas of interest (American Psychiatric Association 2013). The symptoms generally rank from mild (Level 1 autism) to severe (Level 3 Autism) (American Psychiatric Association 2013), but because the categorization of this broad phenomenon is difficult, usually the affected individuals are recognized as being “on the spectrum”. ASD was first widely acknowledged after the 1943 publication by psychiatrist Leo Kanner, which was based on his study of 11 children in his clinic, all of whom had no ambition toward social interaction but instead targeted the majority of their focus on objects. These children had been described as “self-sufficient”, “like in a shell” and “happiest when left alone” when intermitted to the clinic by their parents (Kanner 1944). Autistic disturbances of affective contact, as initially referred to by Kanner, is now known as autism spectrum disorder.

2.1.1 Comorbidities

Alike so many other psychiatric phenomena, ASD rarely occurs alone. In fact, over 70% of those affected also suffer from concurrent conditions (Lai et al. 2014). Neuropsychiatric conditions that often accompany ASD include intellectual delays, depression, anxiety and attention-defict hyperactivity disorder (ADHD). Other neurological disorders, such as tics or seizures are often present in ASD. Other common comorbidities among ASD individuals are gastrointestinal (GI) problems, such as Chron’s disease

(CD) or irritable bowel syndrome (IBS). The effect of comorbidities in many cases surpasses that of core ASD symptoms.

2.1.2 Treatment and diagnostics

There is no cure for ASD, and treatments include behavioral therapy and medicine, either one of these or both combined. Treatment depends heavily on the symptoms, and the commonly used medications generally combat comorbidities. Such medications often include antidepressants, anti-anxiety medications and ADHD-medication. Therapy can consist of applying alternative methods for learning communication skills and developing techniques for alleviating challenging behaviors. ASD is usually diagnosed by the age of three depending on the severity, and it has been established, that early intervention leads to better results (Lai et al. 2014). Diagnosing ASD relies on psychological tests, which are in most countries a part of the general healthcare and integrated into children's routine pediatrician visits at ages 18 and 24 months (Myers and C. P. Johnson 2007). There is currently no medical detection for autism, thus the race for alternative screening methods has sparked a number of studies in many scientific fields such as molecular biology, bioinformatics as well as computer vision (Glessner et al. 2009; Shen et al. 2016; J. Liu et al. 2019).

2.1.3 Epidemiology

The rate of children diagnosed with ASD has been on a steady rise in recent years, reaching a 1 – 2% globally – an increase of over two-fold in the past two decades (Isaksen et al. 2013). The geographic fluctuation in the prevalence of ASD has prompted studies to identify environmental factors that may contribute to its manifestation (Hertz-Picciotto et al. 2018). There is a substantial gender bias in ASD and around 80% of those diagnosed are boys. Protective properties of having a duplicate X-chromosome and prenatal hormones have been suggested as explanatory of the gender ratio (Skuse 2000). In addition, a recent review highlighted how ASD could be incited in boys selectively via male sex hormones that regulate the gut microbiota, to induce a more vehement immune response that leads to neuroinflammation (Kopec et al. 2018).

2.1.4 Etiology

Although the etiology of ASD is still widely obscure, a genetic aspect accompanied by epigenetic modifications and certain environmental triggers appear to be at the root of developing the condition. A reasonable deduction as to where to go from the current state of affairs would be to find robust methods to combine data from different types of experiments to gain a more comprehensive and holistic view of the phenomena that lead to the ASD phenotype.

Autism genes

A plethora of candidate genes have been associated with ASD and the diversity of the resulting list of genes has led to studies on their functional role (Gazestani et al. n.d.). A number of genomic structural variants have also been identified among ASD individuals and twin studies have linked both heritability and the presence of shared environmental factors as contributing to the development of the disorder (Tick et al. 2016a). One challenge in the study of etiology of ASD is that there are no noninvasive methods to study the molecular biology of the brain, and the majority of the transcriptomic and genetic data is obtained from blood samples.

Heritability

In the pursuit of unraveling the heritability of ASD, twin studies have provided insight into the matter and a *meta*-analysis estimated it to range between 64% and 91% (Tick et al. 2016b). The relative risk for a child to develop the ASD has been estimated to be 8.4-fold if they already have a sibling diagnosed (Hansen et al. 2019). The high heritability of ASD has given rise to studies involved with germline mutations, leading to the discovery of a plethora of risk genes and dozens of liable loci (Satterstrom et al. 2020; Levy et al. 2011). Alterations in the identified genomic regions account for around 10 – 20% of diagnosis and an aberration in any single one of these is responsible for as little as 1 – 2% of the cases (Abrahams and Geschwind 2008). For the majority of idiopathic ASD cases there is no clear genetic evidence, hence there is still much to be uncovered in the search for the genotypes accountable for ASD.

Simons Simplex Collection

In the pursuit to reveal *de novo* genetic alterations, the Simons Simplex Collection (SSC) was set up, and consists of samples from families with one child on the spectrum whose parents and siblings remain unaffected (Fischbach and Lord 2010). The most common structure in the family is as follows: the proband in these families is always high functioning and most commonly male, whilst the number of siblings is usually one and they are most often female (Fischbach and Lord 2010).

The Simons Foundation Autism Research Initiative (SFARI), which is part of the same Simons Foundation as SSC, maintains a curated database of genomic features associated with ASD. Their archive acommodates over 1000 genes and 2000 Copy Number Variations (CNVs).

Environmental triggers

Physiological differences between the autistic and normal phenotypes have been established via neuroimaging, but the underlying phenomena associated with the manifestation of ASD can be difficult to detect as the tissue-specific studies are sparse (Lai et al. 2014). Genetic and epigenetic factors, prenatal exposures (hormonal, maternal viruses and toxins) and paternal age have also been suggested as causative.

The involvement of the bidirectional communication of the gut microbiota and the central nervous system (CNS) via the gut-brain-axis at the onset of ASD has gained much attention recently (Hsiao et al. 2013; F. Liu et al. 2019; Matta et al. 2019). Microbiota is a key player in the development of the immune system during the first three years of life. The suggested mechanism leading to ASD involves the recruitment of the microglia to participate in the immune response induced by the microbiota. In this activated state, the microglia do not administer their normal role in the maintenance and pruning of the neurons, and the situation, when prolonged, likely has detrimental effects the neuronal development (Matta et al. 2019).

2.2 Genomics

2.2.1 Copy number variation

Although all individual human genomes are astonishingly similar (99.94%) (Auton et al. 2015), no two genomes are identical. Even monozygotic twins, whom develop from the same zygote and share 100% of their genetic material, have differences in their genomes (Bruder et al. 2008). These genomic variations take many forms and their size can range from single nucleotide changes (SNPs) to large, microscopically visible chromosome aberrations.

Copy number variation (CNV), a form of structural variation, refers to a deviation from the normal diploid copy number of two. These occur in segments that can span from 50bp up to megabases (Mb) in size (Zarrei et al. 2015). A duplication event occurs when these segments contain one or more additional copies and can be classified as a gain or an amplification, depending on the number of copies. A deletion event is the result of either the homozygous loss of two copies or the hemizygous loss of one copy of these genomic segments. The latter is also termed as loss of heterozygosity (LOH). Copy-neutral loss of heterozygosity (cnLOH) occurs when the overall copy number remains unaltered, but the loss of one sister chromatid is compensated with a duplication of the other.

Alterations in the diploid copy number can arise from incorrect non-homologous end joining (NEHJ) and non-allelic homologous recombination (NAHR) following double stranded breaks (DSB) in DNA. Alteration may also result from the incorrect formation of Holliday junctions during meiosis. The number and loci of CNVs differ between individuals and CNVs are a major contributor to the differences identified between the genomes of monozygotic twins (Bruder et al. 2008).

In the general population, CNVs account for an estimate of 4.8 – 9.7% of the genome and may contain entire genes along with their regulatory regions (Zarrei et al. 2015). Variation in such functional loci may impose an effect on gene dosage, although over 100 genes have been identified, whose entire removal poses no implicit phenotypical consequence (Zarrei et al. 2015). Most CNVs overlap with the redundant non-coding

sequences that cover 98.5% of the genome (Auton et al. 2015). These, previously labeled “junk DNA” regions, have more recently been shown to play, for example, a regulatory role amongst others.

Differentiating between the pathogenic and benign copy number alterations involves comparing the genomic structure of cases and healthy controls, whom often consist of immediate family members (Levy et al. 2011; Zarrei et al. 2015). As opposed to copy number alteration, copy number aberration is a term for the amplifications in cancer tissue, where the copy number may be significantly more than doubled. Some ambiguity remains with the nomenclature, but generally copy number variation refers to the germline events and alteration/aberration to the changes that occur during the individual’s lifetime in the somatic cells.

2.2.2 Methylation

Methylation refers to the biological process, where a methyl group is transferred to a DNA cytosine base. This small modification can have direct consequences on transcription. The addition occurs at the fifth carbon of the pyrimidine ring of the cytosine resulting in a 5-methylcytosine and to take place, it usually requires that the following base is guanine. The covalent modification is reversable and the methylated cytosines can be demethylated. Both processes are highly controlled and involve two groups of enzymes called DNA methyltransferases (DNMTs) that handle methylation and ten-eleven translocation (TET) family of methylcytosine deoxygenases, which are key instruments in active demethylation (Moore et al. 2013). Passive demethylation occurs during DNA synthesis, where the methylation step is simply left out. Altering the methylation status allows for swift responses to environmental exposures. Methylation patterns are tissue specific and enable the same genetic material to be utilized accordingly to the needs of different cell types. Methylation has been shown to be trans-generationally heritable and is thought to be the most stable form of epigenetic regulation (Rizzardi and Hickey 2019).

CpG islands and shores

CpG sites, where cytosine is followed by guanine, are scattered along the human genome. Areas with a high density of these dinucleotides are referred to as GpC islands. These are mainly located near transcription start sites and promoters, and because methylation prevents transcription factors from binding, hypermethylation in CpG islands generally has a downregulatory effect on gene expression. CpG sites up to 2kb from CpG islands are titled CpG shores and are likewise associated with reduced gene activity (Ciernia and LaSalle 2016).

Non-CpG methylation

Methylated cytosines can also be followed by other than guanine and are named CpH, where H stands for any nucleotide but G. CpH methylation occurs in different cell types but is particularly abundant in embryonic stem cells, neuronal precursors and neurons (Ciernia and LaSalle 2016). Neuronal CpH methylation has been shown to occur solely postnatally and its levels increase over time, unlike those of CpG methylation (Jang et al. 2017). Accumulation of CpH methylation is particularly dramatic in the frontal cortex during later life (Jang et al. 2017). Although CpH methylation spatially correlates with that of CpG's, it may have independent functions in the brain (Rizzardi and Hickey 2019). Differential methylation of CpH's embedded within gene bodies have been shown to maintain strong associations with differential gene expression (Ciernia and LaSalle 2016).

Methylation in gene bodies and intergenic regions

To make matters less straightforward, intergenic regions that contain CpG- or CpH-sites generally have the opposite effect when methylated compared to the promoter regions: instead of suppressing gene expression, methylation in intergenic sites increases it (Jang et al. 2017). Likewise, hypermethylation within gene bodies increases the expression of nearby genes (Jang et al. 2017). Curiously, methylation in intergenic regions and gene bodies seems to affect lowly expressed genes most and appears to have a function in the fine tuning of gene expression (Rizzardi and Hickey 2019).

2.3 Methods used in this study

2.3.1 Microarrays

Microarrays are, as their name implies, small arrays with oligonucleotide probes attached either to the array surface or to microscopic beads that are randomly interspersed over it. The probes are subjected to the genetic material extracted from the samples and hybridization of complementary sequences takes place (Jaksik et al. 2015). A fluorescent dye attached to the probe and/or target is activated upon hybridization and the resulting signal intensity can be measured. Answers to many types of experimental questions can be sought with microarray technology. The array format depends on the experiment and can be entirely customized to target specific sequences. Disease specific arrays are also readily available for many common diseases such as cancer or psychiatric disorders. Despite the recent developments in sequencing technology, microarrays are still used due to their efficacy and the relatively low cost. The following section focuses on the three microarray types used in this study.

Expression array

In order to measure the gene expression in a cell, the mRNA is extracted and usually converted to the more stable cDNA. The amount of mRNA approximates the genes that are being expressed and proteins being translated and thus, ultimately represents the current functionality of the cell. Because the target mRNA is spliced, the probes correspond to genes exons. Probes are labeled with a fluorescent dye and subjected to sample cDNA and complementary sequences get hybridized. Hybridized probes emit a fluorescent signal, which is recorded. The intensity of the signal represents the level of mRNA in the samples. Some limitations have to be taken to account when designing expression profiling experiments: different cells express different genes, and the cells from the tissue of interest should be used whenever possible. Also, many of the genes that get transcribed belong to a group called “housekeeping” and are essential for the cell to survive and whose production levels are consistent among samples. The levels of mRNA are only indicative of the actual production of proteins as there are regulatory steps that can’t be directly measured on expression array, such as mRNA degradation.

This should be kept in mind when deriving conclusions from the array experiments.

Methylation array

To determine the methylated sequences, the methylated cytosines must be first identified. This is handled with bisulfite conversion where the fragmented DNA of interest is subjected to bisulfite treatment. This will render the unmethylated cytosines into uracil which get further amplified as thymines while the methylated cytosines remain unaltered (Maksimovic et al. 2012). Illumina Infinum HumanRef 27k array design comprises of two types of probes: ones that will hybridize to the methylated sequences with C converted to T and others that hybridize to the original sequence (Maksimovic et al. 2012). The probes terminate at the 3' end with thymine and cytosine, respectively. The hybridization is followed by a single base extension by means of adding a fluorescently labelled nucleotide, either A or G depending on the terminating base downstream of the target of C or T (Maksimovic et al. 2012). The methylation status of the CpG site is measured as the proportion of the signal intensities of the similarly labelled bead types:

$$\beta = \frac{\mathbb{M}}{\mathbb{U} + \mathbb{M} + \alpha} \tag{2.1}$$

where β is the proportional methylation status, $\mathbb{M}$ is the methylated signal, $\mathbb{U}$ is the unmethylated signal and α a small offset to prevent big changes due to small estimation errors (Maksimovic et al. 2012). Often in downstream analysis the more robust M -values are used, which are the $\log_2$ ratios for intensities of methylated and unmethylated probes and are obtained with:

$$M = log_2(\frac{\mathbb{M}}{\mathbb{U}}) \tag{2.2}$$

The 27k array was first introduced in 2008 and targets CpG sites located at proximity to gene promoter regions (Maksimovic et al. 2012). The most recent Illumina methylation array, EPIC, utilizes two types of beads and different fluorescent dyes to keep track of methylated and unmethylated sequences and comprises of 850k probes which also targets CpG's outside of CpG islands as well as CpH's (Pidsley et al. 2016).

aCGH

Array Comparative Genomic Hybridization (aCGH) is a method for detecting copy number changes in sample DNA. The idea is derived from that of the comparative genomic hybridization, in which single stranded sample and reference DNA are tagged with different color fluorophores and with 1:1 ratio competitively hybridize to metaphase chromosomes (Bejjani and Shaffer 2006). The labelled DNA from both sample and reference will bind to their corresponding loci, and signal intensities of these can be measured and compared in order to detect chromosomal differences between the two.

aCGH technique offers better resolution and more efficiency than genomic hybridization. The experiment takes place on array with selected oligonucleotide probes that correspond to known regions in the genome. The array is then exposed, similarly as in the CGH experiment, to the sample and the reference mixture with 1:1 ratio. The reference and sample are differentially labelled, and the copy number status can be interpreted as the ratio of their signal intensities. The probes with significantly higher sample intensities compared to the reference indicate a gain whereas probes with relatively low sample intensities correspond to a loss.

