UNIVERSITY OF CALIFORNIA SAN DIEGO

Exploring the Spectrum of Autism:
Machine Learning and Dynamical Analysis Approaches

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Neurosciences

by

Debha N. Amatya

Committee in charge:

Professor Fred H. Gage, Chair
Professor Neal Swerdlow, Co-Chair
Professor John Kelsoe
Professor Saket Navlakha
Professor Gene Yeo

2019

The dissertation of Debha N. Amatya is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California San Diego

2019

DEDICATION

To Sudha & Ramesh Amatya, who sacrificed willingly and encouraged lovingly on countless occasions along this journey. To Rusty Gage, who taught me that towering achievements are made of humility, effort, & human connections.

I dedicate this to you.

EPIGRAPH

We invoke your name, Manjushri,
to look deeply into the heart of things
so that we may understand the roots of suffering.

May we use the sword of understanding
to cut through the bonds of suffering,
thus freeing ourselves and other species.

— Invoking the Bodhisattvas' Names

TABLE OF CONTENTS

LIST OF SUPPLEMENTAL FILES

File 1. Supplemental Tables for Machine Learning Analysis, Amatya_machine_learning_tables.xlsx
File 1 contains:
1. Variant scoring criteria
2. Primary MSSNG classification performance
3. Secondary SSC classification performance
4. ASD+ and ASD- genes
5. Gene enrichment analysis
6. Gene ontology analysis
7. BrainSpan spatiotemporal enrichment analysis
8. Cell-type enrichment analysis
9. Performance generalization

File 2. Supplemental Tables for Dynamical Analysis, Amatya_dynamical_analysis_tables.xlsx
File 2 contains:
1. Sample metadata
2. MED differential gene expression
3. ASD differential gene expression
4. Comparison gene lists from literature for enrichment analysis
5. MED gene ontology analysis biological processes
6. MED gene ontology analysis cellular components
7. BrainSpan region-stage gene lists
8. BrainSpan spatiotemporal enrichment scores

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I offer my deepest gratitude to all the teachers, family members, and friends that have helped me on my educational path. Without their many kindnesses, this culmination of my education would certainly not be possible. In particular, several people deserve special thanks: I thank my parents, Sudha and Ramesh Amatya, for their constant love and support. I also thank my advisor, Rusty Gage, who gave me invaluable guidance and the opportunity to learn science in one of the finest environments imaginable for a student. I thank my dear friend and peer, Robert Kim, for his incredible knowledge, patience, and willingness to help me with the big and small challenges of my work. Finally, I would like to offer my humble thanks to my spiritual teacher, Sathya Sai Baba, whose guidance has always been a refuge from the many successes and failures of scientific research.

I am grateful for the advice and encouragement of my other doctoral committee members: Neal Swerdlow, Gene Yeo, Saket Navlakha, and John Kelsoe. Their knowledge spans the disciplines of genomics, psychiatry, and computer science, and I am extremely fortunate to have received guidance from such a diverse set of thinkers.

Machine learning, bioinformatics, and stem cell disease modeling were all new topics to me when I began working in the Gage Lab, and I am indebted to the graduate students, postdoctoral researchers, and staff scientists that have taken the time to share their expertise with me. Particularly, I would like to acknowledge Son Pham for encouraging me to take a rigorous approach to bioinformatics and computer science, both in the classroom and laboratory. I would like to thank Simon Schafer, Sara Linker, Carol Marchetto, Yeni Kim, Krishna Vadodaria, Isabelle Guimont, and Beth Coyne for their large contributions to my scientific growth and their roles in these projects. Outside of the lab, several individuals have also gone to great lengths to support my work. I thank Graham McVicker, Arko Sen, Yuansheng Zhou, Tatyana Sharpee, Timothy Tadros, Stephen Scherer, and Bhooma Thiruvahindrapuram for their critical feedback and scientific contributions. I thank M.L. Gage for her editorial comments.

The genome vectorization portion of this dissertation (Chapter 2), was coauthored with Simon Schafer, Timothy Tadros, Saket Navlakha, Bhooma Thiruvahindrapuram, Stephen Scherer, Graham McVicker, Fred Gage. I served as the primary author for these sections, and the presented findings are being prepared for publication in a peer-reviewed journal. Specifically for this project, I thank Jorge Aldana for his technical expertise in optimizing computational resources and Tatyana Sharpee and Terrence Sejnowski for providing access to these resources. Additionally, I would like to thank Autism Speaks and the Simons Foundation for providing the sequencing and phenotypic data used in this study. Financial support for this study was provided by the Mary Jane and Robert Engman Foundation and the JPB Foundation.

The iPSC dynamical analysis (Chapter 3) was coauthored with Sara Linker, Ana Mendes, Renata Santos, Galina Erikson, Maxim Shokhirev, Yuansheng Zhou, Tatyana Sharpee, Fred Gage, Maria Marchetto, and Yeni Kim. I also served as the primary author for these sections, and the presented findings are being prepared for publication in a peer-reviewed journal. The research was primarily supported by a grant of the Korea Health Technology R&D Project funded by the Ministry of Health & Welfare, Republic of Korea (HI18C1077). The project also received support from The Leona M. and Harry B. Helmsley Charitable Trust, The Robert and Mary Jane Engman Foundation, and JPB Foundation.

Chapter 2, in full, is a reprint of the material as it has been written in a manuscript that has been submitted for publication. The authors of this study are Debha Amatya, Simon Schafer, Timothy Tadros, Saket Navlakha, Bhooma Thiruvahindrapuram, Stephen Scherer, Graham McVicker, and Fred Gage. The dissertation author was the primary investigator of this paper.

Chapter 3, in full, is a reprint of the material as it has been written in a manuscript that has been submitted for publication. The authors of this study are Debha Amatya, Sara Linker, Ana Mendes, Renata Santos, Galina Erikson, Maxim Shokhirev, Yuansheng Zhou, Tatyana Sharpee, Fred Gage, Maria Marchetto, and Yeni Kim. The dissertation author was the primary investigator of this paper.

I have been fortunate to receive fellowship funding from the National Institutes of Health (F30 MH115584 and T32 MH020002) for the completion of this dissertation. This financial support was instrumental in making this work possible.

Lastly, I thank the individuals and families affected by autism spectrum disorder. Without their willingness to participate in research endeavors, such work would not be possible.

VITA

2014	Bachelor of Science *with Honors*, Stanford University
2019	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Gutschow MV, Mason JC, Lane KM, Maayan I, Hughey JJ, Bajar BT, Amatya DN, Valle SD, Covert MW. "Combinatorial processing of bacterial and host-derived innate immune stimuli at the single-cell level". *Molecular Biology of the Cell*. 2019 Jan 15;30(2):282-92.

Vadodaria KC, Amatya DN, Marchetto MC, Gage FH. "Modeling psychiatric disorders using patient stem cell-derived neurons: a way forward". *Genome Medicine*. 2018 Dec;10(1):1.

Linker SB, Hsu JY, Pfaff A, Amatya DN, Ko SM, Voter S, Wong Q, Gage FH. "BrainImageR: spatiotemporal gene set analysis referencing the human brain". *Bioinformatics*. 2018 Jul 13;35(2):343-5.

Amatya, DN. "Using Neuroimaging and Optogenetics to Better Understand the Neural Circuit Basis of Major Depression". Stanford Digital Repository. 2014.

Ferenczi EA, Zalocusky KA, Liston C, Grosenick L, Warden MR, Amatya DN, Katovich K, Mehta H, Patenaude B, Ramakrishnan C, Kalanithi P, Etkin A, Knutson B, Glover GH, Deisseroth K. "Prefrontal cortical regulation of brainwide circuit dynamics and reward-related behavior". *Science*. 2016 Jan 1;351(6268):aac9698.

Singh MK, Kesler SR, Hosseini SH, Kelley RG, Amatya DN, Hamilton JP, Chen MC, Gotlib IH. "Anomalous gray matter structural networks in major depressive disorder". *Biological Psychiatry*. 2013 Nov 15;74(10):777-85.

ABSTRACT OF THE DISSERTATION

Exploring the Spectrum of Autism:
Machine Learning and Dynamical Analysis Approaches

by

Debha N. Amatya

Doctor of Philosophy in Neurosciences

University of California San Diego, 2019

Professor Fred H. Gage, Chair
Professor Neal Swerdlow, Co-Chair

Autism spectrum disorder (ASD) is a neurodevelopmental condition that heterogeneously impacts core domains of human function, such as communication, socialization, and cognition. Underlying these symptoms are complex genetic changes in a diverse array of genes and pathways. Developing a mechanistic understanding of ASD remains a major challenge that, left unanswered, will continue to hinder progress in the clinical management of this increasingly prevalent disorder. New approaches are needed to improve our understanding of the neurobiology of ASD, as well as current clinical paradigms.

In this dissertation, I discuss how techniques from the quantitative fields of machine learning and dynamical systems theory may be applied to decipher the cellular and molecular complexity of ASD. In Chapter 2, I address how whole genome sequencing data can be analyzed through machine learning to predict ASD phenotypes. The machine learning analysis describes novel methodology to vectorize genomics data and build interpretable classification models that relate to well-supported findings in the literature. In Chapter 3, I show that neuronal activity can be modeled as a nonlinear dynamical system to yield novel measures of neuronal state and dysfunction. I provide evidence for the minimum embedding dimension (MED) as a marker of diminished dynamical complexity of electrical activity recorded from ASD patient-derived neurons. I also probe the clinical and gene expression correlates of MED, which overlap with known ASD risk genes and pathways related to neurodevelopment in cortical and deep brain structures.

The approaches outlined in this dissertation seek to transfer tools from different quantitative fields to identify convergent cellular and molecular findings that characterize ASD. Genome classification and dynamical analysis improve on existing methods for the identification of disease signatures and reveal a set of common biological pathways and brain regions in ASD.

Chapter 1

Introduction

1.1 The autism spectrum

1.1.1 An expanding definition

Autism spectrum disorder (ASD) is a heritable neurodevelopmental condition that is characterized by a wide range deficits in social, verbal, and behavioral domains [Mayes and Calboun, 1999, American Psychiatric Association, 2013]. This combination of debilitating symptoms impacts the core domains of human functioning and significantly damages the ability of a child to integrate into family, social, and educational communities.

As the name suggests, ASD encompasses a variety of clinical sub-syndromes that exist on a continuum from profoundly disabled, like the first autistic cohort reported by Leo Kanner in 1943, to cognitively high-functioning, such as the patients that Hans Asperger described in 1944 [Kanner et al., 1943, Asperger, 1944]. As time has passed, clinical definitions of autism have broadened, culminating in the creation of the ASD umbrella diagnosis in the latest Diagnostic and Statistical Manual of Mental Disorders [American Psychiatric Association, 2013]. ASD is often further complicated by a constellation of possible neurological, cognitive, gastrointestinal, and immune comorbidities [American Psychiatric Association, 2013]. The most common of these

include intellectual disability, anxiety, sleep disturbances, poor feeding, and seizures [Zachariah et al., 2017].

Therefore, with the aforementioned information, a typical presentation of a child affected by ASD may be as follows: A four-year old boy is brought to see his physician due to poor appetite, trouble sleeping, and worsening unresponsiveness to his name being called. The boy had an uncomplicated birth and was in good health as a toddler. The mother describes that the boy loves watching one particular episode of his favorite show and will cry despite consolation if the television is turned off prior to completing it. The mother states that he has not demonstrated interest in other shows or toys for the past two months. The boy was conversant at an age appropriate level during prior annual clinic visits, but remains silent during the exam today.

The prognostic outlook for most children diagnosed with ASD is as varied as their symptom profile. Though a small portion of children will respond dramatically to behavioral intervention and achieve normal development in adulthood, outcomes are significantly dependent on early diagnosis and difficult to predict. More commonly, adults with ASD reach some degree of independence but continue to live at home with considerable support for daily activities. Almost half of people with ASD will require living in a specialized healthcare setting with supervision and only a small degree of autonomy [Newschaffer et al., 2007].

1.1.2 Societal impact

As definitions and awareness have expanded, the prevalence rates of ASD have increased over the past few decades. According to the most current estimates, ASD impacts 1 of 59 children and 1 of 37 boys in the US, positioning it as a major pediatric health issue [Baio, 2014]. Due to its debilitating nature, early onset, costly treatments, and rising prevalence, ASD is a staggeringly expensive disease that rivals the economic impact of well-known diseases, such as diabetes and attention deficit and hyperactivity disorder. In 2015, the cost of ASD, due to direct medical, indirect medical, and loss of productivity, was estimated to be $268 billion dollars, growing to

nearly $500 billion by 2025 [Lavelle et al., 2014, Leigh and Du, 2015]. Therefore, ASD is a major source of personal and social suffering that has become a key driver for biomedical research efforts.