In cancer studies, the reference DNA usually comes from normal tissue of the same individual as the cancer sample. Another study format is where a “universal” diploid DNA is used as a reference (Bejjani and Shaffer 2006). This is often the case in any non-cancer research, such as the studies of psychiatric disorders. There are several steps to take before any downstream analysis can take place, such as normalization, segmentation and the calling of integer copy number for the segment ratio data. These methods will be further explored in the following sections.

2.3.2 Microarray preprocessing

Preprocessing is required for all raw biological data in order to obtain the true biological signal. This consists of removing the bias and noise that results from the technical artefacts and processes and ensuring the quality of the samples. Preprocessing microarrays often consists of filtering the poorly performing probes or samples from the experiment and removing the background noise that is incessantly present in image

analysis (Jaksik et al. 2015). Any normalization aims to preserve the true biological signal whilst correcting for the unwanted experimental artefacts.

After background correction, two different normalizations may take place: the within- and between-array normalization (Ritchie, J. Silver et al. 2007). The within-array normalization is used in order to make the intensities of the samples within the array consistent. Between-array normalization aims to make the different arrays comparable with one another. The latter is typically not used for two color arrays, such as aCGH due to the presence of natural biological variation in signal intensities between the different samples (Ritchie, J. Silver et al. 2007).

After normalization, data is usually transformed to logarithmic scale as the intensity effects are often multiplicative ratios, which the logarithm turns into additive differences. Logarithmic scale offers a more robust distribution for downstream analysis and modelling as the indifference between two conditions centers to zero. The M -values in equation (2.2) are an example of the logarithmic transformation for better statistical qualities. This section will briefly introduce the preprocessing methods used in this study. More detailed explanations and the derivations of the formulas can be found in the references stated for each of the methods.

Background correction

Microarray signal intensities are first background corrected to remove any signal from other sources than the sample-probe hybridizations. Background can be measured using negative controls, which are surface areas or beads with no probes attached and no hybridizations take place. The normal-exponential convolution background correction model uses a Bayesian model to borrow information across probes and is robust for gene expression experiments where there are many variables and considerably less samples. The signal is considered to have an exponential distribution $exp(\lambda)$ and the background noise to have a normal distribution $N(\mu,\sigma^2)$. Estimate of the true signal, given the observed background subtracted intensities, is the conditional expectation:

$$\mathbb{E}(S|X=x) = \mu_{S\cdot X} + \frac{\sigma^2\phi(s;\mu_{S\cdot X}, \quad \sigma^2)}{1-\Phi(0;\mu_{S\cdot X}, \quad \sigma^2)} \tag{2.3}$$

Where S is the true signal, X is the background subtracted observed signal and Φ the standardized normal distribution (J. D. Silver et al. 2008).

Loess normalization

Within-array normalization is recommended for two-color aCGH data, which can be represented as the $\log_2$ -ratio of the signal intensities from both color channels. These are termed M-values (not to be confused with the M -value in the methylation data) and get centered around zero, which in the case of aCGH data stands for the baseline that corresponds to a normal diploid copy number. The copy number channels are converted to M- and A -values as follows:

$$M = \log_2(\frac{R}{G}) \quad \text{and} \quad A = \frac{1}{2}\log_2(R \times G) \tag{2.4}$$

Where R is the red channel and G is the green channel (Ritchie, J. Silver et al. 2007). The M- and A -values can be plotted against one another in a MA-plot, which often results in a curved figure. With median normalization, the weighted median is subtracted from the M-value, in order to perform the zero-centering. Lowess (locally weighted scatter plot smoothing) normalization is a more sophisticated method, and was inspired by Taylor's Theorem, which states that any continuous function $f(x)$ is essentially a line at a close enough observation. The idea is to fit a curve to the MA-plot and correct the values to output zero-centered M-values so that the curved plot becomes straight.

Quantile normalization

Quantile normalization is a between-array method and forces the distribution of probe intensities for each sample to be the same across the arrays. This method is sufficient for one-color experiments, such as the methylation and expression datasets in this study. The concept is based on the idea in quantile-quantile plots, where the plot is a straight diagonal if the distribution of the two data vectors is the same (Bolstad et al. 2003). In practice, the columns (each sample) of the matrix X are sorted in ascending order, resulting in matrix X_{sort}. The average of each row in the X_{sort} is then calculated and this mean value is assigned to each element in the row resulting in matrix X'_{sort}.

The normalized matrix is achieved by rearranging each column of X'_{sort} to the original ordering of matrix X. This way the quantiles across the samples are forced to be equal thus making the samples comparable with one another for downstream analysis.

Batch correction

Batch effect describes the systematic bias present in samples that is due to the handling of the samples in groups or batches. For example, not all processed in one go, but instead in batches: the samples may be collected at different sites, sent to the processing lab under different conditions and also processed with some variation in the method and reagents. Even the time of the day may affect the batch of samples in any of these steps. In location and scale (L/S) adjustment, the model for location (mean) and scale (variance) of the batches can be adjusted by standardizing their means and variances (W. E. Johnson et al. 2006). Using the Bayesian method to estimate parameters, information is borrowed across genes in each batch and the batch effect parameter estimates shrunk toward the overall mean of the batch.

GC-correction

GC-bias is a significant technical artifact that manifests as a wavy profile instead of a straight one when plotting copy number segments. This phenomenon occurs regardless of the platform used and no single explanation has been pinpointed, although the correlation with the genomic GC-content is evident. The cause is, thus, likely multifaceted, and DNA purity, DNA isolation protocols as well as PCR and dye labeling have been suggested as contributing factors (Leo et al. 2012). The GC bias is unimodal, as both GC- and AT- rich regions appear underrepresented in the results. When such regions are represented on the array with multiple consecutive probes, variation in their hybridization intensity does not reflect the true biological signal. This will pose an effect on the downstream analysis results, as the data from these regions will be skewed away from the expected values (Benjamini and T. P. Speed 2012). Moreover, an increased risk of false-positive as well as false-negative calls is present in these genomic neighborhoods due to the higher asymmetry of the probe signal (Leo et al. 2012). Thus, de-waving by GC-correction is often recommended for aCGH data.

Although GC -content is generally accounted for in the initial probe design, the GC rich areas in the DNA library are nevertheless likely to hybridize to probes with some GC sequences. There are several methods developed for GC-correction, including linear regression, the use of calibration data and mean or lowess centering of windows whose optimal size can be determined (Leo et al. 2012). In the latter the optimal window is selected based on "total variation distance" (TV) -score , which is estimated on GC stratified sample $x \in S_{gc}$ based on the GC content of the reference $gc = GC(x+a,l)$ where x is the position, $x+a$ the beginning of the window, l the width of the window and S_{gc} the strata (Benjamini and T. P. Speed 2012),. The number of fragments or hybridizations, F_{gc}, in S_{gc} is counted for each gc and the rate λ_{gc} is estimated. Choice of GC window estimates $W_{a,l}$ can be used to model predicted counts compared to one another by maximizing the TV distance score:

$$TV(W_{a,l},U) = \frac{1}{2\hat{\lambda}} \sum_{gc=0}^{l} \frac{N_{gc}}{n} |\hat{\lambda}_{gc} - \hat{\lambda}| \tag{2.5}$$

where estimates $W_{a,l}$ stand for the stratified rate and U for the uniform global mean rate in the sample (Benjamini and T. P. Speed 2012). Output of the equation (2.5) is the distance between the distribution of the window and the uniform global distribution and essentially translates to the proportion of hybridizations that were influenced by the stratification. Thus, the higher the TV score, the more dependent the hybridization is on the GC content, and the better the model is to correct for this dependence (Benjamini and T. P. Speed 2012). These windows are centered on the start position of each probe for microarray design and the GC correction is performed using the mean signal intensity of all probes that share the same GC fraction.

Segmenting copy number data

After preprocessing, in order to determine the copy number status in the samples, two further steps are to be taken: segmentation and calling of the integer copy number state. In the segmentation step, consecutive chromosomally ordered probes are considered to belong to the same segment if their $\log_2$ intensities agree. Segments are separated by breakpoints: if the $\log_2$ values differ above a threshold then a copy number transition occurs. Circular Binary Segmentation (CBS) and the use of Hidden

Markov Models (HMM) are the two most common methods used in segmenting CNV data. The latter is a probabilistic model that is used to determine a hidden sequence of states based on a sequence of observations (Fridlyand et al. 2004). The aim is to detect the total number of states in the data as well as the optimal state at each probe with a forward-backward algorithm, which involves three steps: first it computes the forward probabilities, then backward probabilities using the Bayes' rule and thirdly combines the first two steps to calculate a smoothed more accurate result using a K-state HMM with continuous output (Fridlyand et al. 2004). Optimal parameters for HMM are estimated with *Baum-Welch* method or the *EM algorithm*. A potential drawback is the exhaustive use of memory especially when working with large datasets. CBS, which provides a faster segmentation approach, is based on a change-point method and applies statistical testing to identify the breakpoints. Each segment is assigned the copy number derived from the average value of all the probes residing within that segment. The change point is estimated using a maximum likelihood test statistic $T = max_{1 \leq j \leq m} T_{ij}$, and T_{ij} is obtained with the two-sample t-test that compares the mean of the observations at indexes $i+1$ to j with markers m that correspond to the data $X_1, \ldots, X_m$ and the mean of the total number of observations:

$$T_{ij} = \frac{\hat{Y}_{ij} - \hat{Z}_{ij}}{[s_{ij}(j-i)^{-1} + (m-j+i)^{-1}]^{\frac{1}{2}}} \tag{2.6}$$

where $\hat{Y}_{ij} = (X_{i+1} + \cdots + X_j)/(j-i)$, $\hat{Z}_{ij} = (x_1 + \cdots + X_i + X_{j+1} + \cdots + X_m)/(m-j+i)$, and s_{ij}^2 is the mean squared error (Venkatraman and Olshen 2007). In other words, each segment can be conceptualized as a circle by connecting it's endpoints and the likelihood test is used for evaluating whether this circle consists of two complementary arcs with unequal mean values. Statistically significant changes are selected based on p-values smaller than the chosen threshold level α and changepoints i and j that maximize the test statistic (Venkatraman and Olshen 2007).

Copy number calling

The $\log_2$ ratios of copy number segments are commonly assigned integer copy number states. These may be inferred using thresholds or simply by rounding exponential of

the $\log 2$ ratios to the nearest integer. The diploid copy number for a given segment is

$$2 \times 2^n \tag{2.7}$$

where n is the $\log_2$ ratio of the segment. This method is well suited for assigning copy number to germline samples (Talevich et al. 2016). The thresholds are commonly used in somatic samples in cancer studies where the purity of the sample also needs to be accounted for.

2.3.3 Statistical analysis

Statistical analysis consists of a number of tests to answer different biological questions. In microarray experiments common statistical tests include deriving the differential expression or methylation between two or more groups or conditions of interest. Confounding variables such as age and sex may contribute to the outcome, and their effect can be adjusted for by using a regression model in order to unveil the behavior of the variable of interest. Microarray studies generally consist of thousands of genes and a substantially smaller number of samples, which can lead to the multiple testing problem. This is sometimes also referred to as the curse of multidimensionality. Thus, corrections need to be made to identify the genes that are truly significant: the true positives. Next section briefly explains two fundamentals of statistical data analysis: the linear model and the multiple testing problem.

Linear models

Linear regression is a commonly used statistical method to model the relationship between a response variable and one or more explanatory variables. The response and explanatory variables are also called independent and dependent variables, respectively. The case of one independent and one dependent variable is referred to as univariate analysis. When multiple independent variables describe dependent variables, the term used is multivariate analysis (MVA). In the case of MVA the linear model for

$Y_1,\ldots,Y_K$ outputs and $X_0,\ldots,X_p$ predictors can be written as:

$$Y_k = \beta_{0k} + \sum_{i=1}^{p} X_i \beta_{ik} + \epsilon_k \tag{2.8}$$

where ϵ denotes the error term that is assumed to be i.i.d. (Hastie et al. 2016). β_{0k} is the intercept and β_{ik} stands for the unknown coefficients that determine the slope of the linear model and parametrize the average expression in the design (Ritchie, Phipson et al. 2015). In matrix form the equation is simply:

$$Y = XB + E \tag{2.9}$$

For N samples, Y is the $N \times K$ response matrix, X is the $N \times (p+1)$ input matrix with rank N, B is $(p+1) \times K$ matrix with the parameters and E the $N \times G$ matrix of errors (Hastie et al. 2016). Linear regression essentially fits a line through the datapoints in order to approximate the function $E(X|Y)$ (Hastie et al. 2016). The unknown coefficients β can be estimated for each independent variable. A commonly used method is the least squares fitting and the best fit can be found by minimizing the Sum of Squared Residuals (SSR). Least squares estimates often contain low bias, but this comes at the cost of high variance (Hastie et al. 2016).

Identifying significant genes

The significance of the coefficients β in the linear model can be evaluated with a *t*-test. A zero coefficient denotes no dependency between the independent and dependent variables, whereas a large deviation from it means that there is a significant dependency between them (Ritchie, Phipson et al. 2015). In statistical analysis involving microarray or sequencing results, the same linear model is fitted to each gene – often thousands of times over, which can lead to the multiple testing problem. This setting, however, allows for pooling information from the entire dataset to enhance the robustness of the model. This way the variances can be shrunk towards the common mean and a specific moderated *t*-test can be used to obtain a more statistically robust result (Ritchie, Phipson et al. 2015).

Multiple testing problem and the false discovery rate

The statistically significant results from the analysis rely on the test statistic and depend on the chosen threshold α. When the obtained p-value is smaller than α the null hypothesis H_0 is rejected. A common value for α is 0.05, which can be interpreted as accepting that 5% the rejected H_0's are false positives, i.e. cases where the null hypothesis holds. This in many cases is acceptable, however, with a large number of tests for small set of subjects, the rate of false positives becomes very high. In the case of gene measurements where the number of genes runs in tens of thousands and number of samples is hundreds at best, the amount of falsely significant genes would be unacceptably high. This is a case of multiple testing problem.

The *false discovery rate* (FDR) controls the proportion of the false positives. A popular method for FDR proposed by Benjamini and Hochberg (1995) is based on ordering the p-values and using a threshold to correct for the rate of false positives. For *m* ordered p-values $P_1,\ldots,P_m$ from $H_1,\ldots,H_m$ hypothesis tested , the threshold can be identified by finding the largest *k* so that

$$P_k \leq \frac{k}{m}\alpha \tag{2.10}$$

and reject the H_0 for those H_i where $i = 1,\ldots,k$ (Benjamini and Hochberg 1995).

2.3.4 Data integration

The aim in systems biology is to understand the complex interplay of biological functions and gain a more detailed description of the structural architecture of the cell (Subramanian et al. 2020). This is plausible by studying the cells processes that span several biological layers together. Important features can be missed when studying a single source of information in isolation. Integrating multiple types of measurements from the same set of individuals in a meaningful manner provides a better approach and is referred to as multi-omics data analysis. It has been a promising method for disease subtyping, understanding disease biology and aiding biomarker discovery (Subramanian et al. 2020).

Multi-omic approaches have been successful in cancer studies and paved way for per-

sonalized medicine and contributed to developing novel therapies (Karczewski and Snyder 2018). Integration of different genomic data has also been promising in the study of ASD, which as a highly heterogenous disease has been notoriously difficult to reduce to a set of features predictive of the phenotype using a single source of data (Betancur 2011). Common sources to study the molecular processes and interactions include the genomic, transcriptomic, epigenomic and proteomic measurements. In recent years the study of the microbiota using metatranscriptomics and metagenomics has been added to this growing list. Genetic data can be attained from numerous different types of microarrays and sequencing platforms. For metabolites and proteins, NMR or mass spectrometry are commonly used, to name a few technologies. The output from any of these experiments is a multidimensional data matrix and the combination of different omics datasets results in large amounts of data, often with various statistical properties (Subramanian et al. 2020).

Two main courses for data integration can be followed:

1. A multistage method where data is first analyzed individually, followed by evaluation of the associations of the statistically significant features.
2. A meta-dimensional method, where the appropriately preprocessed and normalized data are simultaneously integrated and the significant features identified.

(Vlahou et al. 2017). A draw-back of the multistage method is that its results may not be conclusive and different datatypes analyzed in isolation can be difficult to integrate. The second option, multidimensional methods, can further be divided into:

1. Early-stage integration using concatenation-based method where the different omics data matrices are merged into one
2. Intermediate-stage integration, which comprises of constructing a joint model of the different omics dataset
3. Late-stage integration, where each dataset is modeled individually and fitted values of each model are weighted for inference

(Vlahou et al. 2017). The main obstacle in the first method is the different distributions and scales of different omics data when merging them into one. This would require large scale operations for normalization and potentially lead to loss of information. The

third method involves several separate models and inference is based on choice between Bayesian classification, Random Forest and ensemble classifiers and is only recommended when the first or second methods are not available (Vlahou et al. 2017). Thus, in this study the second method of obtaining a joint model for the datasets using matrix factorization was explored for *supervised* analysis of the different genomics data. Such model can further be used for classification and prediction purposes.

Supervised and unsupervised methods

For multidimensional data obtained from different omics platforms, one central task is to reduce the dimensions and find key features with as little loss of information as possible. A few methods have been developed to achieve this. Supervised methods take into account prior knowledge about the data, such as the phenotype of interest. In machine learning methods this is often presented as a vector of class labels that separate e.g. cases from controls. Class labels can also depict different cell types, disease subtypes or experimental conditions. In the linear model in (2.9) this can be information used to form the design matrix X. In unsupervised techniques no prior knowledge is given to the algorithm and instead any patterns that emerge from data are explored.

Dimensionality reduction

For illustrations purposes the object in data reduction is to find a way to present the data in 2 – 3 dimensions that are as descriptive of the data as possible. Principal component analysis (PCA) is an unsupervised method and performs a linear transformation to a lower dimensional space. In PCA, eigenvectors and -values of the covariance matrix C for mean centered data matrix X are computed using orthogonal projection. Alternatively, the eigenvectors and -values can be obtained with Singular Value Decomposition (SVD). The eigenvalues are ordered decreasingly, and the largest eigenvalue corresponds to the vector denoting the direction of the highest variance in the data. The principal components are obtained by:

$$T = XP + E \tag{2.11}$$

Where P holds the eigenvectors of C, E is the error matrix and T has the reduced dimensions, the *principal components* of data in X (De Bie et al. 2005). The process is iterative, and each subsequent component is constructed by minimizing the error term. In other words, PCA works by preserving the variance from the original data. However, the PCA does not take into account the response Y, and although the principal components in T can be used as regressors of Y, selecting the subset of features that best describe Y is not feasible (De Bie et al. 2005).

Partial Least Squares (PLS) is a family of methods similar to PCA. A key difference is that PLS takes data from both X and Y and projects them onto a new space. In Partial Least Squares Discriminant Analysis (PLS-DA) the dependent variable Y denotes categorical data such as class labels for data in X. The addition of response Y in the formula makes PLS a supervised method for multivariate dimensionality reduction. The decomposition of X and Y using a covariance matrix C of X and Y can be written as:

$$\begin{aligned} X &= TP^T + E \\ Y &= UQ^T + F \end{aligned} \tag{2.12}$$

Here T and U hold the component vectors and P and Q have the loadings of the covariance matrix C of X and Y, respectively (De Bie et al. 2005). The loadings are the coefficients that define the component. PLS components, or *latent vectors*, are identified via a simultaneous decomposition of X and Y using the covariance matrix and hence, in PLS the covariance between X and Y is maximized (Ruiz-Perez et al. 2018). Similar to PCA, the decomposition in PLS can be done with SVD, and now, because the resulting decomposition is obtained using information on Y, performing feature selection from the reduced dimensions can be accomplished (Vlahou et al. 2017).

Shrinkage

Feature selection has two distinct advantages: it regularizes the model to avoid overfitting and identifies the key effectors in the data. This can be achieved by imposing a penalty on the size of the regression coefficients. Particularly when the regression model contains many correlated variables, their coefficients can have both high bias

and variance (Hastie et al. 2016). A large positive coefficient may get entirely cancelled by a countercorrelated large negative coefficient. To alleviate this behavior *lasso* regression uses the size penalty. Penalized SSR are minimized with a complexity parameter λ, which determines the amount of shrinkage. By increasing λ the shrinkage is likewise increased. Following this, the coefficients are shrunk towards zero. In the format of PCA or PLS, this essentially shrinks the last components with the minimal variance/covariance to zero. The formula for lasso estimates is:

$$\hat{\beta}^{lasso} = \frac{argmin}{\beta}\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\} \tag{2.13}$$

Thus, with a larger lambda many of the coefficients will be set to zero and can be discarded. The resultant model, which includes a subset of key features is called the *sparse* model.

Sparse Canonical Correlation Analysis

To integrate different datasets, also termed blocks, measured from the same set of individuals, Generalized Canonical Correlation Analysis (GCCA) provides a useful tool (Tenenhaus et al. 2014). Contrary to what its name implies, GCCA generalizes the Partial Least Squares (PLS) method for the integration of two or more datasets. PLS is based on preserving the covariance of the original data when projecting it onto lower dimensional space. GCCA does this by maximizing the correlation. Instead of utilizing the covariance matrix for calculating the eigenvectors and -values, the correlation matrix is used to find the direction of the maximum correlation between the variables in the different blocks. To obtain a sparse model, the correlated variables are selected using l_1 -penalization and for Q (normalized, centered and scaled) datasets $X^{(1)}(N\times P_1),\ldots,X^{(Q)}(N\times P_Q)$ that measure the $P_1,\ldots,P_Q$ variables from the same *N* samples, sGCCA works by solving the optimization problem for each of the dimensions $h = 1,\ldots,H$:

$$\begin{gathered}\frac{max}{a_h^{(1)},\ldots,a_h^{(Q)}}\sum_{i,j=1,j\neq 1}^{Q} c_{i,j}cov(X_h^{(i)}a_h^{(i)}, X_h^{j}a_h^{(j)}), \\ \text{s.t} \quad \|a_h^{(q)}\|_2 = 1 \quad \text{and} \quad \|a_h^{(q)}\|_1 < \lambda^{(q)} \quad \text{for all} \quad 1 \leq q \leq Q\end{gathered} \tag{2.14}$$

where $c_{i,j}$ denotes the Q design matrix to indicate whether to maximize the covariance between the blocks X_i and X_j: $c_{i,j} = 0$ indicates no relationship between the datasets and $c_{i,j} = 1$ denotes strong relationship and $a_h^{(q)}$ is the coefficient vector of the residual matrix $X_h^{(q)}$ from block $X^{(q)}$. Here $\lambda^{(q)}$ is the complexity parameter presented in (2.13), which controls the shrinkage so that a subset of the variables with non-zero coefficients that define each component score $t_h^{(q)} = X_h^{(q)} a_h^{(q)}$ can be selected. These are the variables that correlate most between and within the blocks. Shown in (2.14) is the sGCCA model for attaining the first dimension, and after obtaining coefficients $a_1^{(1)}, \ldots, a_1^{Q}$ for $h = 1$, the coefficients for the subsequent n components $h = 2, \ldots, n$ to maximize (2.14) are calculated iteratively using the residual matrices $X_n^{(q)} = X_{n-1}^{(q)} \setminus t_{n-1}^{(q)} a_{n-1}^{(q)}, \quad 1 \leq q \leq Q$ until the predefined number of components has been reached. The assumption here is that most of the biological variation can be attained from the component scores $t_1^{(q)}, \ldots, t_h^{(Q)}$ so that the statistical model is not impacted by the unwanted variation that is due to the heterogeneity in the datasets $X^{(q)}$ (Singh et al. 2019).

Discriminant analysis

The aim of discriminant analysis is to partition the data into subsets that best predict the phenotype of interest. For example, in linear discriminant analysis (LDA) the intention is to find the projection which best separates the sample classes and is referred to as the *decision boundary* (Hastie et al. 2016). PLS-DA can also be utilized to identify the decision boundary. In the case of variable selection, the decision boundary can be determined from the selected variables. The model obtained with the selected variables can be validated on how well it is able to discriminate between classes. This involves a set of training data with known class labels for the construction of the model. By introducing new, previously unseen data to the model, predictions can be made on its class. Such validation data can be used to assess the predictive power of the model. For validation data, the class labels are known but not shown to the model. The accuracy of the predictions can be evaluated based on the proportion of correct predictions from the all predictions. Sensitivity and specificity of the model can be plotted against one another in a ROC -curve and the area under the curve (AUC) is often used as a measure of performance.

Pathway analysis

To further explore the biological function of the obtained list of features from the different data measurements, their enrichment in known biological pathways can be evaluated. The features are first annotated, usually to denote genes, and can be compared to pathway databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) and reactome. The idea is to find whether the selected genes appear in a pathway more than would be expected by chance, i.e. they are over-represented in any known pathway. A popular method to obtain the significance of a pathway P is the hypergeometric and is denoted as:

$$P = \sum_{j=x}^{K} S \frac{\binom{K-m}{K_S-j}\binom{m}{j}}{\binom{K}{K_S}} \tag{2.15}$$

where K_S is the length of the list of genes that are tested, x the number of these genes that appear in the pathway and m the total number of genes in the pathway (Evangelou et al. 2012).

3 Objectives

The main objective in this study is to apply a multi-omics approach to the study of ASD and develop a pipeline for the preprocessing and integration of different omics measurements from the same individuals. The omics data need not have identical features measured i.e. the rownames of the data matrices may differ. The pipeline can, with some adjustments, be applied to other data that satisfy these requirements. The raw data is preprocessed and normalized accordingly before converting to suitable matrix format that can be used in the integration. Key features will be selected that best describe the phenotype and also correlate between the different genomics data in an attempt to distinguish if biomarkers could be retrieved by using multiple omics measurements. The feature selection is performed using a supervised analysis technique. Interactions of the key features are further explored in enrichment analysis.

4 Materials and methods

4.1 Overview of the data

The samples used in the experiment were from the Simons Simplex Foundation (SSC), which includes families where a single child is diagnosed with ASD. The foundation was originally established in order to study de novo mutations in ASD (Fischbach and Lord 2010). The three datasets used in the analysis were obtained from NCBI GEO database under the ID's GSE23682, GSE27044 and GSE37772 and include CGH, methylation and expression arrays, respectively. The accompanying information along with the familial ties and the phenotype (ASD or control) included the gender and age at the time of the sample collection. ASD samples are to be referred to as probands and the healthy samples as siblings. All samples were derived from peripheral blood lymphoblasts. Common samples between the three datasets were identified after preliminary preprocessing and quality control (QC) of the raw data, to include only the samples that are both present and sustained good performance in all of the three experiments. This yielded 98 samples from 62 different families. The core of the analysis was done on 72 samples that were discordant siblings from 36 families. The remaining 26 unrelated samples were kept for validation purposes. All the subjects were male and the age distribution for both of the study groups is presented in 4.1.

4.2 Expression data

The expression dataset (GSE37772) included 439 samples from 224 families in the SSC collection and was derived with platform GPL6883 (Illumina HumanRef-8 v3 expression 27k BeadChip) which holds 27 000 probes located in the exons and are thus complementary to the cDNA made from mRNA.

Group	Min.	Median	Mean	Max
ASD siblings	5.1	9.6	9.9	16.8
CTRL siblings	4.5	9.9	10.9	20.3
ASD unrelated	4.3	9.0	9.3	17.7
CTRL unrelated	5.7	10.9	11.6	20.1

Table 4.1. *Age distribution in years of both of the sample groups shows that the age range of the subjects is quite wide. Majority of the subjects were 9 – 10 years of age, and the ASD probands are more often the younger sibling. In the unrelated subjects similarly, the ASD proband is often younger*

4.2.1 Preprocessing expression array

The data was background corrected with normal-exponential convolution model using a Bayesian model as described in section 2.3.2. Negative controls were inferred from the detection p-value assigned to each probe (Ritchie, J. Silver et al. 2007). Detection p-values measure how likely it is that the probe signal differs from the background and a large value denotes no significant difference. Quantile normalization was used, as suggested suitable for a homogenous dataset (Ritchie, Phipson et al. 2015). The quantile normalization effect was visualized with boxplots of the $\log_2$ normalized data. Quality control included removing probes that had a high detection p-value (> 0.05) in more than half of the samples. This resulted in 11130 expressed probes corresponding to 9498 different genes. The data was then corrected for batch using ComBat which utilizes Bayesian modeling as described in section 2.3.2. Correction was made for the batches in which the samples had been handled and mailed to the processing lab as well as the array processing of batches. The coordinates of the probes were updated to match those in the latest genome (hg38) from the original (hg18) using illumina nuID system, which transforms the probe sequence into a unique identifier via a lossless compression that is reversable, so that the identifier can at any time be converted back into the sequence (Du et al. 2007). Finally, the data was filtered so that it only included the shared samples between the methylation and copy number datasets.