1.1.3 Current clinical paradigms

Clinical tools for ASD remain coarse, due to our rudimentary understanding of the biology underlying the disorder [Vorstman et al., 2017]. Though the rate of diagnosis has increased over the past two decades, the process remains mostly subjective and dependent on the emergence of behavioral abnormalities [Fountain et al., 2011]. Molecular markers, such as those used in other fields of medicine, do not yet exist in the neuropsychiatric management of ASD, and as a result, preventative care is not an option. Treatments are similarly imprecise and highly dependent on age at initiation [Butter et al., 2003]. The gold standard treatment for ASD is early intensive behavioral intervention (EIBI), a remarkably expensive cognitive therapy regimen that requires up to 40 hours per week of one-on-one treatment for multiple years [Magiati et al., 2007, Reichow et al., 2012]. Pharmacological interventions are mostly limited to severe cases and are limited to the management of peripheral symptoms, like aggression or anxiety. No medication has yet been shown to be disease modifying. Even after the diagnosis of ASD and initiation of a successful treatment plan, charting the course of recovery or progression is not well-understood. Therefore, the processes of diagnosis, treatment, and prognosis are all clouded by a high degree of uncertainty, much to the frustration of patient families and health professionals.

1.2 Autism genetics

Due to its strong heritable component, a major area of research is identifying the genetic basis of ASD risk. Broadly, this is accomplished by large-scale genetic sequencing efforts that seek to find variants that are disproportionately found in people with ASD. What follows is a

brief explanation of key findings related to ASD genetics studies, as well an introduction to the concept of disease classification based on genomics data.

1.2.1 Heritability and genetic complexity

Autism has been known to be a heritable disorder since the 1970s, when it was first reported that monozygotic twins had an increased likelihood of sharing a diagnosis [Folstein and Rutter, 1977]. Since then, numerous studies have echoed this finding and the modern estimated genetic contribution of ASD is 56% to 95% [Sandin et al., 2017, Colvert et al., 2015]. Despite this understanding, it has been much more challenging to identify the precise genetic loci that make up the heritable risk for developing ASD - a phenomenon often dubbed the missing heritability problem [Manolio et al., 2009]. It has been well-documented that variants associated with ASD target synaptic functioning, chromatin modifying, and regulatory pathways, but individual variants are extremely heterogeneous, with none accounting for more than 1% of idiopathic ASD cases [Geschwind and State, 2015, Sebat et al., 2007, **?**, Devlin and Scherer, 2012]. For ASD and other common traits, like height or intelligence, genetic complexity obfuscates the detection of individual risk genes.

A central goal of biomedical genomic analysis is to elucidate the relationship between disease phenotypes and their genetic underpinnings [Green et al., 2011]. For certain conditions, such as rare monogenic disorders or cancer, genome sequencing has already proven to be clinically useful in risk assessment, diagnosis, and treatment selection [Worthey et al., 2010, Bick et al., 2017, The Clinical Lung Cancer Genome Project and Network Genomic Medicine, 2013, Chapman et al., 2011, Sanford et al., 2018, Khambata-Ford et al., 2007, Slamon et al., 1987]. However, such success has remained more elusive in more common but complex conditions, which are often influenced by environmental factors and characterized by distributed genetic risk [Boyle et al., 2017, Manolio et al., 2009]. In these complex traits, variant burden is spread in a polygenic, or even omnigenic manner, such that diverse genetic perturbations can give rise to

similar pathway or macroscopic phenotypes. Modeling how risk is integrated across the genome remains a critical challenge for complex heritable disorders.

Thus far, research efforts have estimated that ASD involves potentially thousands of risk loci [**?**, Sandin et al., 2017, Colvert et al., 2015]. Whole genome sequencing enables the detection of a range of genetic alterations that occur in the ASD population. Consensus findings from these studies describe an enrichment of *de novo* or rare gene disrupting mutations, the role of genetic background and common variation, structural alterations, and the functional diversity of putative risk genes [Geschwind and State, 2015]. Unfortunately, the inability to understand how loci interactively impart risk for individual patients stymie efforts to map the pathogenic mechanisms that underlie clinical symptoms [Manolio et al., 2009, Geschwind and State, 2015]. Progress has been made in the development of polygenic risk scores, however such methods typically rely on the thresholded hits of genome-wide association studies and do not yet provide clinically actionable predictions [Weiner et al., 2017, Clarke et al., 2015]. New tools that interpret sequencing data to provide clinical insight for at-risk children are needed, affording them a greater chance for healthy development.

1.2.2 Machine learning for genomic classification

Research interest in ASD has sparked the creation of massive genomics databases that contain high quality DNA sequencing and clinical data for both cases and controls [C Yuen et al., 2017, Fischbach and Lord, 2010]. These growing databases coincide with the development of powerful and scalable machine learning methods that excel in finding discriminative patterns in big data. Unlike traditional association studies, machine learning methods are equipped to flexibly combine input features to optimize prediction [Libbrecht and Noble, 2015, Leung et al., 2016]. Machine learning-enabled programs have already demonstrated best in class (sometimes exceeding human-level) performance in a wide variety of tasks, such as image classification, natural language processing, and reinforcement-based tasks [LeCun et al., 2015].

A growing number of sequencing-based machine learning analyses have used a variety of data representations, including primary base pair sequence, genotype arrays, and RNAseq transcript counts, to predict biomedically relevant information, such as cancer diagnosis, RNA splicing patterns, and variant impact [Furey et al., 2000, Wang et al., 2005, Jaganathan et al., 2019, Zhou and Troyanskaya, 2015, Poplin et al., 2018]. A related field of study, polygenic risk scoring, has attacked the problem of estimating individual genomic risk in ASD; however, such techniques often depend on the selection of arbitrarily thresholded hits from GWAS and thus far have been more useful for complex trait association than prediction [Clarke et al., 2015, Weiner et al., 2017, Dudbridge, 2013]. In addition to genomics, other sources of data such as patient behavioral measurements, blood chemistry, imaging, and movement data have been utilized for the automated classification of ASD [Jacob et al., 2019, Heinsfeld et al., 2018, Emerson et al., 2017, Howsmon et al., 2017, Wall et al., 2012, Crippa et al., 2015].

Despite the existence of big genomic data and machine learning tools, automated genomic classification has not yet demonstrated robust and reproducible results for neuropsychiatric disease prediction. Genetic heterogeneity, low statistical power, and data dimensionality are common issues encountered in such studies [Geschwind and State, 2015]. In part, these problems stem from a lack of computationally efficient and biologically relevant representations of an individual genome [Leung et al., 2016, Bzdok and Meyer-Lindenberg, 2018, Libbrecht and Noble, 2015]. A vector cast in primary sequence or variant space may have a prohibitively large dimensionality, whereas smaller representations may not sufficiently encode the complexity of the disease signature. Machine learning applications in complex conditions like ASD may be facilitated by the development of genomic representations that balance dimensionality with biological information content.

1.3 Cellular studies in ASD

Our understanding of ASD neurobiology is limited by our knowledge of the human brain and its billions of individual components, neurons. Therefore, another major effort to understand ASD is rooted in the small-scale but deep exploration of neurons obtained from patients affected by ASD. These cellular studies offer an important bridge between findings from sequencing studies and clinical phenotypes, because they allow scientists to peer into the critical cellular units that translate molecular alterations into cellular disease mechanisms. This section will describe findings from post-mortem and iPSC-derived neuronal studies in ASD. I will conclude by discussing how stem cell models of ASD give us living, dynamic neurons that enable functional readouts, namely electrical activity.

1.3.1 Post-mortem analysis

Post-mortem tissue harvesting and analysis is the traditional means by which neuronal changes related to ASD can be studied. Though there are inherent advantages in using primary tissue, its non-functional nature typically limits studies to assessing morphology, cell-type abundance, or transcriptomic changes [Courchesne et al., 2011, Voineagu et al., 2011, Gupta et al., 2014]. Some hybrid studies have sought to take DNA sequencing results from large ASD cohorts and map them to potential transcriptomic consequences in large datasets that have serially RNA sequenced numerous brain regions at various timepoints of development [Willsey et al., 2013, Parikshak et al., 2013, Sunkin et al., 2012]. Together, these studies have consistently identified gene networks and pathways associated with synaptic functioning, immune regulation, cortical development, and cell motility [Voineagu et al., 2011, Gupta et al., 2014, Willsey et al., 2013, Parikshak et al., 2013, Luo et al., 2012, Garbett et al., 2008]. Additionally, some of these studies also report overlap with hits from DNA sequencing efforts. Unfortunately, the specific sets of differentially expressed genes varies across studies and a consensus transcriptomic profile

has not yet been identified. Therefore, post-mortem studies tell us that molecular heterogeneity is still present at the gene expression level, but it is likely that affected gene networks and pathways are conserved across cohorts.

1.3.2 iPSC technology and functional studies

The combination of molecular heterogeneity and the relative inaccessibility of human neural tissue has posed a major challenge to building a mechanistic understanding of ASD. In part, this challenge has been addressed by advances in stem cell biology that have enabled the generation of neurons from human subjects using induction and differentiation techniques [Vadodaria et al., 2018]. Therefore, a path forward in the cellular mechanistic exploration of complex heritable conditions, like ASD, is through neuronal studies derived from stem cell models of phenotypically homogeneous subgroups. Previous studies have shown a correlation between early brain overgrowth, as measured by head circumference and neuroimaging, and an increased risk for ASD [Courchesne et al., 2003, Hazlett et al., 2011, Shen et al., 2013]. In this subset of the ASD population, postmortem analysis has revealed an excess of neurons in the first three years of life, and seminal iPSC-derived neuronal studies have uncovered increased proliferation of neural progenitors and reduced synaptogenesis [Courchesne et al., 2011, Liu et al., 2017, Marchetto et al., 2016, Mariani et al., 2015].

A key advantage of iPSC-derived neurons is the ability to record functional readouts from otherwise inaccessible living tissue [Vadodaria et al., 2018]. Multielectrode arrays (MEAs) enable high-throughput, longitudinal recordings of extracellular electrical dynamics from populations of neurons at millisecond resolution, facilitating the analysis of neurons in vitro [Kreuz, 2013, Samengo, I., Elijah, D., 2013, Baudry and Taketani, 2006]. The analysis of electrical recordings typically involves the quantification of spiking related variables, such as firing rate, spike morphology, and network measures. Electrophysiological analyses of neurons derived from ASD patients have revealed functional defects, such as reductions in neuronal firing, disrupted

postsynaptic currents, and imbalanced excitatory/inhibitory tone [DeRosa et al., 2018, Liu et al., 2017, Marchetto et al., 2016, Mariani et al., 2015].

1.3.3 Neuronal activity as a dynamical system

Despite these advances, less attention has been devoted to the nonlinear dynamical analysis of neuronal signals, which arise from the activity of numerous interacting neurons [Röschke and Başar, 1988]. Traditional spiking analyses capture some important features, but more sophisticated methods from the field of dynamical systems theory may extract novel discriminative variables for the characterization of individual or group electrical activity. One such technique, minimal embedding dimension (MED) analysis, is a mathematical algorithm to determine the dynamical complexity of a time-series recording [Kennel et al., 1992]. In this framework, a neuronal network is thought of as a system comprised of a finite number of n differential equations that govern all possible states of the system. The purpose of MED analysis is to empirically estimate this n value by iteratively forming d embeddings on an electrical time-series recording, which are simply time delayed versions of the original recording. When the embedding dimension, d, matches the native complexity of the system, n, the system is described as being unfolded and the minimum embedding dimension is identified [Kennel et al., 1992, Takens, 1981].

Thus, MED is a quantitative indicator of the dynamical complexity of electrical activity. Subject recordings may be characterized by average MED values, similar to traditional spiking variables, and group-wise statistics or downstream analysis can be performed. Though this dissertation describes the first application of MED analysis to neuronal models to our knowledge, similar techniques have been applied to electroencephalogram and magnetoencephalogram recordings in autism, schizophrenia, bipolar, and depression [Akar et al., 2015, Bosl et al., 2011, Fernández et al., 2018, Jeong et al., 1998]. These studies conclude that dynamical complexity is elevated in mood disorders, such as bipolar disorder and depression, and it is decreased in developmental conditions, like schizophrenia and ASD.

The human brain is a complex nonlinear system that is comprised of multiple interacting components at a variety of spatiotemporal scales. This complexity poses a challenge to scientific investigation, but it also provides an opportunity to apply dynamical analysis tools to characterize neuronal activity in a new way.