4.2.2 Statistical tests for differential expression

The samples were split into paired data with the siblings (n=72) for the core of the analysis and the remaining unrelated samples (n=26) were analyzed separately. In order to evaluate whether the expression levels between the probands and siblings differ in any specific genes, a linear model (equation 2.8) was fitted to the data and a paired moderated *t*-test was performed for each of the sibling pairs using R-package limma (Ritchie, Phipson et al. 2015). The *t*-statistic in the paired *t*-test evaluates the difference of the coefficients β_{ik} between the siblings. The effect of age was tested both as categorical and continuous variable, and the latter was selected because of the nature of the paired *t*-test in order to keep more information in the model. The unpaired data underwent a moderated *t*-test with the age covariate as a categorical variable sectioned into four age groups: toddler, elementary- and secondary -schools and young adults. The p-values were FDR corrected using the BH method as presented in 2.3.3.

4.3 Methylation data

The methylation profiling dataset (GSE27044) originally included 1128 samples from four different cohorts and had been acquired with GPL8490 platform (Illumina HumanMethylation27k BeadChip) which includes 27 000 probes located in gene promoters. The reasoning for this is that methylation in the promoter region directly affects the expression of genes.

4.3.1 Preprocessing methylation array

Quality control included keeping only the samples with median $\log_2$ signal intensity taken from both methylated and unmethylated probes exceeding 10.5, as implemented in Bioconductor package minfi (Aryee et al. 2014). This is slightly less stringent than the method used in the original article, where the samples with the two channels average signal intensity under 2000 were discarded (Alisch et al. 2012). The data was then background corrected and quantile normalized using R package minfi, which normalizes the methylated and unmethylated signal intensities separately (Aryee et al. 2014).

98 poorly performing probes with high detection p-values (> 0.01) were removed. After these steps 17013 probes remained corresponding to 12127 different gene promoter regions. Batch correction steps were taken in the following order to correct for:

1. Position of the array
2. Batch in which the arrays were processed
3. Delivery group indicating the batches in which samples were delivered to the research lab for processing

as was suggested by Price and Robinson (2018). The coordinates of the probes were updated to the latest genome (hg38) using a chain file from University of California Santa Cruz (UCSC) genome browser (https://genome.ucsc.edu 18.3.2020). After pre-processing the data was filtered to include the common samples between the datasets.

4.3.2 Statistical tests for differentially methylated genes

The analysis for differential methylation was performed for paired and unpaired samples separately. For statistical testing the M -values of the methylation intensity signals obtained with equation 2.2. For the data matrix with unique genes as rows, the average value of multiple probes mapping to the same gene was used. Statistical analysis included fitting a linear model similar to the expression data 2.8 and DE genes identified using the paired moderated *t*-statistic for the siblings and unpaired for the unrelated samples. The age was adjusted for as a continuous variable for the siblings and categorical for the unrelated samples as in expression data. The p-values were FDR corrected with the BH method as described in 2.3.3.

4.3.3 Differentially methylated regions

In search for differentially methylated regions (DMRs) the data matrix with M -values and all the probes from the dataset was used. The R-package *Bumphunter* accepted data from a 27k array. Different settings for the gap allowed in the clustering of probe regions were tried, but the algorithm found no significant regions.

4.4 Copy number data

Copy number data included 3852 samples from SSC families that had been obtained with GPL10815 platform (NimbleGen Human 2.1M array) with over 2M 60bp probes that correspond to loci interspersed along the genome excluding the highly repetitive regions such as centromeres and telomeres. Due to the volume of the data, the common samples were filtered before the preprocessing steps.

4.4.1 aCGH preprocessing

The Nimblegen copy number array was a two-color array with separate channels for the competitive hybridization signal for the reference and the sample. The signal from the two channels was first converted to M and A -values using the equations in 2.4 and background correction and loess normalization was performed on with the default settings in R package limma -function *normalizeWithinArrays* (Ritchie, Phipson et al. 2015; Smyth and T. Speed 2003). The normalized intensity values were then GC -corrected using *ArrayTV* R-package with window size 60bp. The optimal window was first measured from five random samples using the TV -score (equation 2.5) for 3 different window sizes (60, 600 and 6000) and performed best at the size of the probe fragment as suggested in the literature (Benjamini and T. P. Speed 2012). Points with values > 4 standard deviations (SD) from the neighbouring regions were considered outliers and were shrunk to these neighbouring values with R-package *DNAcopy*. Using the same package, segmentation of the smoothed data was attained with CBS with equation 2.6. Any change point with SD < 3 were removed. The integer copy number for each segment was then called with Python package *CNVkit* using the options to median center the segments and derive the copy number by rounding the result from equation 2.7 to the nearest integer (Talevich et al. 2016). The segments with < 10 probes were removed as not having strong enough evidence of the copy number status. Also, regions where the copy number for all the samples was normal were not considered informative and were thus removed from downstream analysis.

4.4.2 Copy number differences between the groups

The frequencies of copy number gains and losses for both cases and controls were visualized and the genomic regions of these annotated, in order to distinguish any difference between the copy number status between the groups. Common regions of gains and losses were determined for the ASD probands and their siblings separately using R-package *CNVRanger*, which uses the method described in Mei et al. 2010 for identifying the recurrent regions.

4.4.3 Data integration

Multistage approach

The expression and methylation arrays had altogether 5905 common genes. In order to identify any common ground between the individual analyses, results from these were visualized together. This included finding the intersection of the genes identified from expression and methylation datasets as most differential between the probands and siblings. These were further separated to hypo- and hypermethylated genes for the methylation data and over- and underexpressed in the expression data according to the log fold change between the sample groups. The intersections were visualized to see whether the hypomethylated genes would be over expressed or hypermethylated underexpressed in the probands. The total number of most differential genes between the groups were also intersected with a curated list of 1079 ASD -related genes from the SFARI -archive (https://gene-archive.sfari.org/database/human-gene/, read 12.4.2020).

For copy number data, a permutation test (n=1000) was made to evaluate whether the recurrent copy number alterations occurred in the regions identified as most variable in the methylation and expression data more often than would by chance alone. This was implemented in the R-package *regioneR* and the regions were randomized in a per chromosome basis.

Metadimensional approach

To proceed with the metadimensional data integration, the expression and methylation data were filtered to include the most variable genes as suggested in literature (Singh et al. 2018). The variances were visualized in a histogram, and the thresholds for the proportion of the genes to include was decided upon these. The conclusion was to attain the top 25% most variable genes resulting in 3032 genes in the methylation data and 2375 in the expression data.

Copy number data existed as a list of segments for each sample and these regions differed for each of the samples. To form a matrix with uniform rownames, the genome was split into windows of 50kb. These windows were set as the rows of the matrix and the samples as columns, as was the setup in the expression and methylation data. The values of the sample segments overlapping a window by more than 5kb were added to the corresponding row. In the case of multiple segments overlapping a row region, the weighted average of the log_2 values of these segments was assigned. The lengths of the corresponding overlapping segments were used as the weights for calculating the average. This resulted in 4195 rows corresponding to 50kb size genomic regions.

Rationale behind the 50kb window was that after filtering out segments with evidence of less than 10 probes, the segments were all likely longer than 10kb because the average probe distance was about 1000kb. For 50kb windows, the shortest segments would not become too long when constructing the matrix, while keeping the matrix still at a reasonable size (number of rows $<$ 10 000) for the data integration part as suggested in literature (Singh et al. 2018).

Metadimensional integration of the data was done using R-package *mixOmics*, which utilizes sGCCA algorithm (2.14) for feature selection. First the number of components to use for the model were chosen by performing PCA on the separate datasets and the variance explained by each principal component was visualized to aid the decision. Correlation between the datasets was evaluated using GCCA before proceeding to variable selection to construct the design matrix $c_{i,j}$ as described in section 2.3.4. In the design, the correlation to the phenotype was set to maximum (=1) to emphasize the features that contribute to the phenotype. The number of features to select from each of

the components was assessed using cross validation. The samples were split into 10 subsets and in an iterative process each of these was used as validation data once and the model constructed with the rest. Several sparse models were tried using differing number of components and features, and the best performing model was finally chosen based on its accuracy when using the data from the unrelated samples for validation.

To take the analysis further, and to find the functionality behind the selected features, enrichment analysis provides a useful tool. The copy number regions were annotated, and the corresponding genes added to the list with genes selected from the expression and methylation blocks. The enrichment analysis was performed using the hypergeometric implemented in the R-package *reactomePA* to identify whether the number of selected genes belonged to pathways in two databases – KEGG and Reactome, more than would be expected by chance alone.

4.5 Overview of the workflow

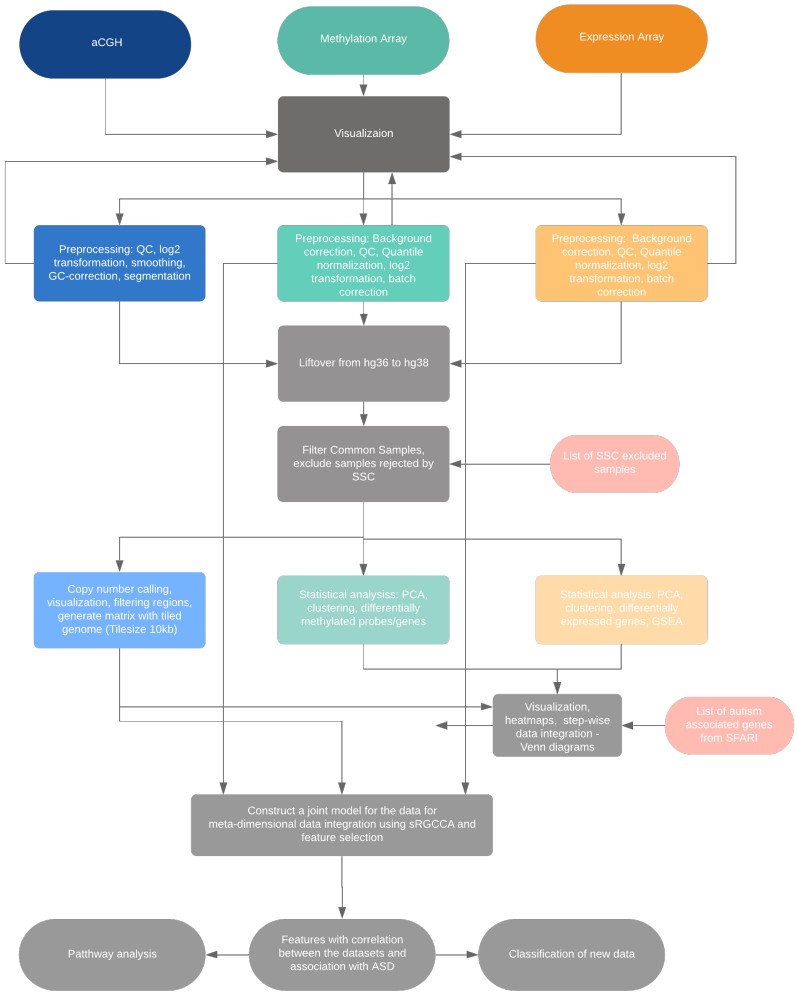


Figure 4.1. *The workflow of the experiment from top to bottom begins with preprocessing the raw data, coordinate liftover to the most recent genomic build and the identification of the common samples between the datasets followed by individual statistical analysis and finally the metadimensional data integration.*

5 Results

5.1 Expression dataset

The dataset showed no statistically significant DE genes. After FDR the adjusted p-values were above 0.67. The log fold change between the sibling pairs was small, which is a reasonable explanation for the negative results. The visualization of the genes using a heatmap in *figure 5.1a* shows that the probands and siblings do not separate to their own clusters. Similarly, the unpaired data did not include any significant DE -genes, but the clustering in *figure 5.1b* does set most of the subject groups into their own cluster.

5.2 Methylation dataset

The paired methylation data, similarily to the expression data, yielded no significant results when testing for differentially methylated genes, and after adjustment even the smallest p-values were as high as 0.999. The log fold change between the study groups was also very small, which could have been one reason why no significant difference in the methylation status was detected. The genes closest to being DM were visualized in a heatmap which clustered the two groups quite nicely in both paired and unpaired data as *figures 5.2a* and *b* show.

5.3 Copy number data

The copy number data from the siblings was visualized using frequency plots to determine any difference between the two groups. The plots in *figures 5.3a* and *5.3b* were very similar, although one of the control samples displayed Kleinfelter syndrome. The

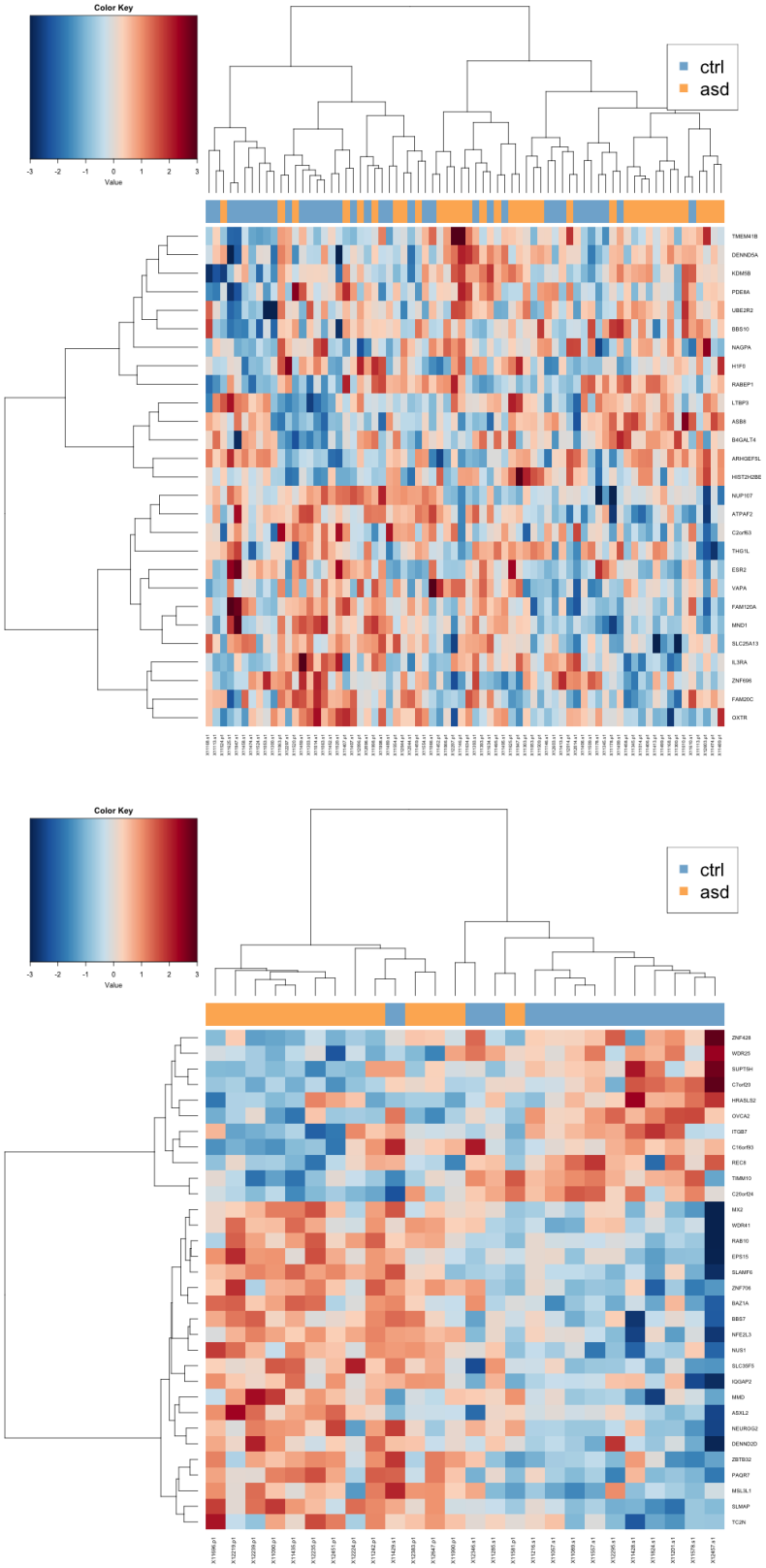


Figure 5.1. *Heatmaps visualizing the genes closest to being DE in siblings (a) and unrelated samples (b) shows that the siblings do not form distinct clusters. The number of genes visualized was based on attaining the clearest result with hierarchical clustering using average linkage with Pearson correlation as the distance function.*

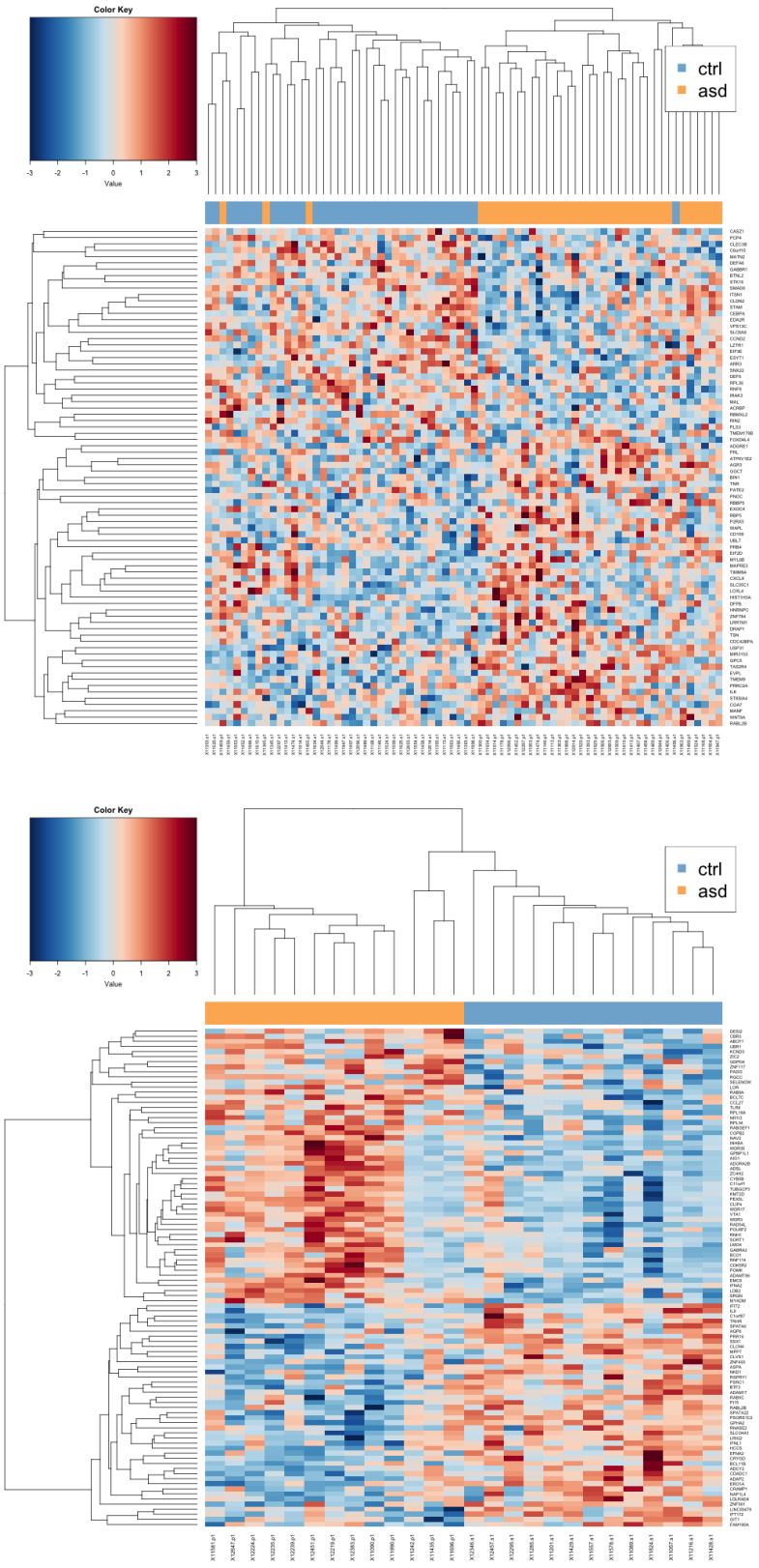


Figure 5.2. *Heatmaps of the closest to DM genes from both the siblings and the unrelated samples show that some difference exists between the groups. The number of genes in to visualize was based on the clearest result with hierarchical clustering using average linkage and Pearson correlation.*

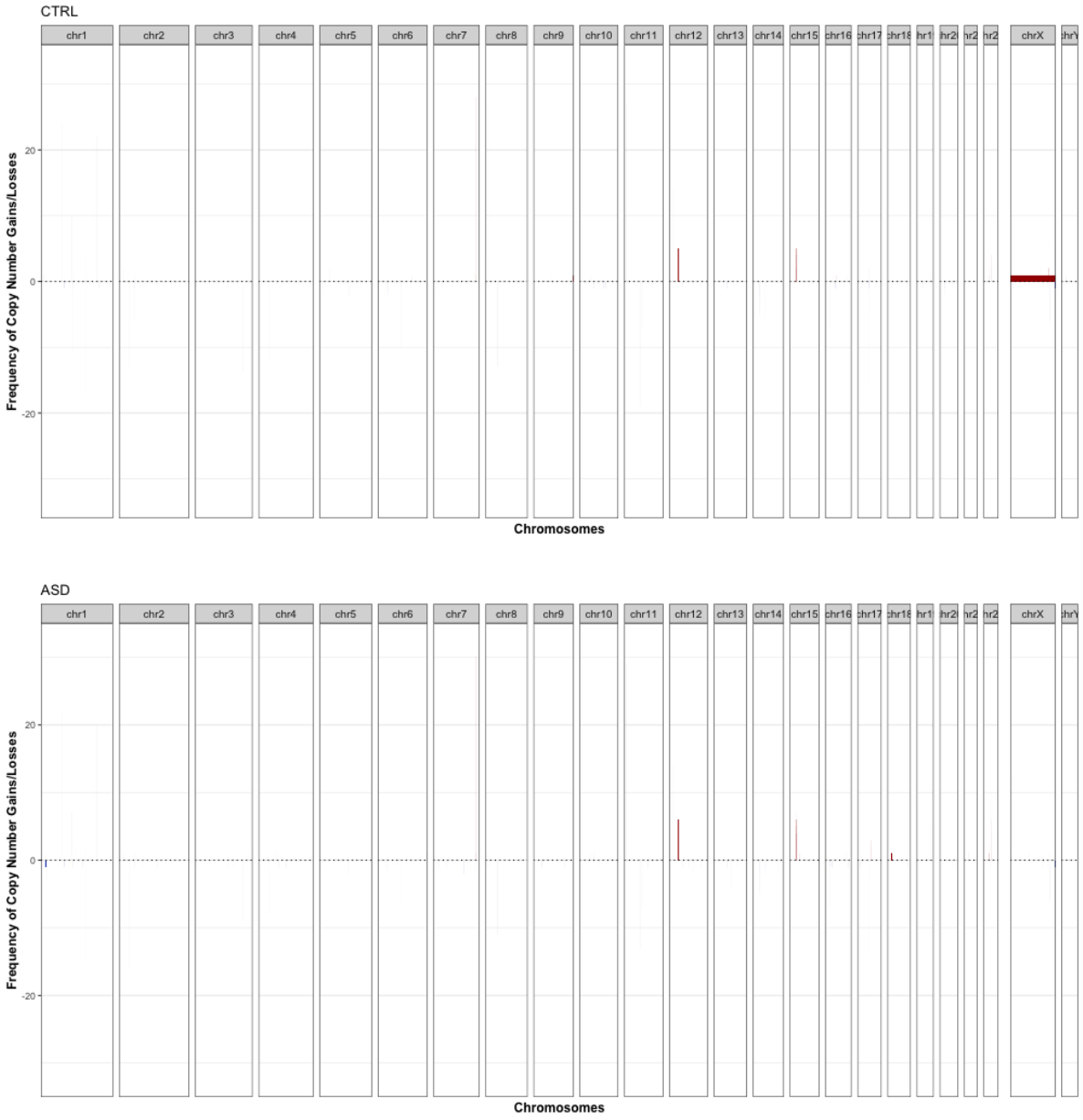


Figure 5.3. *Frequency plots of the healthy siblings (a) and probands (b) from the paired data shows some recurring gains (pictured red) in chromosomes 12 and 15. One sibling in (a) has Kleinerfelter syndrome.*

condition had also been identified in the original article (Levy et al. 2011), and had not been a reason to exclude the family from the study. Therefore, the sibling pair was kept for the remaining analysis.

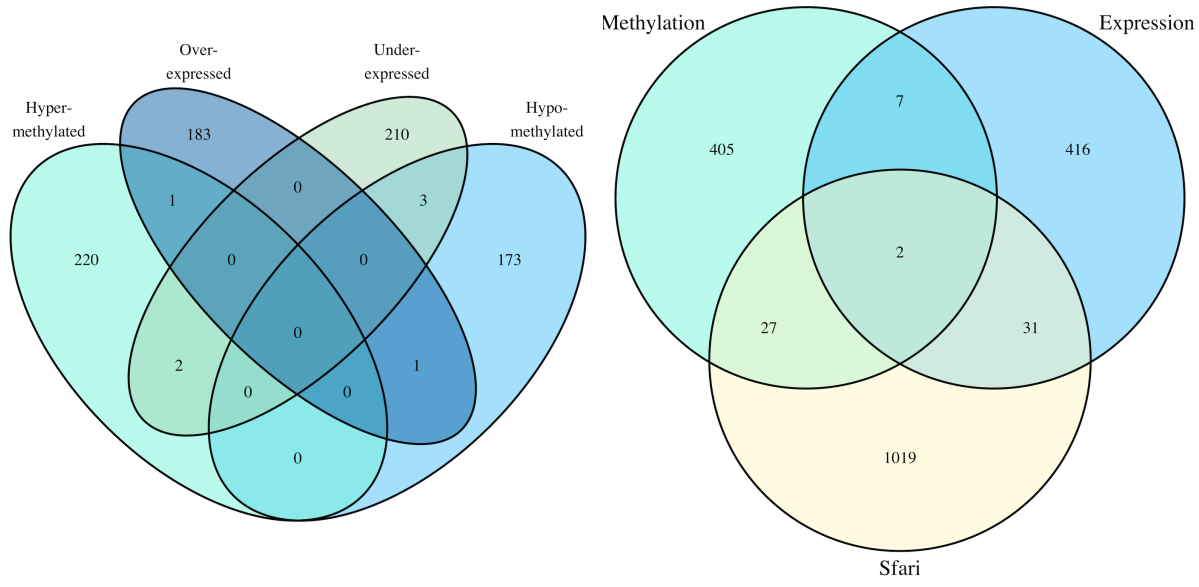
Both *figures 5.3a* and *figure 5.3b* show recurring gains in chromosomes 12 and 15. The loci in chromosome 15 has been well established as a risk loci for ASD, and the SFARI database (https://gene.sfari.org/database/cnv/15q11.2, read 11.4.2020) had 83 articles that associated it with ASD.

5.4 Step-wise data integration

Common ground between the data was investigated from the individual analysis of the methylation and expression datasets. Intersections between the closest to DE and DM genes are shown in the Venn -diagrams in *figure 5.4*.

Because there were no significant genes, the genes with a significant p-value before FDR were selected. The selection included 441 genes from the methylation data and 456 from the expression data, but only 9 were shared in these as can be seen in *reffig:venna*.

The separation of most differential genes between the groups into hypo and hypermethylated and under and overexpressed subgroups are shown in the venn diagram in *5.4b*.



(a) ***(b)***

Figure 5.4. *Venn diagrams showing the overlapping genes from the statistical analysis of the methylation and expression datasets. Figure 5.11a shows the hyper/hypomethylated and over/underexpressed genes. Figure 5.11b shows the intersection of genes identified as significant before FDR and the curated list of ASD associated genes from the SFARI database.*

A permutation test between the recurrent regions in the copy number data and the results from the differential expression and methylation data showed that the ASD population had slightly more recurring copy number alterations in these regions as figures *5.5a,b,c* and *d* show.