Chapter 2

Genome Vectorization for Machine Learning Applications to ASD

2.1 Background

2.1.1 Approach

In this section, I propose a gene-based vectorized representation for the classification of genomes belonging to ASD and neurotypical subjects. In contrast to traditional genomic analyses, I describe a novel methodology that allows for the simultaneous consideration of gene variant risk across the entire annotated coding and non-coding genome. I demonstrate that variant burden vectors can be used to train highly sensitive and specific classification models for idiopathic autism that provide a new approach to the measurement of genome-wide risk in ASD. Additionally, I explore how machine learning models can be "reverse engineered" to identify putative ASD risk genes and mechanistic pathways that align with findings in the ASD genomics literature. Finally, I explore anatomical correlates throughout neurodevelopment that localize predicted ASD genes to prenatal cortical growth stages. Combining these factors, this study presents a new method for the delineation of genomic mechanisms through machine learning. Additionally, genome

vectorization presents a potential path forward for the holistic assessment of genomic ASD risk that could precede the development of symptoms and complement standard clinical evaluations, which remain costly and subjective [Lord et al., 2000, Vadodaria et al., 2018]. I believe that such a framework will prove useful for the molecular interrogation of ASD as well as other heritable traits and diseases. Prior to presenting the methods and results of this framework, key background concepts related to DNA sequencing and machine learning will be covered.

2.1.2 Whole genome sequencing

Whole genome sequencing (WGS) is a modern technology that allows for the comprehensive detection of DNA sequence variants in individual samples. Compared to earlier methods, like Sanger sequencing, WGS is high-throughput, automated, and much less costly [Metzker, 2010]. Additionally, unlike single nucleotide polymorphism (SNP) sequencing methods, WGS allows for the detection of all types of sequencing variation (e.g. insertions, deletions, and structural variation), both common and rare [Cirulli and Goldstein, 2010]. The typical outputs of a WGS experiment are billions of short reads, which must be aligned to a reference genome to identify sequence differences in a given sample, known as variant calling. Because WGS covers both coding and noncoding regions, variants are found in exons, regulatory regions, and many unannotated regions.

The goal of variant calling is to identify sequence variants that are characteristic of your sample. In the biomedical research setting, this is accomplished by amassing large collections of case and control variant calls and performing association statistics for each variant to test for significance. However, due to the high dimensionality of the variant space and the finding that complex traits and diseases are characterized by many small effect size variants, WGS studies often fail to identify robust, replicable loci in conditions like ASD, thus limiting the usefulness of sequencing in research or clinical applications.

2.1.3 An algorithm for genome vectorization

A useful way to quantitatively describe individual variant call data is in the form of a vector of m dimensions or features. Vector valued quantities facilitate the comparison of observations and are required for numerical applications, like machine learning. Importantly, vectors also simultaneously represent relevant features, such that data can be holistically compared across many features at once. Most intuitively, I can describe each subject using a three billion dimensional vector that describes each base pair value in the genome. Or, instead, I could chose to focus only on variants to reduce M by a few orders of magnitude. Regardless, due to the large dimensionality of these vectors, computational calculations and statistical tests are very costly at this scale. In this chapter, I propose the construction of a gene based vector, which compresses the feature set to within an order of magnitude of the number of subjects, N, available in the largest ASD sequencing datasets, facilitating computation. The pseudocode for this simple algorithm, called genome2vec, is given below:

Algorithm 1: genome2vec for genome vectorization

Input: A set of variants S_i for N subjects, a scoring function F, and an index of genes M

Output: A set of gene-based variant burden vectors V

for $i = 1$ **to** N **do**
    $V_i \leftarrow$ a 0 vector of index M
    foreach $s_i \in S_i$ **do**
        $(index, impact) = F(s_i)$
        update $V_i(index) \leftarrow V_i(index) + impact$
return $V = \{V_1, V_2, ...V_n\}$

2.1.4 Classification theory

Vector valued inputs enable data analysis through an ensemble of tools called machine learning. The goal of all machine learning analyses are to automatically construct a model that

is descriptive of given data and predictive for future data. As such, machine learning models are particularly useful in contexts with large feature spaces, imperfect knowledge of feature importance, and large quantities of data [Domingos, 2012, LeCun et al., 2015]. Three key use cases of machine learning include regression, clustering, and classification. For the purposes of this dissertation, I will focus on classification, which is the task of predicting a categorical class membership from vector-valued data. In this case, genome classification refers to the prediction of ASD or control status from variant burden vector.

To construct a classification model, three key ingredients are needed: 1) data representation, 2) model optimization, and 3) model evaluation [Domingos, 2012]. Consider the popular naive Bayes classifiifier:

$$\hat{C} = \underset{i \in \{1,\ldots,K\}}{\operatorname{argmax}} \prod_{k=1}^{M} P(x_k|C_i) \tag{2.1}$$

In this probabilistic framework, classification is cast as maximum likelihood estimation problem. Data, X, is treated as a random variable dependent on the class, C, it is drawn from. Data features are considered independently and multiplied together based on prior models for each feature to identify the probabilistically maximal class. Evaluation is performed by compared predicted class, $\hat{C}$, to true class, C. Similarly, other classification models cast this problem in the mathematical framework of a separating hyperplane in support vector machines, decision tree in random forest classification, or curve fitting problem for logistic regression.

2.2 Methodology

2.2.1 Whole genome sequencing data

The primary data for this study were derived from the Autism Speaks MSSNG database (MSSNG), a large collection of paired genomic and clinical data for ASD and control subjects. The MSSNG database used in this study (DB5) contained whole genome sequencing data from

7,187 individuals. Most typically, MSSNG individuals belonged to a trio family structure, composed of neurotypical parents and a child affected with ASD (3,786 subjects), though other family structures were also present. Whole genome sequencing was performed for each subject using a mixture of Illumina and Complete Genomics platforms. Though Complete Genomics samples were associated with higher error, classification performance still remained acceptably high (Appendix Figure A.3A). Ninety-one percent of the DNA samples were acquired from blood, and the remaining samples were acquired from cell lines (5%) and white blood cells (4%) [C Yuen et al., 2017].

A secondary whole genome sequencing dataset, the SFARI Simons Simplex Collection (SSC), was used for replication of the machine learning analysis. The version of the SSC used in this study (phases 1, 2, and 2) contained 7,400 samples, composed of 1,850 quad families. The quad family structure contained one ASD proband, one neurotypical sibling, and two neurotypical parents. More sample information is provided in Table 2.1.

Recruitment, informed consent, and study protocols were governed by the respective institutional review boards of the original creators of the MSSNG and SSC databases. The present study was exempted from the institutional review board of the Salk Institute, due to the data having already been collected and de-identified prior to access.

2.2.2 Variant filtering

Annotated variants for the MSSNG subjects, which were called using an analysis pipeline described by Yuen et al. (2017), were acquired through MSSNG's Google Cloud BigQuery service [C Yuen et al., 2017]. Variants were included if they survived filtering criteria for allele frequency (minor allele frequency $\leq 10\%$) and effect (damage prediction, conservation, and known clinical relevance). The precise filters used for the creation of this annotated variant table are described in the original publication [C Yuen et al., 2017]. Additionally, the annotated variant table that was used for the bulk of this study excluded variant calls present only in control samples,

thereby focusing on disease-relevant variants only.

SSC whole genome sequencing variant calls were filtered as similarly as possible to the primary MSSNG dataset. Additionally, due to the case-control class imbalance caused by the quad family structure, random resampling was performed before model training to balance the number of ASD (proband) and control (mother, father, sibling) observations.

2.2.3 Variant scoring and gene-based vectorization

Filtered variants were subsequently scored using the Variant Effect Predictor (VEP) tool developed by Ensembl [McLaren et al., 2016]. VEP uses a wide range of bioinformatics databases to assess the impact of both coding and noncoding variation in sequencing data. VEP was used to score the consequence of each annotated variant on a scale of 1 to 4, as defined by the Sequence Ontology [Eilbeck et al., 2005]. A score of 1 was assigned to "Modifier" variants, e.g., intergenic variants or minor regulatory region modifications. A score of 2 was assigned to "Low" impact variants, e.g., synonymous substitutions. A score of 3 was assigned to "Moderate" impact variants, e.g., missense mutations and inframe insertions or deletions. Finally, a score of 4 was assigned to "High" impact variants, e.g., frameshift mutations or transcript ablations. A comprehensive catalog of variant types and VEP impact scores is included in Supplemental File 1, Table 1.

Gene-based vectorization was performed by calculating the average VEP score for each gene for a given individual's set of filtered variants. The gene was selected as the basis for vectorization for its relatively low dimensionality, yet sufficient specificity in capturing relevant disease-associated patterns. Averaging was performed to correct for differences in gene length and the number of variants per subject. In total, each subject variant burden vector had a dimensionality of 30,676 genes for the MSSNG data.

Alternative features spaces were constructed to assess the performance of higher and lower dimensional representations. Using the MSSNG filtered set, the variant-based vector was a binary presence/absence array of variants present in at least 50 subjects and thresholded by the

top half variance variants. For the same data, the chromosome-based vector was constructed by calculating the average impact score across the autosomal chromosomes for all subjects.

2.2.4 Implementation of machine learning classification models

The individual subject vectors were concatenated as rows to construct a $7,187 \times 30,729$ gene-based variant burden matrix. To further reduce the dimensionality of this matrix, unbiased variance filtering was performed to select the top half of higher variance genes. Machine learning models were trained on this native gene space matrix to facilitate model interpretation in later portions of the study. Principal component analysis was only performed for the visualization of vectors in a two-dimensional space.

Classifiers were trained using five different models: logistic regression (LR), support vector machines (SVM), multilayer perceptron (neural network), Naive Bayes, and random forest. The goal of each of these models was to predict the case-control status of a sample given its variant burden vector. For each model, training was performed using standard 10-fold cross-validation, during which 90% of the data was iteratively used for training and 10% was used for validation [Friedman et al., 2001]. Of note, a linear kernel was used for the SVM classifier and no priors were supplied to the Naive Bayes classifier. Exact model hyperparameters and trained models are available online at GitHub (see "Availability of Code" section). All models were implemented using Python 3 and its freely available modules, namely NumPy, SciPy, Pandas, scikit-learn, Keras, and TensorFlow [Jones et al., 2014, Oliphant, 2006, McKinney et al., 2010, Pedregosa et al., 2011, Chollet et al., 2015, Abadi et al., 2015].

2.2.5 Analysis of classification performance

Accuracy was measured as the fraction of correct case-control predictions on the validation or test data only. Accuracy was measured for each fold of cross-validation and displayed on

a boxplot. Also, using the trained models and validation or test data, I constructed receiver operating characteristic curves to assess model robustness of balancing sensitivity and specificity. A representative curve from the final fold of 10-fold cross-validation is shown in Figure 2.6.

2.2.6 Selection and investigation of salient classification genes

The SVM model was selected for further analysis, due to its high accuracy and consistency across cross-validation trials ($93 \pm 0.005\%$). A genome-wide ranking was assigned for each gene, based on the hyperplane weights learned by the model during training. The top and bottom quintile (SVM, ASD+ and SVM, ASD-) genes were chosen as representative gene lists for ASD-relevant and control-relevant genes, respectively. Both of these lists contained 3,067 genes. In a similar manner, top and bottom quintile genes were selected from the logistic regression model (LR, ASD+ and LR, ASD-).

These lists were compared to existing sets of putative ASD genes (Princeton), evidence-based ASD genes (SFARI), and highly brain-expressed genes using a one-tailed binomial test for overrepresentation [Krishnan et al., 2016, Abrahams et al., 2013, Uhlén et al., 2015]. Significance was set at $p \leq 0.10$, for each test. A set of highly expressed genes in the human liver was also included as a negative control in this analysis [Uhlén et al., 2015]. The exact gene sets and results for this portion of the study are included in Supplemental File 1, Table 5.

2.2.7 Gene ontology (GO) analysis

The SVM, ASD+ genes were further studied with the Panther Database online tool to identify biological processes, molecular functions, and cellular components involved with the selected list [Thomas et al., 2003, Thomas et al., 2006, Ashburner et al., 2000]. The two-tailed Fisher's exact test with false discovery rate correction was used to identify significantly enriched modules. A complete table of modules with $p \leq 0.05$ can be found in Supplemental File 1, Table

6.

2.2.8 Spatiotemporal enrichment analysis

Spatiotemporal enrichment of the SVM gene rankings was assessed using gene expression data from the BrainSpan Atlas of the Developing Human Brain [Sunkin et al., 2012]. Normalized gene transcript counts were acquired for brain samples that varied across multiple time points and neuroanatomical regions. Twelve developmental stages ranging from early prenatal to adulthood were included. Regionally, 16 discrete brain structures were included. For a more detailed description of the BrainSpan samples, documentation for the "Developing Transcriptome" dataset can be found at http://www.brainspan.org.

In a process similar to one described by Krishnan et al. (2016), representative genes sets were chosen for each region-stage pair by calculating the modified z-score of a given gene in the distribution of counts for all region-stage pairs [Krishnan et al., 2016]. The modified z-score was calculated using the median and median absolute deviation (MAD) in lieu of the average and standard deviation, because the median provides a better measure of centrality for the counts, which are typically not normally distributed.

Enrichment was calculated in two steps: 1) the identification of representative stage-region genes in the BrainSpan Atlas and 2) permutation testing to assess whether representative stage-region genes were highly ranked in genome-wide classifier rankings. The first step was performed with a modified z-scoring of genes using the median of reads per kilobase of transcript, per million (RPKM) count values (M) and median average deviation (MAD), due to the non-normality of counts data. The procedure follows:

1. Select a stage-region from the BrainSpan Atlas.

2. For each gene, i, in this stage-region, j, compute:

$$z_{i,j} = \frac{0.6745 \times (RPKM_{i,j} - M_i)}{MAD_i} \tag{2.2}$$

3. For stage-region j, select genes with $z_{i,j} \geq 2$ as the representative gene list.

To accomplish the second step, the gene rankings derived from the SVM model were set to an exponential scale, such that the topmost gene was assigned a value of 1 and the bottommost gene was assigned a value close to 0. Then, the following procedure was used to compute the observed ranking difference (d_{obs}) and a distribution of permuted ranking differences (d_{perm}):

1. Convert classifier gene ranking $rank_i$ to exponential rank, s_i, with the following equations:

$$s_i = \frac{b^{r_i} - 1}{b - 1} \tag{2.3}$$

with $b = 100$ and

$$r_i = \frac{N - rank_i + 1}{N} \tag{2.4}$$

with N set to the total number of genes in the ranking.

2. For stage-region j, select its representative genes. Also select a set of random genes of the same length.

3. Calculate:

$$d_{obs} = \bar{s_{r1}} - \bar{s_{r2}} \tag{2.5}$$

where $r1 \in$ representative genes and $r2 \in$ random genes.

4. For 100,000 iterations, shuffle the genes in the representative and random gene lists and calculate d_{perm}.

5. Determine significance with

$$z = \frac{d_{obs} - mean(d_{perm})}{SD(d_{perm})} \tag{2.6}$$

The p-value of the d_{obs} was calculated using the z-score derived from the d_{perm} distribution. P-values were adjusted for multiple comparisons with Bonferroni correction, and significance was assigned to adjusted p-values ≤ 0.10. Results for all 192 region-stage pairs are presented in Supplemental File 1, Table 7.

2.2.9 Cellular type enrichment analysis

Cell-type enrichments were calculated using a set of hand curated marker gene lists from a large-scale, human brain transcriptomic study [Kang et al., 2011]. Twenty cell types (primarily classes of GABAergic and glutamatergic neurons, but also glial types) were tested using a permutation test identical to the method for spatiotemporal enrichment. The cell-type enrichment results are presented in Supplemental File 1, Table 8 and Appendix Figure A.6.

2.2.10 Performance Generalization

Classification generalizability was tested by creating variant burden vectors that included all control and case variation that survived the previously mentioned damage, allele frequency, and conservation filters. Vectors from the MSSNG and SFARI dataset were standardized to include shared genes, and batch correction was performed with the ComBat tool from the Surrogate Variable Analysis (sva) toolbox [Leek and Storey, 2007]. Additionally, the SFARI SSC data was matched to the MSSNG data by only including parental controls and ASD probands in the analysis. Next dimensionality reduction was performed on the concatenated MSSNG and SFARI variant burden vectors with principal component analysis, retaining 99% of the variance in the

data. Finally, the MSSNG and SFARI vectors were separated, and the former was used for model training and cross-validation, while the latter was saved for model testing. The same models and performance metrics as described earlier in the methods were used. The full results for this analysis can be found in Supplemental File 1, Table 9.

2.2.11 Plotting and figure construction

Images were produced using R and its associated packages (ggplot2, pheatmap, CerebroViz). Aesthetic editing was performed using Adobe Illustrator.

2.2.12 Availability of code

The code for genome vectorization, machine learning model implementation, and miscellaneous functions are available on GitHub: https://github.com/damatya/genome2vec.

2.2.13 Availability of data

All data and results to support the findings of this study are presented in the results and Appendix materials. Primary data from the MSSNG and SSC databases are available through protected access from Autism Speaks (https://research.mss.ng/) and SFARI (https://www.sfari.org/resource/simons-simplex-collection), respectively.

2.3 Results

2.3.1 A machine learning framework for genomic classification

Classification is a general machine learning problem in which labeled data are used to train a model that can predict the classes of new, unlabeled data. In the context of this study, genomic classification refers to the prediction of ASD or control status from a representation of

an individual genome. Figure 2.1 describes an end-to-end genomic classification framework in which raw DNA sequencing data are automatically processed into a vectorized representation and passed into a classifier to yield individualized ASD risk prediction and insight into genetic risk features.

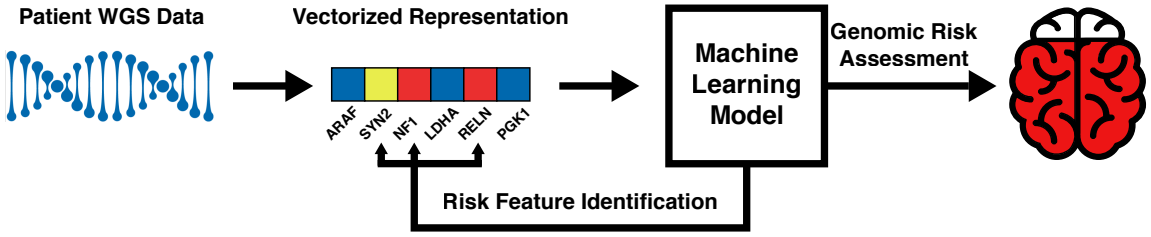


Figure 2.1: A framework for genomic classification of ASD. This schematic illustrates an end-to-end solution for a machine learning approach to the classification of ASD sequencing data. Whole genome sequencing (WGS) variant calls are transformed into a compact, vectorized representation that can easily be fed into a machine learning model. Trained models offer both insight into genetics mechanisms of ASD and assessment of genome-wide risk. Model investigation reveals how genes are utilized for prediction, allowing biological plausibility and implications to be assessed.

The creation of a trained genomic classifier requires 1) the availability of a large collection of labeled case-control genomes and 2) a computationally feasible and biologically relevant vectorized representation of the genome. To address the first requirement, I used the Autism Speaks MSSNG database, one of the largest joint repository of ASD and control genomes in existence (Table 2.1) [C Yuen et al., 2017]. The MSSNG data can be accessed at varying levels of analysis, ranging from raw sequencing reads to sample summary statistics. For this study, I utilized coding and non-coding subject variant calls that had been filtered for quality, allele frequency, and predicted impact (Figure 2.4). Of note, the filtered variants included both rare and common variants, up to a minor allele frequency of 10%. Unannotated variants, often found in non-coding regions, did not meet the filtering criteria and were thus excluded from this study.

Given the wide variety of feature scales (primary sequence, variant, gene, etc.) at which the analysis could be performed, as illustrated in Figure 2.3, the elimination of irrelevant or benign variation was a crucial step in reducing the dimensionality of the problem. To do so, variants were

Table 2.1: Whole genome sequencing samples. Many case and control samples are required to capture the diversity of genomic signatures during model training. The primary data analyzed in this study belong to the Autism Speaks MSSNG database, and a secondary validation dataset was acquired from the SFARI SSC. A key difference between the datasets is that MSSNG is primarily comprised of neurotypical parent controls and ASD probands, i.e. a trio family structure, whereas SSC is comprised of neurotypical parent and sibling controls and ASD probands, i.e. a quad family structure. Gender and group splits are written as a percentage of the dataset sample size. Age is given as the average age, plus or minus one standard deviation.

	Autism Speaks MSSNG ($N = 7,187$)		**SFARI Simons Simplex Collection** ($N = 7,400$)	
	Control	ASD	Control	ASD
Male	26.2%	37.4%	36.7%	21.7%
Female	26.2%	10.2%	38.2%	3.3%
Age (years)	41.9± 6.5	9.3± 4.8	N/A	9.6± 3.5
Data Structure	Trio: father, mother, proband (52.7%)		Quad: father, mother, sibling, proband (100.0%)	

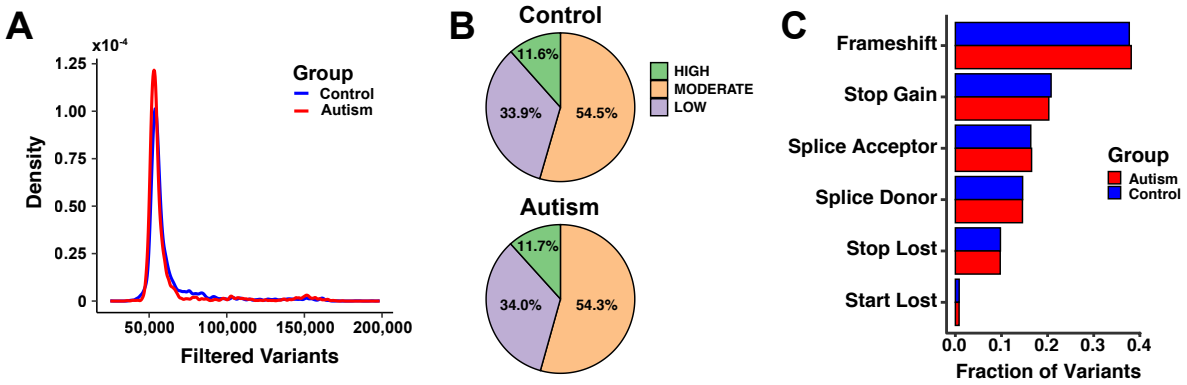


Figure 2.2: Characteristics of whole genome sequencing variants. A. Density plot of variants that survive filtering across groups. Coarse examination of variant call quantity indicates that both ASD (red) and control (blue) groups appear to be similar. B. A pie chart of variants, scored by damage impact, reveal similar distributions in controls (top) and cases (bottom). C. Closer examination of protein coding variants, binned by functional impact, also yield no differences between ASD (red) and control (blue) samples. Simplistic measures, such as variant quantity, predicted impact, and functional consequence, lack the resolution to separate the groups in a meaningful manner.

included in the primary training data only if they were present in at least one ASD sample call set. I relaxed this constraint for the final portion of this analysis (Figure 2.12), in which generalization across datasets was tested. After filtering for rareness and damage, the distribution of variant calls per subject was very similar between ASD and control subjects (Figure 2.2A). Additionally,

no significant differences were observed between categories of variant impact (Figure 2.2B), as measured by the Ensembl Variant Effect Predictor, or functional consequence of protein coding variants between cases and controls (Figure 2.2C). Together, these results indicated a lack of obvious differences in quantity and quality at the variant level between groups.

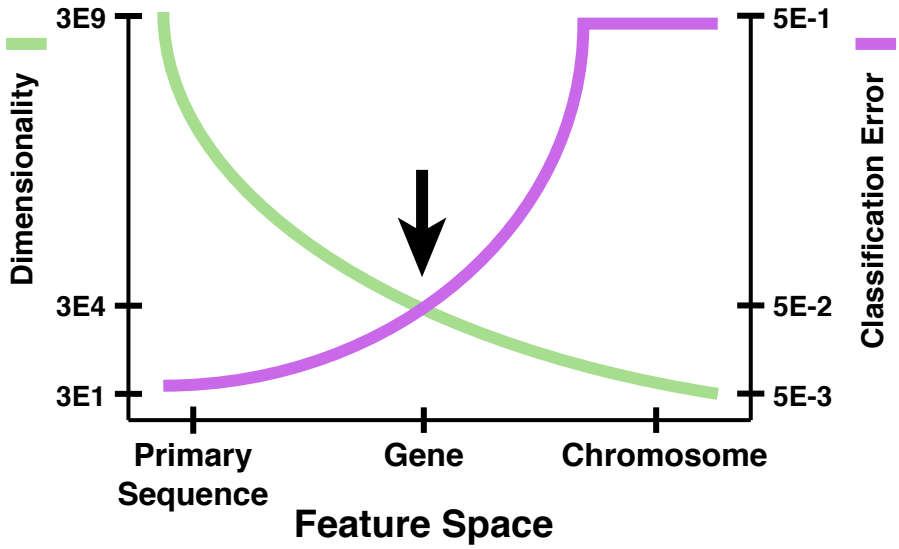


Figure 2.3: Feature selection of vector scale. This conceptual demonstration shows that the choice of dimensional scale for vectorization impacts the predictive value of the data. These axes may theoretically exist on a continuum of maximally fine-grained (e.g., base pair or variant level) to coarse (e.g., chromosomal blocks).