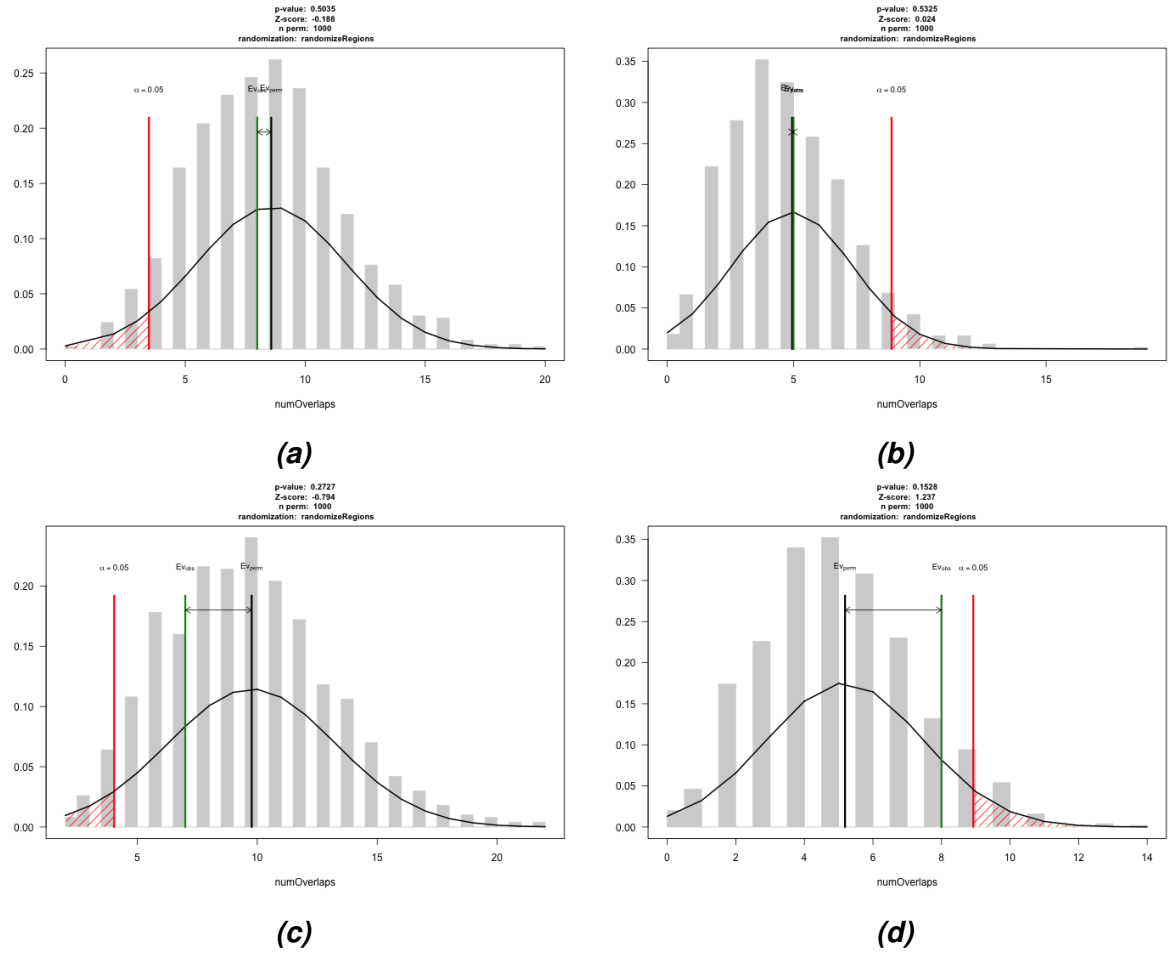
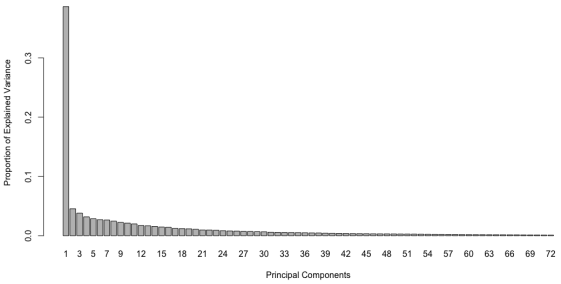


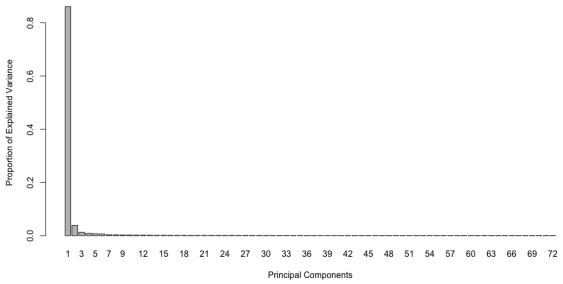
Figure 5.5. *Overlap analysis of the most variable genes from expression (a, b) and methylation (c, d) data with the recurring copy number regions between the asd probands (a, c) and siblings (b, d) shows that not many overlaps occurred in these regions. However, there is slight difference between the sample groups in the methylation data (c, d).*

5.5 Metadimensional data integration

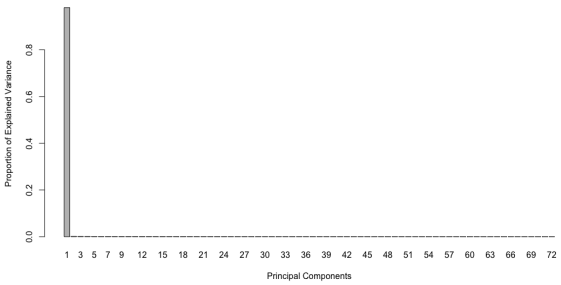
The decisions to be made for constructing the integration model include the number of components to use and the number of features to select. PCA was used to evaluate the amount of variance in the data components and visualized in *figures 5.6a,b* and *c*. Most of the variance in expression and methylation data appeared to be explained by the first component *(figures 5.6a* and *5.6c)*, whereas in copy number data this figure was just above 50



(a) *Copynumber*



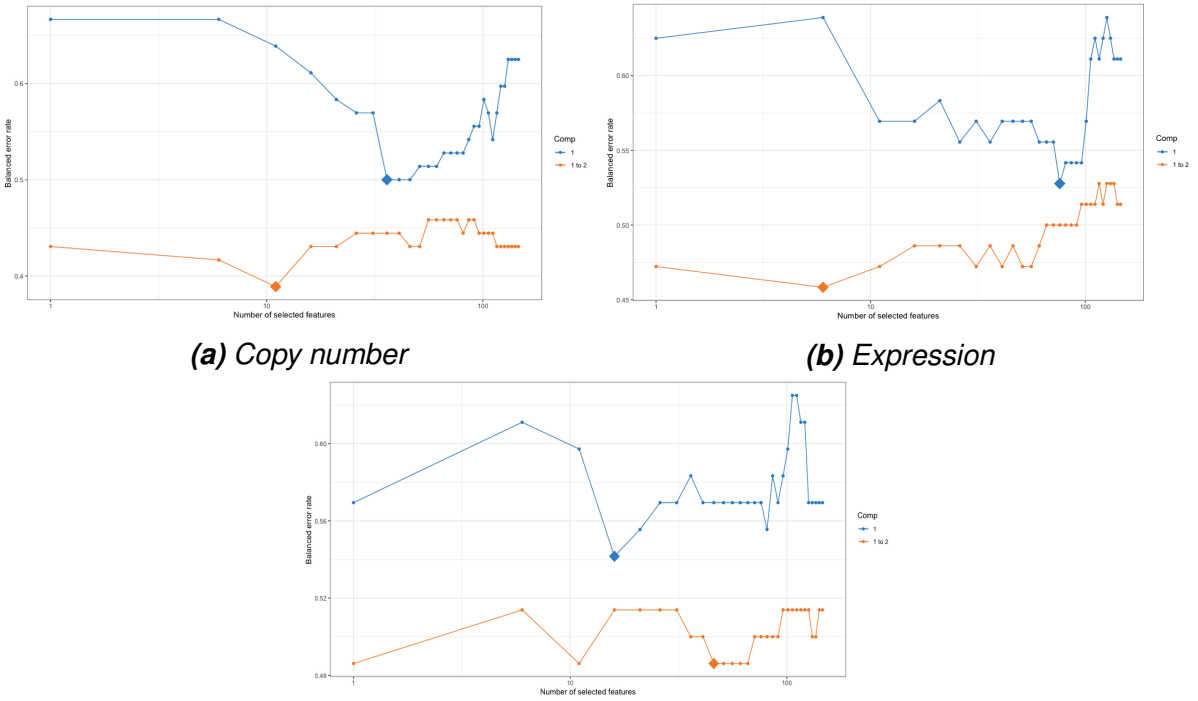
(b) *Expression*



(c) *Methylation*

Figure 5.6. *The variance explained by each of the datasets principal components as histogram shows that only the first component is informative in the methylation (b) and expression (c) datasets. In copy number data the first component explains less than half of the variance.*

The second step was to choose the number of features to select from the two components chosen for constructing the model. The number of features to select from each component in each dataset was evaluated using cross validation and the result was visualized in *figure 5.7*. Overall, 210 features were selected with 80 from the expression data (70 from first and 10 from second component), 100 from the methylation data (90 from the first and 10 from the second), and 30 from the copy number data (20 from the first and 10 from the second). The results from cross validation and the optimal number of features to select is visualized in *figures 4.10a-c*.



(a) *Copy number*

(b) *Expression*

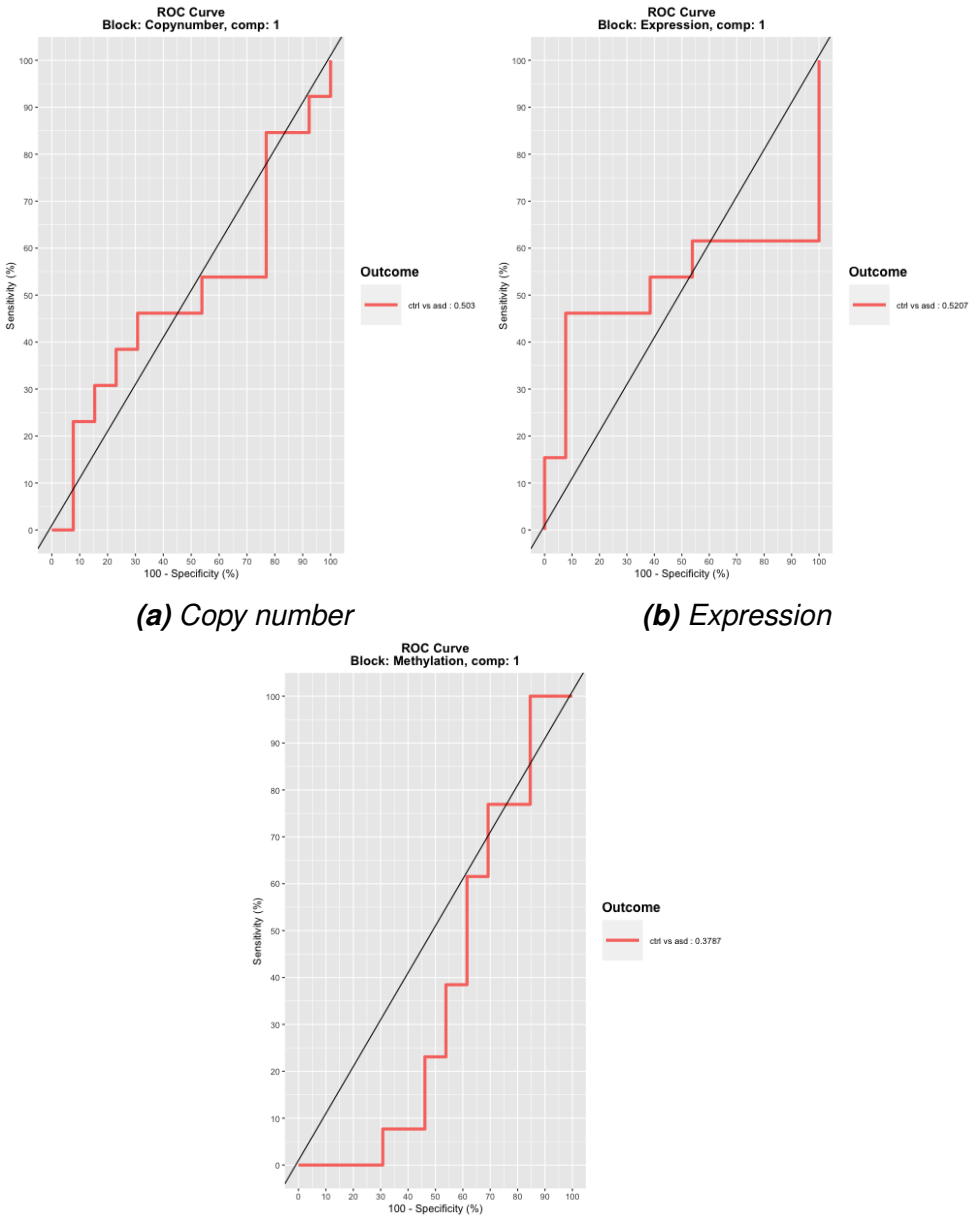
(c) *Methylation*

Figure 5.7. *The optimal number of variable to select from each component (1=blue, 2=orange) of each dataset (a-c) with x-axis denoting the number of variables and y-axis the resulting balanced error rate.*

Accuracy of the model was evaluated using the unrelated samples. It was not high and at best the model performed little better than a random classifier. The receiver operator curve (ROC) for each dataset shows that the first component from the expression data in the model separated the samples slightly better (AUC= 0.52) than those of copy number (AUC = 0.38) and methylation (AUC = 0.50). It appears from *table 5.1* that most of the samples get predicted as siblings.

Ground truth	Predicted as CTRL	Predicted as ASD
CTRL	5	8
ASD	1	12

Table 5.1. *Confusion matrix with the predicted and ground truth values*



(a) *Copy number*

(b) *Expression*

(c) *Methylation*

Figure 5.8. *The ROC curves show specificity (x-axis) and sensitivity (y-axis) for first components of the three datasets in the model. Expression data (b) performed better than copy number (a) and methylation (c).*

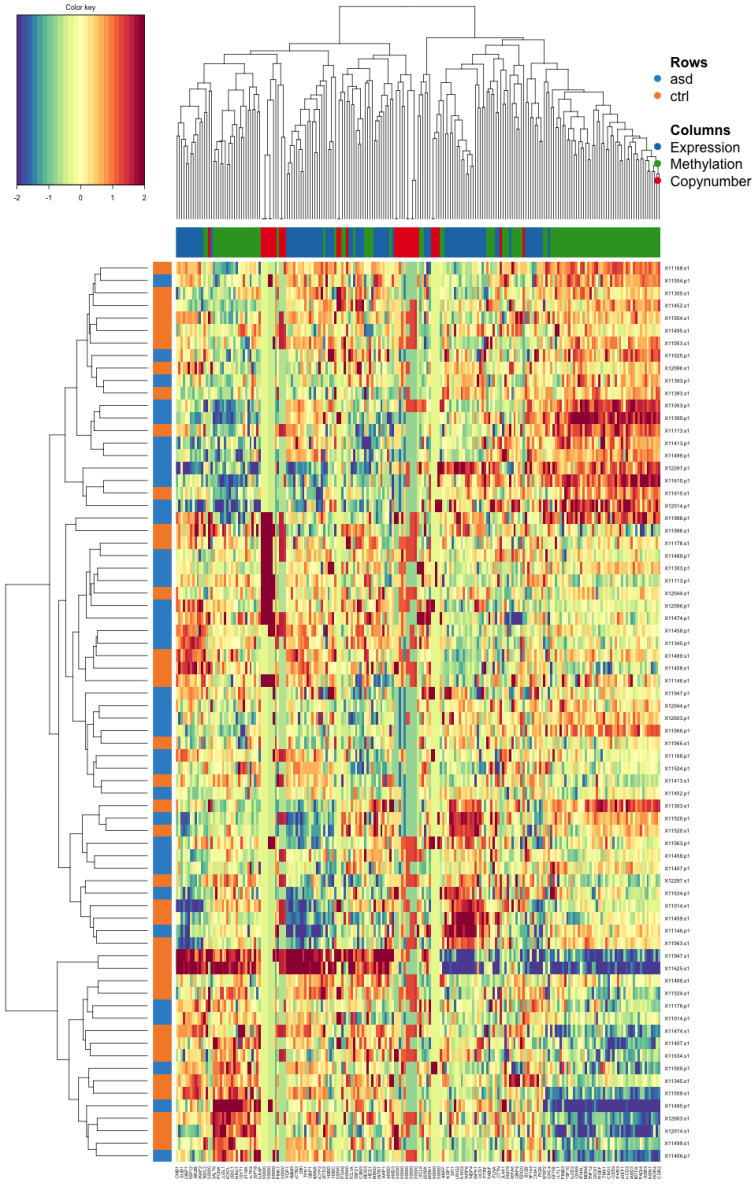


Figure 5.9. *Heatmap of the most discriminant features selected by sGCCA for the model shows that the samples (left) didn't cluster into distinct groups.*

Figure 5.10 shows the negative and positive correlation identified in the selected features by the sGCCA algorithm. The selected features were further visualized in a heatmap in *figure 5.9* with complete hierarchical clustering using Euclidean distance.

To take the data integration further, pathway enrichment analysis can reveal the functionality of the features identified in the sGGCA algorithm. Indeed, the selected genes seemed to be enriched in pathways involved in digestion and immunity as seen in *figures 5.11a and b*.

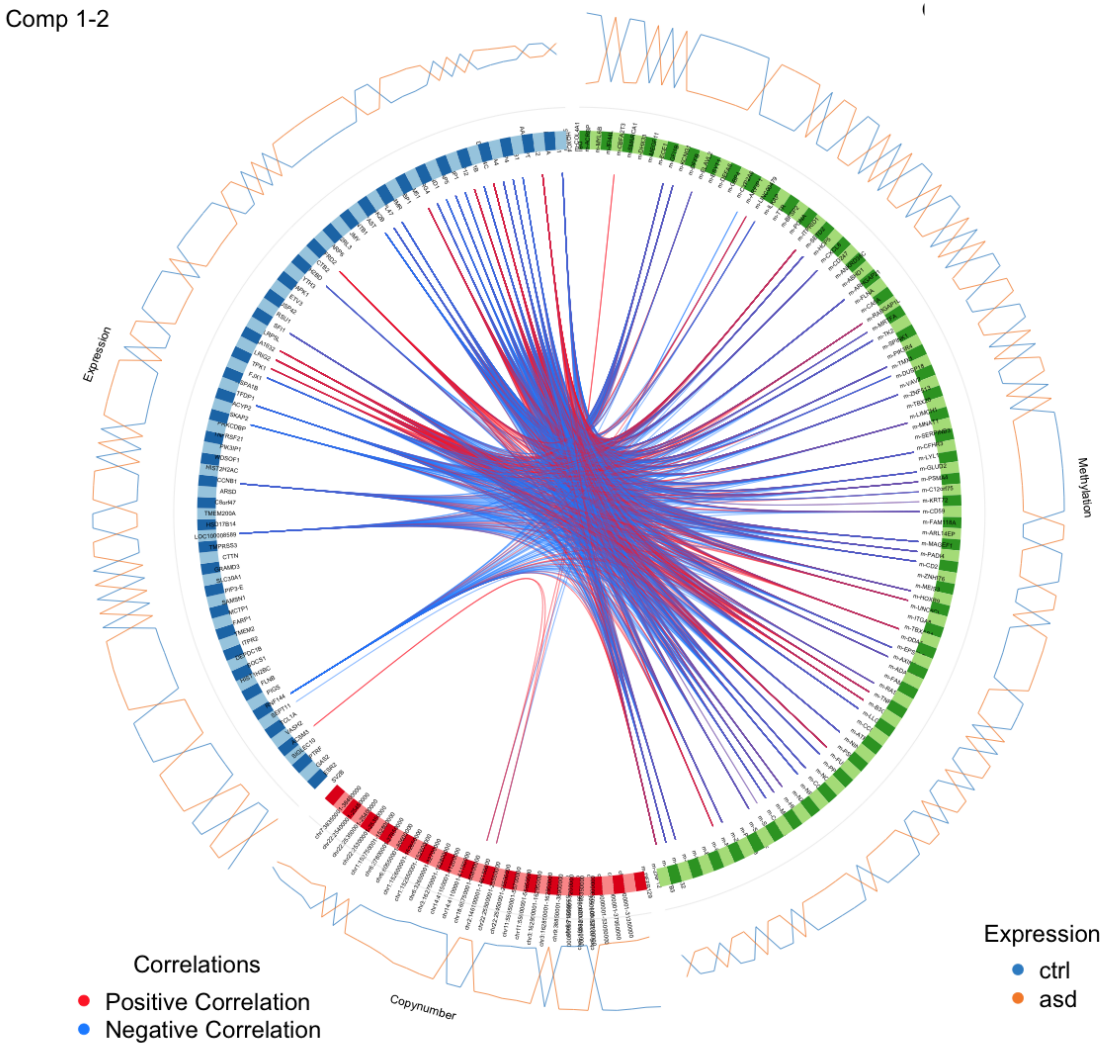
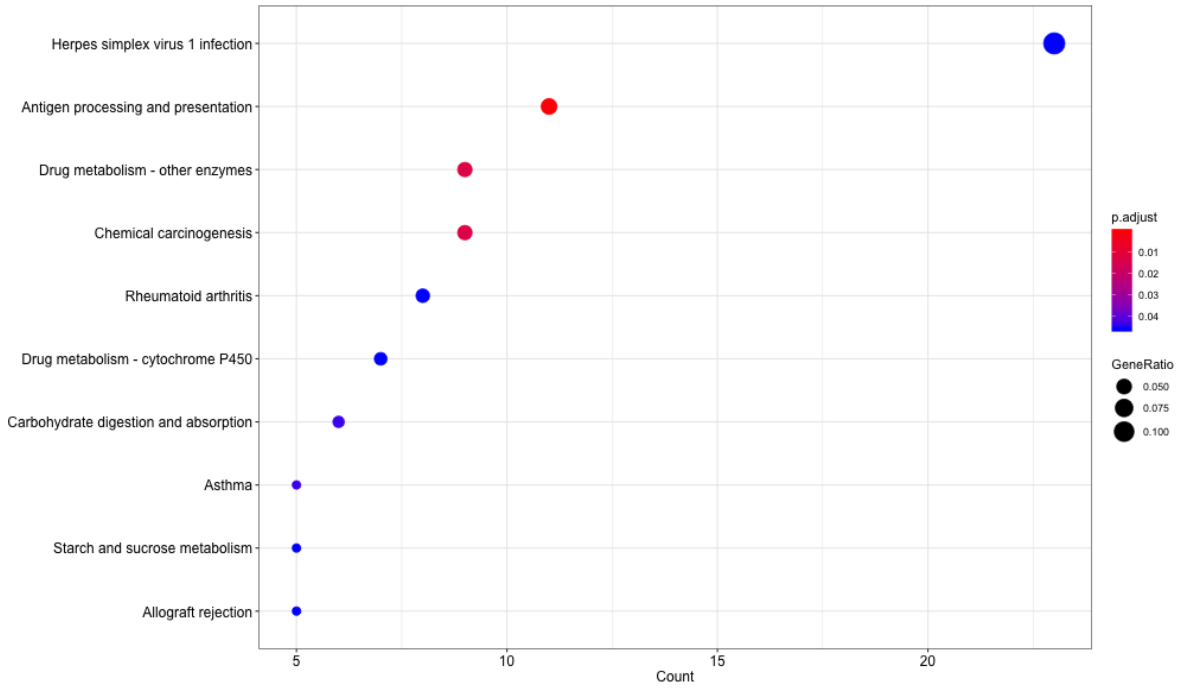
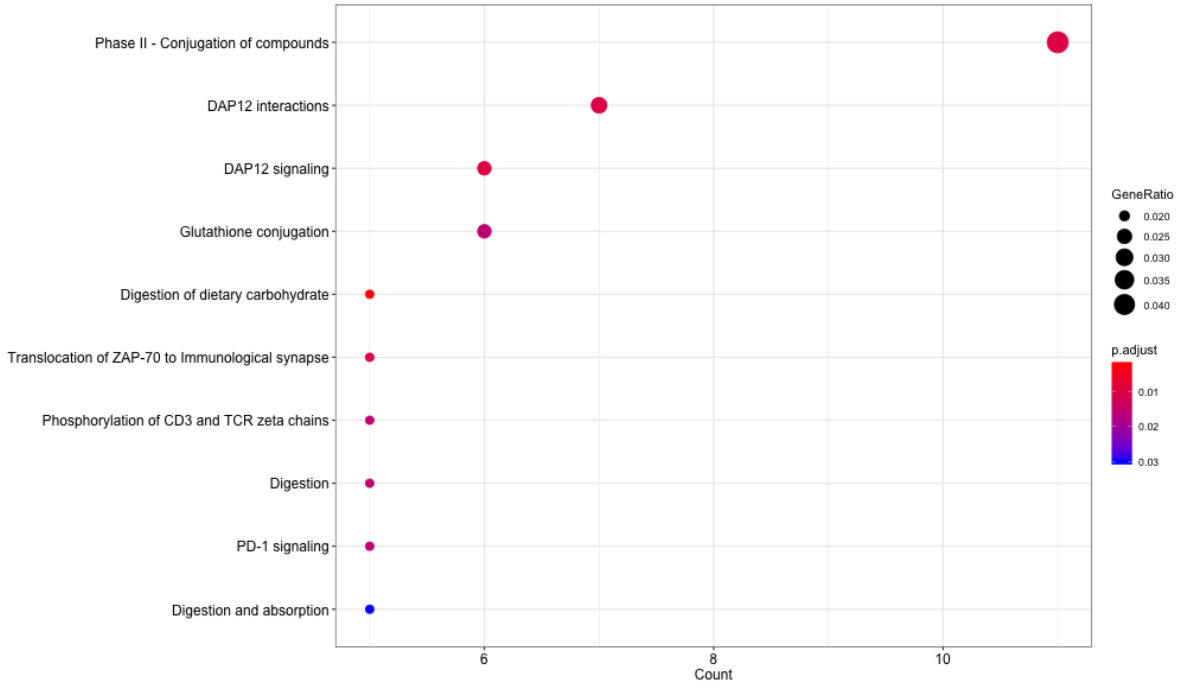


Figure 5.10. *The lines connecting the selected features denote correlation. Positive and negative correlation are shown in red and blue, respectively. Surrounding the circle is relative expression of the selected features between the sample groups (ASD shown in orange and CTRL in blue).*



(a) *KEGG*



(b) *Reactome*

Figure 5.11. *The results from hypergeometric tests for pathways from KEGG (a) and Reactome (b) both include immune related pathways.*

6 Discussion

This study supports the notion that studying ASD is not always straightforward (Betancur 2011). The differences between the sample groups in this data were minor, and perhaps more detailed knowledge of the sample groups, particularly the probands, could have yielded more accurate results. This information could include the severity level of ASD and details of the possible comorbidities. Also, due to the limited coverage of the expression and methylation datasets, some relevant information may have been missed.

Obtaining negative results from the individual analysis of the different datasets was not surprising, as the original articles accompanying the data had been focused on very different type of analysis. There were three articles on the aCGH data, two of which were involved with studying ASD, but concentrated on identifying *de novo* mutations in individual families (Levy et al. 2011; Gilman et al. 2011). The article on the expression data was determining the effect of copy number alterations from a separate aCGH study on the gene expression patterns (Luo et al. 2012). Their approach of obtaining DE genes was based on Z-statistic from the *scale* function in R, but in this study the paired t-test used due to its robustness and the paired nature of this study (Stevens et al. 2018). The article published on the methylation data was not about ASD at all, and instead focused on the age-related methylation patterns during the childhood and adolescence (Alisch et al. 2012). There were no existing studies on the integration of these three datasets.

Determining the transcriptomic profile in the brain from the blood samples is not feasible, however, studying the gene expression in the lymphocytes of autists can give vital information on the neuroinflammation process that has been shown to occur during the manifestation of ASD (Matta et al. 2019). This occurs during the development of ASD,

although the persistence of a low-grade inflammation into the adulthood has also been reported (Estes and McAllister 2015). Indeed, the gene set enrichment analysis performed on the expression dataset (*supplementary table 3*) did involve many immune related pathways, and such were also present in the pathway enrichment analysis of the integrated data.

The process of constructing a model with the sGCCA algorithm involves selecting the number of features to be extracted from each dataset, and this can have a large impact on the performance of the model. Too many features may lead to overfitting and too few may not be informative enough. In this study, the cross-validation process for optimizing the number of features to select gave strikingly different results on each run, perhaps due to the low variation in the data between the probands and healthy siblings. Each dataset was very homogenous indeed, and perhaps for the expression and methylation arrays different normalization methods could be tested instead of the quantile normalization, which forced the samples to have the same distribution and could potentially involve large adjustments to the data (Bolstad et al. 2003).

Classification accuracy of the model was not good, and this could be due to the small differences observed between the sample groups, but also the small number of validation data. Although the model constructed with this data did not discriminate between the sample classes, the developed pipeline can be easily accommodated for other studies where different measurements from the same individuals are to be integrated. Indeed, in the GEMMA project, which this thesis study was a part of, multiple measurements will be made from ASD siblings and the pipeline is applicable with some adjustments (https://www.gemma-project.eu, read 29.4.2020).

7 Conclusions

In this study the main aim was to integrate different omics datasets for a more holistic understanding ASD and to develop a pipeline for this process. Although the model constructed with the available data did not predict the sample classes using the validation data, the integration method did seem promising in that the sGCCA algorithm was able to find correlation between the datasets and select features that likely contribute to the phenotype. The problem with the accuracy likely lies in the fact that no large differences existed between the study groups as was evident from the individual analysis.

Although this study did not find features that discriminate the ASD phenotype from the samples, the developed pipeline is applicable to further studies with multiple measurements from the same set of individuals. More information on the phenotypes would be useful, such as severity of ASD symptoms and details of possible comorbidities, particularly the gastrointestinal symptoms that elevate the peripheral inflammation levels in the subjects. Other data types could be added to the study, particularly the metatranscriptomic and metabolomic, and these directions, amongst others, will be explored in the GEMMA -project that this thesis study was a preliminary part of.

References

American Psychiatric Association, A. (2013). “Diagnostic and statistical manual of mental disorders”.

Kanner, L. (1944). “Early infantile autism”. *Journal of Pediatrics*, 211–217.

Lai, M.-C., Lombardo, M. V. and Baron-Cohen, S. (2014). “Autism”. *The Lancet* 383.9920, 896–910.

Myers, S. M. and Johnson, C. P. (2007). “Management of Children With Autism Spectrum Disorders”. *Pediatrics* 120.5, 1162–1182.