The MSSNG variant calls comprised the primary study data, and another major ASD sequencing database, the SFARI Simons Simplex Collection (SSC), was similarly prepared for replication of this analysis and generalization of methodology (Table 2.1). Both the primary MSSNG and secondary SSC data were subsequently vectorized and analyzed by machine learning.

2.3.2 Gene-based vectorization for whole genome sequencing data

The second necessary component for this analysis is a vectorized representation that can be utilized to train machine learning classification models. Vectorization requires the choice of a consistent set of genomic axes with which distinct samples can be represented, compared,

and analyzed. As shown in Figure 2.3, a vectorization at the base pair or variant level would maximally encode genetic risk at the cost of computational efficiency and statistical power. Due to the sparsity of high dimensional representations at limited sample sizes, overfitting to individual subject differences pose a major challenge to the training of generalizable classification models. In contrast, a vector that summarizes chromosomal variant burden would be computationally efficient but too coarse for biological predictions. Appendix Figure A.1 shows that both variant-based and chromosome-based vectorizations failed to separate cases and controls in the MSSNG data. Due to the intractability of higher-dimensional representations, our selection of a feature space was constrained to a more compact scale [Keogh and Mueen, 2011, Sherry et al., 2001]. A gene-based vector was a seemingly optimal trade-off of numerical efficiency and biological relevancy for this analysis (Figure 2.3). At a scale of roughly 30,000 genes, the vector dimensionality is sufficiently small such that statistical inference is feasible with the available sample size, and the convergence of variant burden at the gene level affords flexibility in modeling the heterogeneity of how biological states might be impacted in ASD [Crick, 1970].

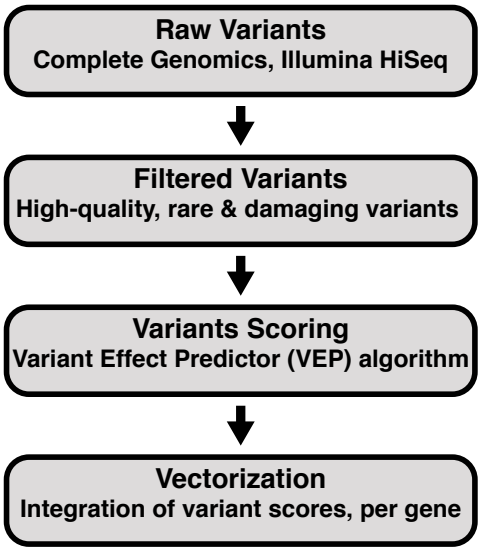


Figure 2.4: Pipeline for variant processing. Filtered variants that were selected for quality, rareness, and damage criterion were scored and averaged for each gene and subject.

I performed gene-based vectorization by scoring variants and summarizing burden by

gene for each subject, as described in Figure 2.4. Using the Ensembl Variant Effect Predictor algorithm, variants were scored on a four-tier scale: "Modifier", "Low", "Moderate", and "High" consequence (see Methods). Next, these labels were converted to numerical values and averaged per gene for each individual, resulting in our gene-based variant burden vector. Finally, the subject variant burden vectors were then subsequently stacked on each other to yield a variant burden matrix, which had the shape of 7,187 subjects $\times$ 30,729 genes. A portion of this matrix, in which ith subject and jth gene burden were colorized on a standardized scale, is shown in Figure 2.5A. Though differential genomic trends between cases and controls may be difficult to ascertain with the human eye, machine learning models excel at finding patterns in such data. To visualize the separability of the classes, I performed principal component analysis (PCA) on the variant burden matrix and plotted the top two principal components (Figure 2.5B). Two distinct clusters based on sequencing platform were observable (Appendix Figure A.2). More importantly, though the groups did not entirely form distinct clusters in this view, an ASD-control decision boundary was apparent, which was suggestive of amenability to classification.

2.3.3 Classification of variant burden vectors

I achieved classification of ASD and control variant burden vectors with five different models: logistic regression, support vector machines (SVM), multilayer perceptron neural network, naive Bayes, and random forest. Though these models differ in their implementation, each one functions by receiving a variant burden vector input and producing a class prediction as output. Prior to training, I filtered the variant burden matrix based on gene column variance to further reduce dimensionality in an unbiased manner. As shown in Figure 2.6, I constructed our classification models using a 10-fold cross-validation scheme that iteratively utilized 90% of the data for training and the remainder for validation.

All five models demonstrated high mean accuracies, ranging between 85% and 95%, which generalized well to the validation data (Figure 2.7A). Logistic regression, SVM, and the

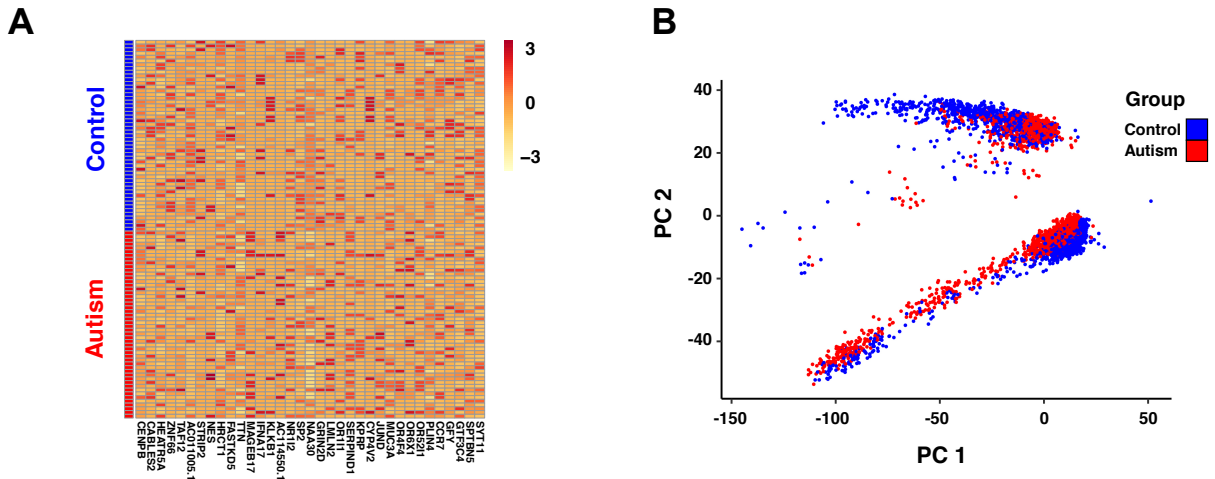


Figure 2.5: Visualization of variant burden vectors. A. Depiction of real vectors from MSSNG dataset. Each row represents an individual subject's variant burden vector and a set scale of gene columns. Damage is colorized on a standard scale, and blue and red row labels correspond to control and ASD subjects, respectively. B. Principal components (PC) analysis plot of variant burden vectors reveals a decision boundary between the ASD and control groups. Two distinct clusters are formed by samples sequenced on Illumina and Complete Genomics platforms, as shown in the Appendix.

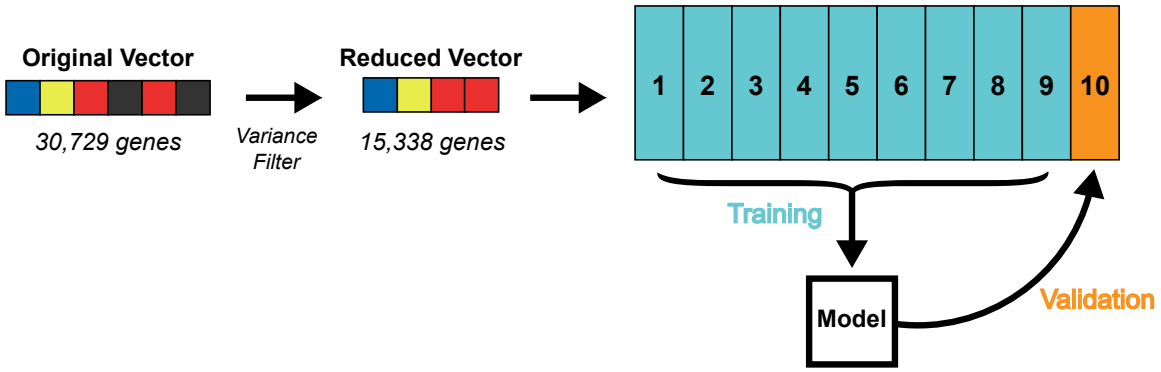


Figure 2.6: Classification model training and testing scheme. Vectors were purged of low variance genes then used to train classification models with 10-fold cross-validation. In this scheme, 90% of the data is iteratively used to train the model and 10% is left to test performance.

artificial neural network exceeded an average of 90% accuracy. Of these, the SVM model had the least variance across folds (93±0.005%, mean±SD accuracy across folds). Cross-validation accuracies for each classifier are described in Supplemental File 1, Table 2. Further, I examined classifier sensitivity and specificity through the receiver operating characteristic curves shown

in Figure 2.7B, revealing strong performance, particularly in the logistic regression, SVM, and neural network models. This level of performance suggests robust vector separability and the successful identification of ASD-relevant genomic patterns.

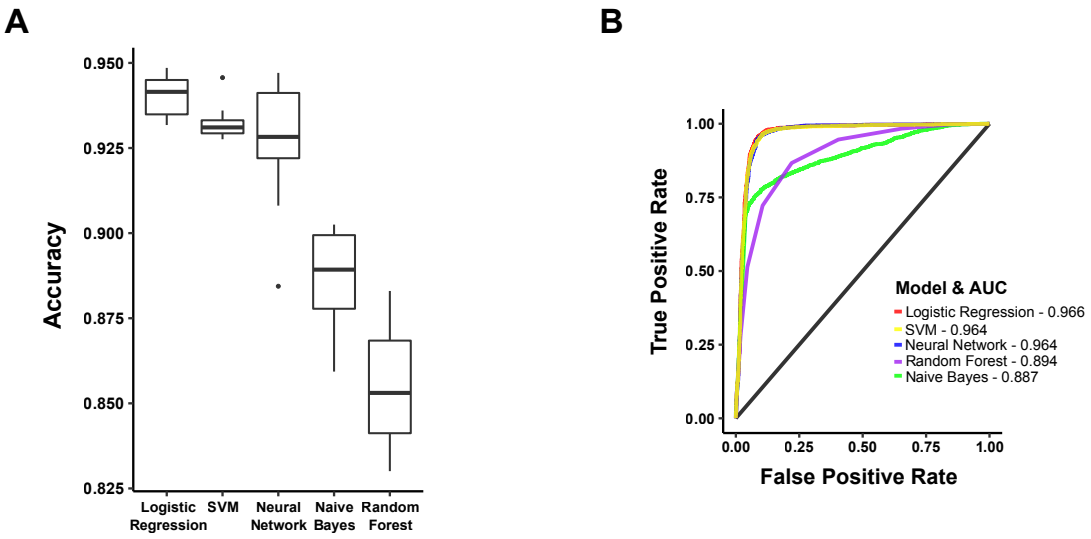


Figure 2.7: Classification model performance for MSSNG data. Vectors were purged of low variance genes then used to train classification models with 10-fold cross-validation. In this scheme, 90% of the data is iteratively used to train the model and 10% is left to test performance. A. Average model accuracy. Five different classification models performed well across cross-validation. Particularly, the SVM model demonstrated high accuracy with the most consistency across folds (93$\pm$0.005%, mean$\pm$SD). B. Representative receiver operating characteristic curves show balanced model sensitivity and specificity. The classifier curve colors are specified in the legend, and the black curve represents a random classifier. Area under the receiving operating characteristic curve (AUROC), a performance measure of binary classification, is also listed in the legend for each model.

The classification performance within various covariate groups (e.g., sequencing platform and gender) does not substantially differ in accuracy, as shown in Appendix Figure A.3. Therefore, the classification result is robust to variability in experimental conditions. Additionally, both scrambling of group labels and variant burden vector columns per subject were sufficient to break classification performance, indicating that the ground truth labeling and distribution of variant burden across the genome is structured in a meaningful way (Appendix Figure A.4).