Glessner, J. T. et al. (2009). “Autism genome-wide copy number variation reveals ubiquitin and neuronal genes”. *Nature* 459, 569–573.

Shen, L., Lin, Y., Sun, Z., Yuan, X., Chen, L. and Shen, B. (2016). “Knowledge-Guided Bioinformatics Model for Identifying Autism Spectrum Disorder Diagnostic MicroRNA Biomarkers”. *Scientific Reports* 6 (1).

Liu, J., He, K., Wang, Z. and Liu, H. (2019). “A Computer Vision System to Assist the Early Screening of Autism Spectrum Disorder”. *Communications in Computer and Information Science* 1006, 27–38.

Isaksen, J., Diseth, T. H., Schjølberg, S. and Skjeldal, O. H. (2013). “Autism Spectrum Disorders – Are they really epidemic?”: *European Journal of Paediatric Neurology* 17.4, 327–333.

Hertz-Picciotto, I., Schmidt, R. J. and Krakowiak, P. (2018). “Understanding environmental contributions to autism: Causal concepts and the state of science”. *Autism Research* 11.4, 554–586.

Skuse, D. (2000). “Imprinting, the X-Chromosome, and the Male Brain: Explaining Sex Differences in the Liability to Autism”. *Pediatric Research* 47 (1), 9.

Kopec, A., Fiorentino, M. and Bilbo, S. (2018). “Gut-immune-brain dysfunction in Autism: Importance of sex”. *Brain Research* 1693. Where the gut meets the brain, 214–217.

Gazestani, V., Pramparo, T. and Nalabolu, S. e. a. (n.d.). “A perturbed gene network containing PI3K–AKT, RAS–ERK and WNT–-catenin pathways in leukocytes is linked to ASD genetics and symptom severity.” *Nature Neuroscience* 22 (10), 1624–1634.

Tick, B., Bolton, P., Happé, F., Rutter, M. and Rijsdijk, F. (2016a). “Heritability of autism spectrum disorders: a meta-analysis of twin studies”. *Journal of Child Psychology and Psychiatry* 57.5, 585–595.

– (2016b). “Heritability of autism spectrum disorders: a meta-analysis of twin studies”. *Journal of Child Psychology and Psychiatry* 57.5, 585–595.

Hansen, S. N. et al. (2019). “Recurrence Risk of Autism in Siblings and Cousins: A Multinational, Population-Based Study”. *Journal of the American Academy of Child & Adolescent Psychiatry* 58.9, 866–875.

Satterstrom, F. K. et al. (2020). “Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism”. *Cell* 180.3, 568–584.

Levy, D. et al. (2011). “Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders”. *Neuron* 70.5, 886–897.

Abrahams, B. S. and Geschwind, D. H. (May 2008). “Advances in autism genetics: on the threshold of a new neurobiology”. *Nature Reviews Genetics* 9, 341.

Fischbach, G. D. and Lord, C. (Oct. 2010). “The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors”. *Neuron* 68.2, 192–195.

Hsiao, E. Y. et al. (2013). “Microbiota Modulate Behavioral and Physiological Abnormalities Associated with Neurodevelopmental Disorders”. *Cell* 155.7, 1451–1463.

Liu, F., Li, J., Wu, F., Zheng, H., Peng, Q. and Zhou, H. (2019). “Altered composition and function of intestinal microbiota in autism spectrum disorders: a systematic review”. *Translational Psychiatry* 9.1, 43.

Matta, S. M., Hill-Yardin, E. L. and Crack, P. J. (2019). “The influence of neuroinflammation in Autism Spectrum Disorder”. *Brain, Behavior, and Immunity* 79, 75–90.

Auton, A. et al. (2015). “A global reference for human genetic variation”. *Nature* 526.7571, 68–74.

Bruder, C. et al. (2008). “Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles”. *The American Journal of Human Genetics* 82.3, 763–771.

Zarrei, M., MacDonald, J. R., Merico, D. and Scherer, S. W. (Feb. 2015). “A copy number variation map of the human genome”. *Nature Reviews Genetics* 16, 172.

Moore, L. D., Le, T. and Fan, G. (Jan. 2013). “DNA methylation and its basic function”. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* 38.1, 23–38.

Rizzardi, L. F. and Hickey, P. F. (2019). “Neuronal brain-region-specific DNA methylation and chromatin accessibility are associated with neuropsychiatric trait heritability”. *Nature Neuroscience* 22.2, 307–316.

Ciernia, A. V. and LaSalle, J. (2016). “The landscape of DNA methylation amid a perfect storm of autism aetiologies”. *Nature Reviews Neuroscience* 17.7, 411–423.

Jang, H. S., Shin, W. J., Lee, J. E. and Do, J. T. (May 2017). “CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function”. *Genes* 8.6, 148.

Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J. and Kimmel, M. (Sept. 2015). “Microarray experiments and factors which affect their reliability”. *Biology direct* 10, 46.

Maksimovic, J., Gordon, L. and Oshlack, A. (2012). “SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips”. *Genome Biology* 13.6, R44.

Pidsley, R. et al. (2016). “Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling”. *Genome Biology* 17.1, 208.

Bejjani, B. A. and Shaffer, L. G. (Nov. 2006). “Application of array-based comparative genomic hybridization to clinical diagnostics”. *The Journal of molecular diagnostics : JMD* 8.5, 528–533.

Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A. and Smyth, G. K. (Aug. 2007). “A comparison of background correction methods for two-colour microarrays”. *Bioinformatics* 23.20, 2700–2707.

Silver, J. D., Ritchie, M. E. and Smyth, G. K. (Dec. 2008). “Microarray background correction: maximum likelihood estimation for the normal–exponential convolution”. *Biostatistics* 10.2, 352–363.

Bolstad, B. M., Irizarry, R. A., Åstrand, M. and Speed, T. P. (Jan. 2003). “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias”. *Bioinformatics* 19.2, 185–193.

Johnson, W. E., Li, C. and Rabinovic, A. (Apr. 2006). “Adjusting batch effects in microarray expression data using empirical Bayes methods”. *Biostatistics* 8.1, 118–127.

Leo, A., Walker, A. M., Lebo, M. S., Hendrickson, B., Scholl, T. and Akmaev, V. R. (Nov. 2012). “A GC-Wave Correction Algorithm that Improves the Analytical Performance of aCGH”. *The Journal of Molecular Diagnostics* 14.6, 550–559.

Benjamini, Y. and Speed, T. P. (Feb. 2012). “Summarizing and correcting the GC content bias in high-throughput sequencing”. *Nucleic Acids Research* 40.10, e72–e72.

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G. and Jain, A. N. (2004). “Hidden Markov models approach to the analysis of array CGH data”. *Journal of Multivariate Analysis* 90.1, 132–153.

Venkatraman, E. S. and Olshen, A. B. (Jan. 2007). “A faster circular binary segmentation algorithm for the analysis of array CGH data”. *Bioinformatics* 23.6, 657–663.

Talevich, E., Shain, A. H., Botton, T. and Bastian, B. C. (Apr. 2016). “CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing”. *PLOS Computational Biology* 12.4, 1–18.

Hastie, T., Tibshirani, R. and Friedman, J. (2016). “The Elements of Statistical Learning”. Springer New York Inc.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. and Smyth, G. K. (Jan. 2015). “Limma powers differential expression analyses for RNA-sequencing and microarray studies”. *Nucleic Acids Research* 43.7, e47–e47.

Benjamini, Y. and Hochberg, Y. (1995). “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1, 289–300.

Subramanian, I., Verma, S., Kumar, S., Jere, A. and Anamika, K. (2020). “Multi-omics Data Integration, Interpretation, and Its Application”. *Bioinformatics and Biology Insights* 14, 1177932219899051.

Karczewski, K. J. and Snyder, M. P. (2018). “Integrative omics for health and disease”. *Nature Reviews Genetics* 19.5, 299–310.

Betancur, C. (2011). “Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting”. *Brain Research* 1380, 42–77.

Vlahou, A., Mischak, H., Zoidakis, J. and Magni, F. (2017). “Integration of omics approaches and systems biology for clinical applications”, 1–362.

De Bie, T., Cristianini, N. and Rosipal, R. (Jan. 2005). “Eigenproblems in Pattern Recognition”.

Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K. and Narasimhan, G. (2018). “So you think you can PLS-DA?”: *2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, 1.

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J. and Frouin, V. (Feb. 2014). “Variable selection for generalized canonical correlation analysis”. *Biostatistics* 15.3, 569–583.

Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J. and Lê Cao, K.-A. (Jan. 2019). “DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays”. *Bioinformatics* 35.17, 3055–3062.

Evangelou, M., Rendon, A., Ouwehand, W. H., Wernisch, L. and Dudbridge, F. (July 2012). “Comparison of Methods for Competitive Tests of Pathway Analysis”. *PLOS ONE* 7.7, 1–10.

Du, P., Kibbe, W. A. and Lin, S. M. (May 2007). “nuID: a universal naming scheme of oligonucleotides for illumina, affymetrix, and other microarrays”. eng. *Biology direct* 2, 16.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D. and Irizarry, R. A. (Jan. 2014). “Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays”. *Bioinformatics* 30.10, 1363–1369.

Alisch, R. S., Barwick, B. G., Chopra, P., Myrick, L. K., Satten, G. A., Conneely, K. N. and Warren, S. T. (Apr. 2012). “Age-associated DNA methylation in pediatric populations”. *Genome research* 22.4, 623–632.

Smyth, G. K. and Speed, T. (2003). “Normalization of cDNA microarray data”. *Methods* 31.4, 265–273.

Mei, T. S., Salim, A., Calza, S., Seng, K. C., Seng, C. K. and Pawitan, Y. (Nov. 2010). “Identification of recurrent regions of Copy-Number Variants across multiple individuals”. *BMC bioinformatics* 11, 147.

Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J. and Lê Cao, K.-A. (Jan. 2018). “DIABLO: from multi-omics assays to biomarker discovery, an integrative approach”. *bioRxiv*, 67611.

Gilman, S. R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M. and Vitkup, D. (2011). “Rare De Novo Variants Associated with Autism Implicate a Large Functional Network of Genes Involved in Formation and Function of Synapses”. *Neuron* 70.5, 898–907.

Luo, R. et al. (2012). “Genome-wide Transcriptome Profiling Reveals the Functional Impact of Rare De Novo and Recurrent CNVs in Autism Spectrum Disorders”. *The American Journal of Human Genetics* 91.1, 38–55.

Stevens, J. R., Herrick, J. S., Wolff, R. K. and Slattery, M. L. (2018). “Power in pairs: assessing the statistical value of paired samples in tests for differential expression”. *BMC Genomics* 19.1, 953.

Estes, M. L. and McAllister, A. K. (2015). “Immune mediators in the brain and peripheral tissues in autism spectrum disorder”. *Nature Reviews Neuroscience* 16.8, 469–486.

8 Appendices

Gene Symbol	logFC	AveExpr	t	P.Value	adj.P.Val
ATPAF2	-0.2378	6.5587	-4.1535	0.0002	0.6742
HIST2H2BE	0.3102	7.1717	4.1131	0.0002	0.6742
ZNF696	-0.2079	7.3158	-4.0878	0.0002	0.6742
C2orf63	-0.2673	6.7322	-3.9317	0.0003	0.7369
SLC25A13	0.3071	10.2186	3.6856	0.0007	0.9999
DENND5A	-0.2378	6.5587	-3.6114	0.0019	0.9999
OXTR	-0.3284	8.4793	-3.3273	0.0023	0.9999
VAPA	0.1933	6.5359	3.2688	0.0026	0.9999
BBS10	0.1242	6.4370	3.2250	0.0026	0.9999
KDM5B	0.3109	8.4759	3.2151	0.0028	0.9999

Table 8.1. *Small set of the top results from the differential expression analysis shows that there were no significant genes after adjusting for FDR.*

Gene Symbol	logFC	AveExpr	t	P.Value	adj.P.Val
MIR3153	0.1721	0.4033	3.8780	0.0004	0.9988
MYL6B	0.1874	-3.9189	3.6941	0.0006	0. 9988
CDC42BPA	0.1540	-4.2193	3.5629	0.0009	0. 9988
RBBP5	0.1432	-0.7356	3.5433	0.0010	0. 9988
C6orf15	-0.1341	-4.1323	-3.5306	0.0010	0. 9988
TMEM9	0.1288	-2.8403	3.5248	0.0010	0. 9988
WNT8A	0.1695	-4.3070	3.5151	0.0011	0.9998
RIN1	-0.1894	-4.4084	-3.5137	0.0011	0.9998
SNX22	0.3755	-2.3682	-3.4465	0.0013	0.9998
PNOC	0.1943	-1.4969	3.4439	0.0013	0.9998

Table 8.2. *Small set of the top results from the differential methylation analysis shows that there were no significant genes after adjusting for FDR.*

hsa04514 Cell adhesion molecules (CAMs)
hsa04640 Hematopoietic cell lineage
hsa04672 Intestinal immune network for IgA production
hsa04512 ECM-receptor interaction
hsa00980 Metabolism of xenobiotics by cytochrome P450
hsa00260 Glycine, serine and threonine metabolism
hsa04610 Complement and coagulation cascades
hsa04970 Salivary secretion
hsa04010 MAPK signaling pathway
hsa04630 Jak-STAT signaling pathway
hsa00564 Glycerophospholipid metabolism
hsa00601 Glycerosphingolipid biosyhhesis – lacto and neolacto series
hsa04974 Protein digestion and absorption
hsa04972 Pancreatic secretion
hsa04380 Osteoclast differentiation
hsa04976 Bile secretion
hsa00380 Tryptophan metabolism
hsa00512 Mucin type O-Glycan biosynthesis hsa04070 Phosphatidylinositol signaling system
hsa04062 Chemokine signaling pathway
hsa00830 Retinol metabolism
hsa02010 ABC transporters
hsa00561 Glycerolipid metabolism
hsa00562 Inositol phosphate metabolism
hsa00565 Ether lipid metabolism
hsa00500 Starch and sucrose metabolism
hsa04975 Fat digestion and absorption

Table 8.3. *Gene set enrichment analysis from paired expression data*

Description	Package	Version
Data import	GEOquery	2.52.0
Batch correction	SVA (ComBat)	3.32.1
Preprocessing, analysis	Limma	3.40.6
Preprocessing	Minfi	1.30.0
GC-correction	ArrayTV	1.22.0
Segmenting	DNAcopy	1.58.0
Copy Number Call	CNVkit (Python)	0.9.6
Liftover	Rtracklayer	1.44.4
Annotation	annotatr	1.10.0
CNV summary	CNVRanger	1.0.3
Venn diagram	VennDiagram	1.6.2
CNV visualization	GenVisR	1.16.1
Data integration	MixOmics	6.8.5
Pathway analysis	ReactomePA	1.28.0
Gene set enrichment	gage	2.34.0

Table 8.4. *Package dependancies*