I repeated variant burden vectorization and machine learning analysis in the SFARI SSC

to study the replicability of these methods in independent data. The random forest, logistic regression, and SVM classifiers converged on highly accurate solutions (greater than 85% mean accuracy over 10-fold cross-validation). However, the naive Bayes and neural network models were unable to achieve similar performance as their counterparts trained in the MSSNG cohort. In part, this gap may stem from an increased difficulty in learning the subtle differences between genomic patterns in healthy versus affected sibling pairs, which are present in this quad family structured data. Performance metrics for the SFARI SSC validation data are presented in Appendix Figure A.5 and Supplemental File 1, Table 3.

2.3.4 Relevance of salient classification genes to ASD neurobiology

A common issue in machine learning is the seemingly uninterpretable nature of complex models, often referred to as the "black box" problem [Castelvecchi, 2016]. Despite the performance presented in this paper, a thorough interrogation of the biological plausibility of the result is necessary for informing future clinical and biological applications. Therefore, the learned genome-wide weights in the SVM and logistic regression models were used as a measure for classification salience. The top and bottom weighted quintiles of genes were extracted to create lists associated with positive (ASD+) and negative (ASD-) ASD prediction, as shown in Figure 2.8. The complete list of ASD+ and ASD- genes for both SVM and logistic regression models can be found in Supplemental File 1, Table 4.

Given the presumed importance of the ASD+ lists, I hypothesized that these genes would be enriched for known ASD risk genes and pathways, as well as genes specifically relevant in brain function. Figure 2.9A shows that enrichment analysis of the classifier ASD+ list revealed significant convergence with genes implicated in ASD and brain function, as reported by SFARI and the Human Protein Atlas. Additionally, significant overlap ($p = 0.000524$, one-tailed binomial test) was found between the classifier ASD+ list and a set of putative ASD genes from a genome-wide prediction study by Krishnan et al. (2016) [Krishnan et al., 2016]. Conversely, classifier

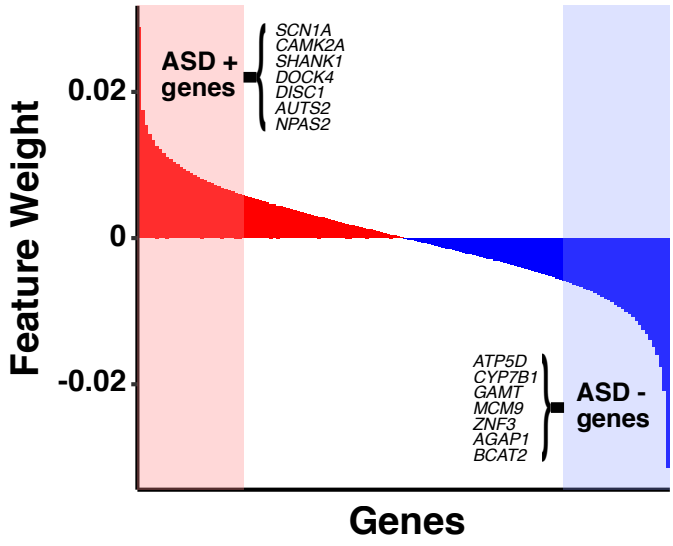


Figure 2.8: Extraction of salient classification genes. A quintile plot of SVM hyperplane weights is shown. The top quintile ASD+ list is composed of genes deemed to be important for ASD classification, and the bottom quintile ASD- list is attributed to a control class prediction.

ASD- genes were not enriched for SFARI and brain-related gene sets, and a negative control set of liver expressed genes was not enriched in either ASD+ or ASD- genes. Comprehensive results and gene lists for the gene enrichment analysis are in Supplemental File 1, Table 5.

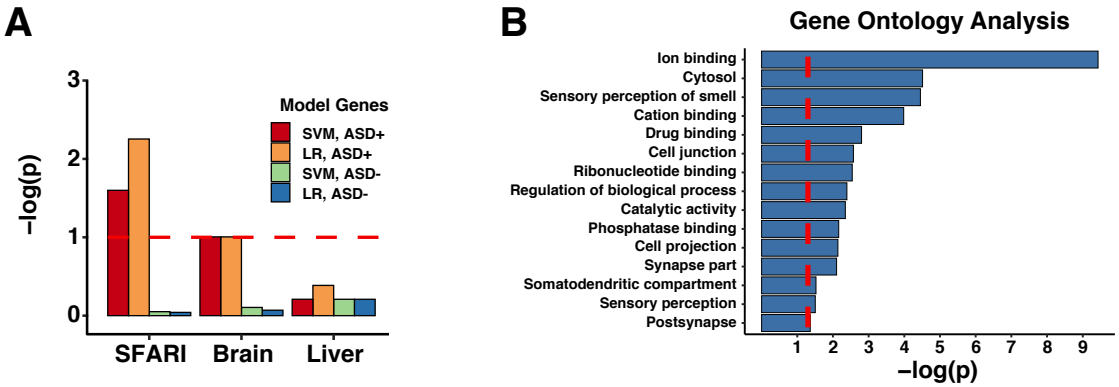


Figure 2.9: Analysis of ASD+ genes and biological pathways. A. ASD+ classifier genes are enriched for SFARI and brain-related gene sets. Enrichment was calculated using a one-tailed binomial test with a p-value significance cutoff of 0.10. B. Gene ontology analysis suggests neurodevelopmental pathway involvement. SVM, ASD+ genes were tested for enrichment using a two-tailed Fisher's exact test with false discovery rate correction with a significance cutoff of 0.05. A portion of relevant results include ion binding, synaptic, and sensory perception terms.

I conducted gene ontology analysis to examine the enrichment of biological pathways related to the ASD+ genes [Thomas et al., 2003, Ashburner et al., 2000]. Given the consistency of the SVM classifier, I tested the SVM, ASD+ genes for overlap with annotated biological processes, molecular functions, and cellular components. Figure 2.9B shows a selection of significant findings, in which specific biological processes associated with neuron development and function, such as ion binding, regulatory pathways, and synaptic terms, are among the most apparent categories. In contrast, the SVM, ASD- list was not associated with enrichment in any biological process. Complete results for the gene ontology analysis of the SVM, ASD+ genes are in Supplemental File 1, Table 6.

2.3.5 Spatiotemporal and cellular localization of salient classification genes

ASD is a complex neurodevelopmental condition that is hypothesized to impact cortical neural circuits during prenatal time points [Willsey et al., 2013, Parikshak et al., 2013, Zhou and Troyanskaya, 2015, Courchesne et al., 2018]. To asses the spatiotemporal relevance of the genes predicted by our model, I constructed a permutation distribution to test whether the genome-wide rankings corresponded to 12 developmental stages and 16 brain regions derived from the BrainSpan Atlas of the Developing Human Brain (Figure 2.10A) [Sunkin et al., 2012, Zhou and Troyanskaya, 2015]. This analysis revealed a significant involvement of early mid-fetal stages (13-18 pcw), particularly in a cluster of specific cortical regions that included the primary visual cortex (V1C), primary auditory temporal cortex (A1C), and inferior temporal cortex (ITC) regions, as well as the primary somatosensory cortex (S1C). The regional enrichments of the most significant developmental stage, early mid-fetal 2, are visualized *in situ* in Figure 2.10B. Primarily, cortical regions were associated with high genome-wide average rank. However, some inner structures, such as the hippocampus and striatum, were also significantly enriched in the ASD+ ranking.

I also examined the convergence of the SVM genome-wide rankings with cell type-specific

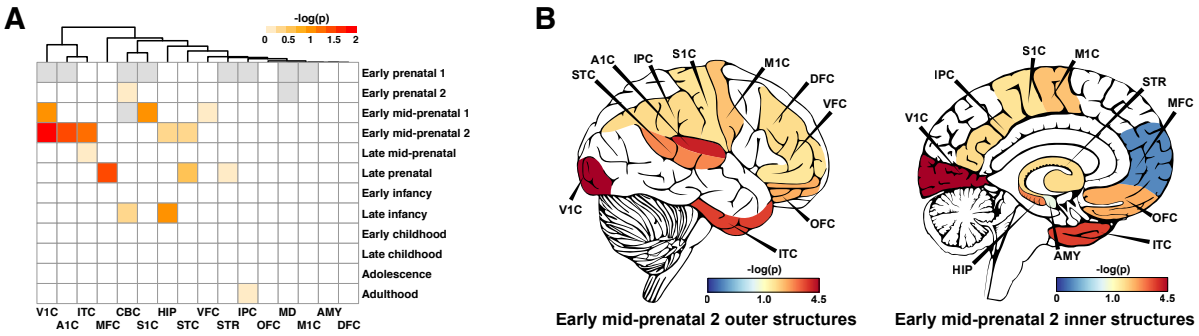


Figure 2.10: Brain-wide spatiotemporal enrichment of ASD+ genes. A. Cortical regions during early mid-fetal development are enriched for highly ranked classifier genes. Significance was calculated with a permutation testing procedure, as described in the Methods. In this heat map, the inverse log of the adjusted p-values is shown, after Bonferroni correction for multiple comparisons. The grayed out cells in the heat map correspond to brain structures absent in the early fetal brain. B. Neuroanatomical visualization of putative ASD brain regions. The raw permutation test p-values were plotted for the developmental stage with the most significance, early mid-prenatal 2 (16-18 pcw). Broad patterns of cortical enrichment are evident. The included brain regions are the primary visual cortex (V1C), primary auditory cortex (A1C), inferior temporal cortex (ITC), medial frontal cortex (MFC), cerebellar cortex (CBC), primary somatosensory cortex (S1C), hippocampus (HIP), superior temporal cortex (STC), ventral frontal cortex (VFC), striatum (STR), inferior parietal cortex (IPC), olfactory cortex (OFC), mediodorsal nucleus of thalamus (MD), primary motor cortex (M1C), amygdala (AMY), dorsal frontal cortex (DFC).

transcriptional programs [Kang et al., 2011]. Twenty neural cell types, including astrocytes, inhibitory and excitatory neuronal types, microglia, and oligodendrocytes, were analyzed for enrichment using permutation testing. Of these, two types of GABAergic interneurons (CCK and CALB2), GABAergic progenitors, and cortical layers 2-4 and layer 6 glutamatergic neurons were significantly elevated in our classifier rankings (Appendix Figure A.6). Further analysis revealed that excitatory and inhibitory cell type risk was not uniformly distributed across the ASD sample. These findings demonstrate developmental convergence on specific excitatory and inhibitory neuronal cell-types, consistent with previous reports [Willsey et al., 2013, Parikshak et al., 2013, Courchesne and Pierce, 2005]. Additionally, our results suggest that mechanistic heterogeneity may in part be a function of differing variant burden in diverse neural cell types.

2.3.6 Generalizable ASD Classification

Apart from interpretability, another major challenge in machine learning is model generalizability. That is, whether a classification model trained on one dataset is robust enough to perform well in another dataset. Thus far, I have demonstrated highly accurate performance during repeated training and validation within MSSNG and SFARI SSC cohorts that have been purged of control only variation (Figure 2.7 and Appendix Figure A.5). Here, I sought to assess classification generalization across MSSNG and SFARI datasets in the context of including all control and case variants to mitigate class bias (Figure 2.11). Please note, the MSSNG and SFARI datasets were balanced, such that primarily healthy parent controls and ASD probands (trio structures) were considered.

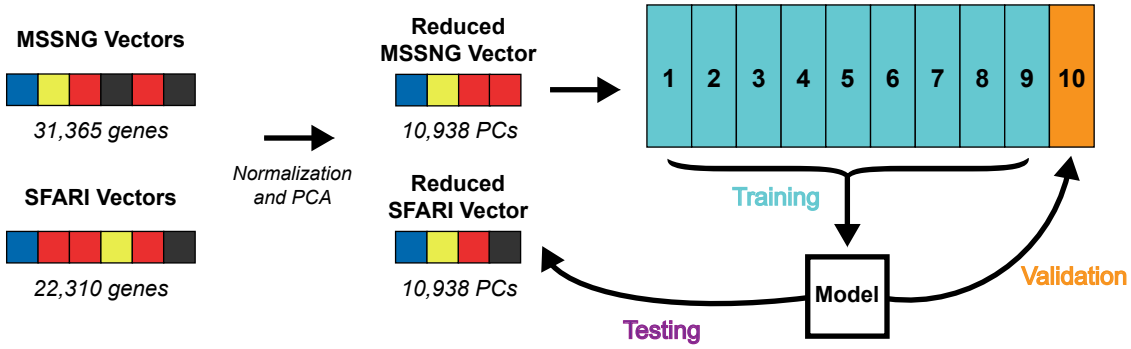


Figure 2.11: Training and testing scheme to analyze generalizability across MSSNG and SFARI datasets. The primary dataset used in this study excluded variants found only in controls, thus facilitating classification performance. Models were retrained using datasets that included all filtered variants from both ASD cases and controls, thus mitigating class bias. Variant burden vectors were transformed using batch correction and principal component analysis (PCA) with inclusion of 99% of data variance. The MSSNG data was used for training classification models with 10-fold cross-validation, and the SFARI data was used exclusively for testing.

In Figure 2.12A and B, I show that the Naive Bayes classifier accurately discerns ASD and control vectors when trained and cross-validated on the MSSNG dataset and tested on the SFARI dataset. The remaining models either performed poorly in cross-validation or were unable to generalize to the test data and might require further optimization or additional training to achieve performance similar to the Naive Bayes model. Importantly, this generalizable performance

was achieved by batch correction across datasets and dimensionality reduction through principal component analysis (Figure 2.11). The demonstration of moderate but generalizable classification performance is an important confirmation of the value of the genome vectorization technique, and it suggests utility for scenarios in which prior case or control status is unknown, such as the clinical setting.

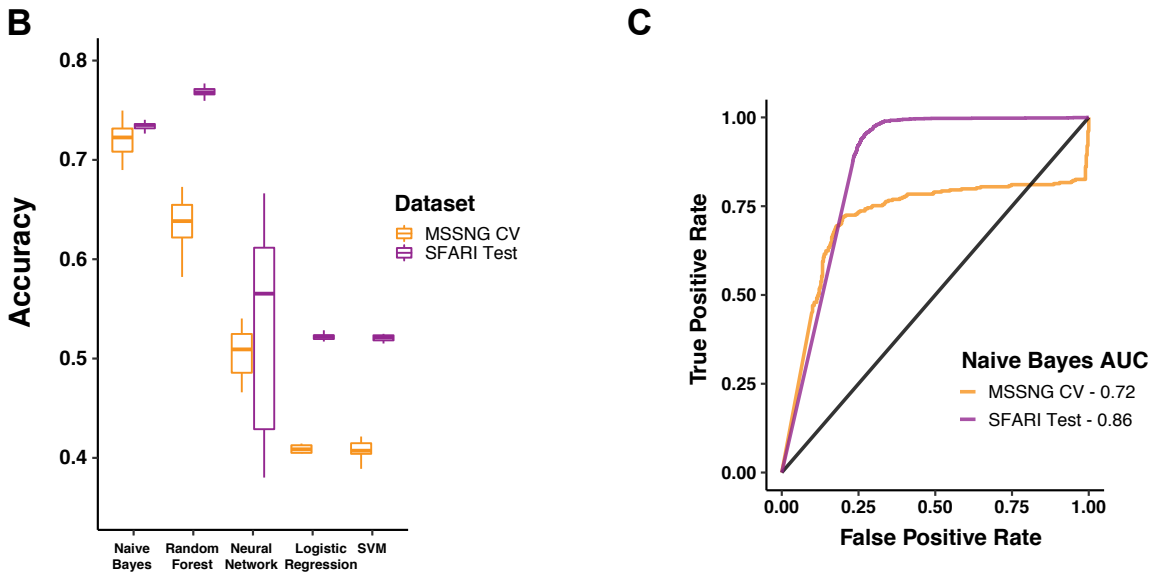


Figure 2.12: Generalizable classification of ASD vectors. A. Average model accuracy. Five different classification models were tested, but only the Naive Bayes model consistently performed well in both cross-validation and testing (CV: 72±1.8% and Test: 73±0.4%, mean±SD). B. The receiver operating characteristic curves of the Naive Bayes model shows balanced model sensitivity and specificity. The cross-validation and testing curve colors are specified in the legend, and the black curve represents a random classifier. Area under the receiving operating characteristic curve (AUC), a performance measure of binary classification, is also listed in the legend for each model.

2.4 Discussion

In this study, I report a novel genomic representation that successfully enables a machine learning-based analysis of ASD. The introduction of these tools for the exploration of complex

heritable diseases like ASD is timely. Sequencing databases now routinely possess thousands of subject observations. This abundance of data supports machine learning applications, which are designed for the automatic identification of salient patterns among feature-rich data, such as the variant burden vectors presented here. By considering integrated variant risk simultaneously across all genes, genome vectorization provides an alternative to GWAS or polygenic risk scoring, which are better suited for identifying highly penetrant variants in weakly polygenic contexts and limited in predictive power. Additionally, by including both coding and noncoding, rare and common, and *de novo* and inherited variation, I maximize the pool of information which classification models can leverage. Genome vectorization is a new solution to the problem of mapping genotype to phenotype for complex heritable disorders like ASD.

An omnigenic framework for modeling ASD implies that disease risk is a function of high impact-driver mutations working in tandem with common variants in virtually any gene [Boyle et al., 2017]. Due to the interconnectedness of biological networks, diverse molecular changes may converge on stereotyped cellular states and disease phenotypes [Geschwind and State, 2015]. By modeling genome-wide risk, our vectorization method in essence tests whether an omnigenic model can be fit to sequencing data acquired from ASD subjects. I derive our rationale for this approach from the ASD genomics literature, which has convincingly shown that disease risk is aggregated through a combination of rare gene disrupting mutations and common variation at heterogeneous loci [Geschwind and State, 2015, Weiner et al., 2017, Sebat et al., 2007, Iossifov et al., 2014]. Vectorization was performed on a large cohort of ASD and control genomes, and the resulting variant burden vectors were successfully used to train highly accurate classification models. Classification generalized across cross-validation folds, diverse model types, and covariate subgroups of the data, indicating a robust result that supports an omnigenic model for ASD.

An emphasis of this study was to probe the interpretability of the classification models to facilitate biological understanding. The ASD+ list, a set of salient classification gene features,

was automatically chosen based on the learned model weights. These ASD+ genes demonstrated a specific and statistically significant overlap with known ASD risk genes, brain-expressed genes, and neuronal biological pathways that recapitulate pertinent findings from standard RNA and DNA sequencing analyses in the literature [**?**, Parikshak et al., 2013, Willsey et al., 2013, Fischbach and Lord, 2010]. Given the genetic heterogeneity within the ASD population, variant burden vectors account for a spectrum of molecular changes, ranging from numerous, distributed low impact mutations to a smaller number of deleterious protein-coding events. Though risk may be distributed widely across the genome, our results suggest that the most salient ASD genes play important roles in neural development and synaptic function.

The hypothesis that the genetic programs implicated in ASD affect prenatal development in cortical regions and cell types is well-supported [Parikshak et al., 2013, Willsey et al., 2013, Courchesne and Pierce, 2005, Courchesne et al., 2011, Chow et al., 2012, Marchetto et al., 2016]. Using our novel machine learning approach, I present concordant evidence that salient genes for ASD classification are enriched in cortical development, particularly during midfetal time points at 13-18 pcw. At this stage of gestation, the generation of new neurons occurs at an exponential rate, representing a critical period of neuronal growth and circuit formation [Rabinowicz et al., 1996, Gohlke et al., 2007, Courchesne et al., 2018]. Therefore, disruptions in significantly enriched regions, such as the auditory and somatosensory cortices, at this critical juncture could initiate pathological programs with far-reaching effects later in pediatric development. The identified cortical regions have been previously tied to ASD neurobiology through both molecular and neuroimaging studies [Gaffrey et al., 2007, Gervais et al., 2004, Schultz et al., 2000, Khan et al., 2015, Tabuchi et al., 2007, Willsey et al., 2013]. Postmortem sequencing and gene knockout studies have correlated high-confidence ASD gene mutations to transcriptomic and synaptic disruptions in the somatosensory cortex [Willsey et al., 2013, Tabuchi et al., 2007]. Functional imaging studies have linked abnormal connectivity in the auditory and inferior temporal cortices to specific ASD-related cognitive deficits, such as deficits in voice and facial recognition, respectively

[Gervais et al., 2004, Schultz et al., 2000]. The ability to recapitulate these regions, which were identified using diverse modalities, is a validation of both our machine learning approach and existing findings in the field.

The ASD+ genes identified by the classification models suggest that both inhibitory GABAergic and excitatory glutamatergic neuronal cell types are impacted by variant burden differentially across the ASD population (Appendix Figure A.6 and Supplemental File 1, Table 8). Previous studies have implicated both of these cell types in ASD, supporting the theory that excitatory-inhibitory dysregulation is an important mechanistic driver in this disease [Willsey et al., 2013, Parikshak et al., 2013, Marchetto et al., 2016, Mariani et al., 2015, Chao et al., 2010]. The advantage of performing our analysis at a population level is that the distribution of excitatory and inhibitory neuronal risk can be analyzed, revealing variant burden heterogeneity that may explain differences in neural circuit and patient behavioral phenotypes.

Despite its considerable potential, genome vectorization can be improved and further explored. First, as a novel machine learning method, the success of this approach should be replicated in other ASD sequencing datasets, as well as datasets of related conditions. It is reassuring that the findings presented in this study replicate across two large, independent datasets and map nicely to existing knowledge in the field of ASD neurobiology. Confidence in this machine learning technique will grow as it is tested in the ever growing iterations of ASD sequencing databases and those of other heritable neurodevelopmental and psychiatric conditions, such as bipolar disorder and schizophrenia. Second, the ASD+ gene list is a ranking of salient model genes that may inform biological experimentation. By combining this genome-wide ranking with knowledge of tissue or cell specific contexts, our approach could be further augmented by targeting mutations in cellular models to disrupt highly ranked and molecularly relevant genetic pathways. Additionally, the creation of quadratic or polynomial vector feature spaces could reveal salient gene-gene interactions that are suggestive of nonlinear ASD risk and epistatic phenomena that may be tested *in vitro*. Finally, this study provides evidence that variant

burden vectors encode sufficient information for the accurate and generalizable classification of ASD cases from healthy controls, i.e. unaffected parents and siblings. The clinical implications of this result are enticing. However, expectations should be tempered till the demonstration of high performance in a controlled prospective analysis with unrelated subjects and a standardized medical genomics pipeline. Further validation is critical for the development of new clinical tools that utilize genome vectorization.

This initial study provides convincing preliminary evidence that our novel method facilitates molecular classification and risk gene identification in the context of ASD. By applying this tool to large sequencing databases, I was able to demonstrate that genome-wide variant burden is a useful vectorized quantity that encodes disease-specific risk and informs the analysis of genes, pathways, neurodevelopmental time points, brain regions, and cell types related to ASD. Using tools such as genome vectorization, I believe that the automated detection of molecular risk patterns has the potential to guide biological research and clinical management for complex heritable conditions moving forward.

Chapter 2, in full, is a reprint of the material as it has been written in a manuscript that has been submitted for publication. The authors of this study are Debha Amatya, Simon Schafer, Timothy Tadros, Saket Navlakha, Bhooma Thiruvahindrapuram, Stephen Scherer, Graham McVicker, and Fred Gage. The dissertation author was the primary investigator of this paper.

Chapter 3

Dynamical Analysis of Neuronal Electrical Activity in ASD

3.1 Background

3.1.1 Approach

In this chapter, I used minimum embedding dimension (MED) analysis to quantify dynamical complexity differences in the electrical activity of individuals with ASD, comorbid with early brain overgrowth, and neurotypical controls. Given the evidence of synaptic disruption in previous studies, I hypothesized that ASD samples would exhibit reduced dynamical complexity during neuronal differentiation [Liu et al., 2017, Marchetto et al., 2016, Mariani et al., 2015]. I further explored how MED captured the spectrum of healthy and abnormal states present in this cohort through gene expression analysis. Unlike case-control labels, I expected that MED would reveal more robust patterns of differential gene expression, because it was cellularly derived and encoded both individual and ASD-related features. Through the application of dynamical analysis techniques to patient-derived neuronal recordings, I suggest a novel paradigm for investigation of the disease-related signatures in ASD.

3.1.2 Dynamical complexity

Electrical activity arising from cultured neurons can be modeled as a dynamical system. Doing so is advantageous, because these methods excel at capturing the inherent nonlinearity of such data, which allow for new measures of neuronal function to be gleaned. The dynamical complexity of an electrical recording is one measure that can be quantified in the MED variable.

The false nearest neighbors (FNN) method was proposed by Kennel et al. (1992) to find the MED for time-series dynamic systems [Kennel et al., 1992]. In this framework, a time-series arises from a dynamical system composed of n differential equations. All possible states of the system can be geometrically represented by an attractor, which is a regular shape formed when the n state variables (e.g. x, y) are plotted against one another. The governing equations and state space trajectory of the Hénon map, a simple two variable system are presented in Figure 3.1.

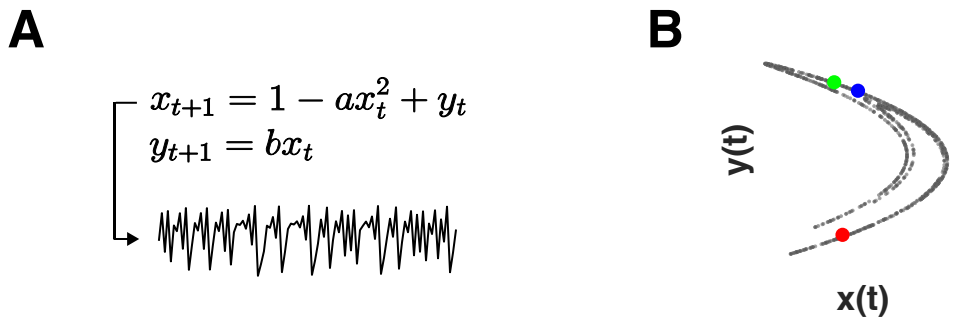


Figure 3.1: Using the Hénon map to understand MED. The goal of the MED technique is to estimate the dimensional complexity of the system using only a time-series recording from a single variable in the system. This can be accomplished by using an algorithm to minimize the number of false nearest neighbors (FNNs) identified by iteratively plotting the time series over embedded versions of itself. A. The Hénon map is a 2-variable (x and y) nonlinear dynamical system that exhibits chaotic behavior in time, simulated to recorded electrical activity. B. A state space reconstruction of this system reveals a smooth and regular attractor (right), in which some (x, y) coordinate pairs are closer together in space (blue and green) relative to other pairs (red). This attractor is a geometric representation of the 2-variable system, which can be summarized as having a dimensionality, or complexity, of 2, because it is fully unfolded in the (x, y) coordinate space.

When only a single variable is accessible, I can instead use d delayed embeddings of this variable to reconstruct the attractor. For example, for $y(n)$ from time-series $x(t)$:

$$y(n) = (x(n), x(n+T), ..., x(n+(d-1)T)) \tag{3.1}$$

where T is time delay and the trajectory evolves by $y(n), y(n+1)$. If the embedding dimension d is lower than the native embedding dimension, the system is not fully unfolded and the reconstruction fails to represent the dynamics of the system, which leads to false nearest neighbors (FNN) in the geometric space, as shown in Figure 3.2.

When iteratively increasing the embedding dimension from d to $d+1$, the FNN percentage decreases during the unfolding process until the MED is reached and the state space is fully represented. Therefore, plotting the FNN percentage against the embedding dimension, reveals the MED, a quantitative indicator of the intrinsic complexity of the system. Figure 3.2 visually

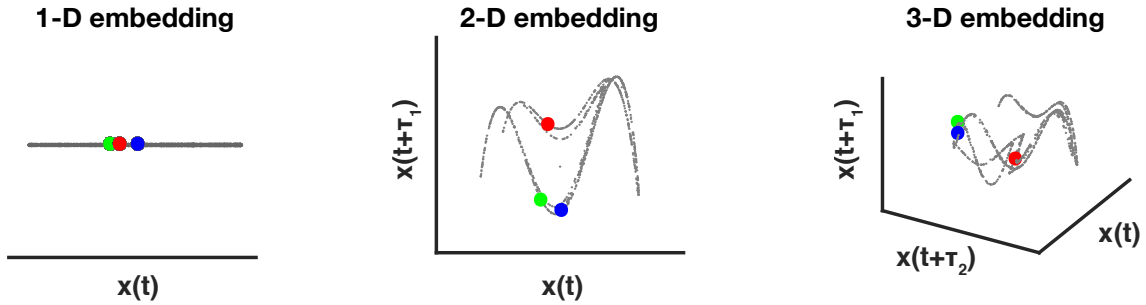


Figure 3.2: The false nearest neighbors algorithm. Demonstration of the false NN procedure to find MED. The recording of a single variable from a dynamical system, such as x in the Hénon system or electrical activity in a system of neurons, can be embedded with delayed versions of the time series to reconstruct the original attractor of the entire system. This allows for the iterative testing of embedding dimensions and measurement of conserved neighboring points until a minimum is reached. In the example of the Hénon system, 1-dimensional embedding clusters all 3 points together. A 2-dimensional embedding appropriately separates the red point from the blue and green points. Finally, a 3-dimensional embedding fails to create a further separation of these points, signifying that sufficient complexity is captured by a 2-dimensional embedding for the Hénon system. In the case of MEA recordings, MED provides a measurement of dynamical complexity of neural activity, potentially providing new insights into the neurobiology of development and ASD.

demonstrates these concepts in an intuitive manner.

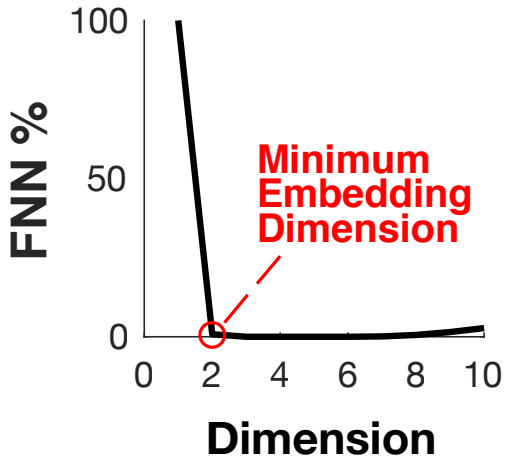


Figure 3.3: The MED recapitulates the dimensionality of a dynamical system. The false nearest neighbors (FNN) algorithm can be used to find the minimum embedding dimension (MED). Percentage of false NN is minimized when a time series is embedded in its native dimensionality. For the Hénon system, the MED occurs at 2, because it is a 2-variable system.

3.2 Methodology

3.2.1 iPSC samples and neuronal differentiation

iPSC lines were derived from two sources. First, 13 samples from Marchetto et al. (2017) comprised the bulk of the lines [Marchetto et al., 2016]. These included 8 lines derived from ASD patients with an early brain overgrowth phenotype and 5 age-matched neurotypical control lines. Second, 2 additional control lines from neurotypical adults were added to balance the groups. The samples from the first cohort included standardized clinical and functional assessments, including the Autism Diagnostic Observation Schedule, Wechsler Intelligence Scale, and Vineland Adaptive Behavior Scale. Sample metadata and clinical scores are given in Supplemental File 2, Table 1. Samples were obtained with approval by the internal review board of the Salk Institute for Biological Studies and informed consent of all subjects. Neuronal differentiation was accomplished through differentiation of the iPSCs into neural progenitor cells, followed by the removal of FGF2 to drive maturation into neurons, as described in the literature [Marchetto et al., 2016].

3.2.2 MEA recordings and spike analysis

Ninety-six-well MEA plates from Axion Biosystems (San Francisco, CA, USA) were used to record electrical activity of neurons derived from all 15 subjects. Each subject's cells were plated in replicates of 6 and seeded with 10,000 neural progenitor cells (NPCs) that were induced into neuronal differentiation the next day. Wells were coated with poly-ornithine and laminin before cell seeding. Cells were fed every other day and measurements were taken twice a week before feeding. The Maestro MEA system and AxIS software (Axion Biosystems) were used to record neuronal electrical activity from the plates. Voltages were recorded at a frequency of 12.5 kHz and bandpass filtered between 10 Hz and 2.5 kHz. Spike detection was performed using an adaptive threshold set to 5.5 standard deviations above the mean activity of each electrode.

Following 5 minutes of plate rest time, recordings were performed for 10 minutes. A total of 15 separate recordings were performed over a timespan of 47 days, for each subject.

Multielectrode data analysis was performed using the Axion Biosystems Neural Metrics Tool, which calculated standard spike-related variables. Bursts were detected with an adaptive Poisson algorithm for high spiking activity that occurred on a single electrode. Variables were averaged across subject replicates and plotted by group for each day of the recordings. The 95% confidence interval around the mean was also calculated and plotted for each day to aid in the discovery of significant trends.

3.2.3 False nearest neighbors and embedding dimension analysis

Two criteria were used to define the FNNs, as described by Kennel et al. (1992) [Kennel et al., 1992]. The first criterion is the tolerance threshold parameter, R_{tol}, which measures how the neighbor distances change relative to previous distances when increasing d to $d+1$. If the neighbors are false, adding another dimension would largely increase the neighbor distances during the unfolding process. The second criterion models noise in the data. I introduced a threshold parameter A_{tol} to compare the neighbor distances in dimension d+1 with the size of the attractors in dimension d. By combining these two criteria, the MED for time-series data contaminated by noise can be effectively found.

I set $T = 2$, and I used the same threshold parameters as in Kennel's original study $R_{tol} = 15$ and $A_{tol} = 2$ [Kennel et al., 1992]. FNN was implemented in MATLAB R2017b (Natick, MA, USA) and performed on both real MEA data and simulated data from the Hénon map [Hénon, 1976, MATLAB, 2017].

3.2.4 RNA sequencing

At 15 days post-differentiation, the neurons positive for PSA-NCAM (Anti PSA-NCAM Antibody, Miltenyi Biotec) were sorted using flow cytometry to isolate neuronal fate-committed cells. RNA was extracted after cell sorting on TRIzol LS reagent (Invitrogene) from all 15 neuronal samples. Total RNA was extracted using DNA Free RNA Kit (Zymo Research) according to the manufacturer's instructions. RNA quality was assayed using Agilent Technologies 2200 TapeStation and samples with integrity superior to RIN 8.5 were used for library preparation. Stranded mRNA sequencing libraries were prepared using the Illumina TruSeq Stranded mRNA Library Prep Kit according to the manufacturer's instructions. RNA with a poly A tail was isolated using magnetic beads conjugated to polyT oligos. mRNA was then fragmented and reverse transcribed into cDNA. dUTPs were incorporated, followed by second strand cDNA synthesis. The dUTP-incorporated second strand was not amplified. cDNA was then end repaired, index adapter ligated and PCR amplified. AMPure XP beads (Beckman Coulter) were used to purify nucleic acid after each step of the library prep. All sequencing libraries were then quantified, pooled and sequenced at single-end 50 base-pair (bp) on Illumina HiSeq 2500 at the Salk Institute Next Generation Sequencing Core.

1000ng of RNA was used for library preparation with the Illumina TruSeq RNA Sample Preparation Kit. The RNAs were sequenced on Illumina HiSeq2000 with 50 bp paired-end reads, generating 50 million high-quality sequencing fragments per sample, on average. Sequenced reads were quality-tested using FASTQC [Andrews, 2010] and aligned to the hg19 [Lander et al., 2001] human genome using the STAR aligner [Dobin et al., 2013] version 2.4.0k. Mapping was carried out using default parameters (up to 10 mismatches per read, and up to 9 multi-mapping locations per read). The genome index was constructed using the gene annotation supplied with the hg19 Illumina iGenomes collection [Illumina, 2015] and sjdbOverhang value of 100. Raw gene expression was quantified across all gene exons (RNA-Seq) using the top-expressed isoform as proxy for gene expression. Transcripts per million (TPM) values were calculated for each

sample:

$$TPM_i = \frac{X_i}{l_i} \times \frac{1}{\Sigma \frac{X_j}{l_j}} \times 10^6 \tag{3.2}$$

where X_i was the count estimate for gene i and l_i was the length of the transcript as determined by querying the Ensembl mart with dataset set as "hsapiens_gene_ensembl", using biomaRt in R version 3.5.1. TPM values were $\log_2(TPM_i + 1)$ transformed for downstream analyses.

Two subject outliers were detected by visual inspection of a principal component analysis plot of the normalized gene counts matrix (Appendix Figure B.3). Thirteen samples were retained for further analysis, including 7 ASD lines and 6 control lines.

3.2.5 Differential expression analysis

Gene-based read counts were analyzed for differential expression using the R DeSeq2 package, which uses variance stabilization techniques and the negative binomial distribution to detect expression changes across experimental conditions (Boston, MA, USA) [Love et al., 2014]. Here, two conditions were analyzed for differential expression, using two different model matrices. First, differential expression with respect to MED was analyzed:

$$X_i \propto MED \tag{3.3}$$

Second, autism-associated genes were gleaned by testing a model that included autism and autism-MED interactions:

$$X_i \propto MED + MED \times ASD + ASD \tag{3.4}$$

A total of 24,162 genes was examined for differential expression, and genes were deemed significant if $p_{FDR} \leq 0.05$. An expanded set of genes with $p_{raw} \leq 0.05$ were used to asses broad

trends in the data with follow up analyses Plotting of individual gene expression results, top gene summaries, and heatmaps were performed using transcript per million normalized counts that were corrected for covariates with the R ComBat package (Boston, MA, USA) [Leek and Storey, 2007]. All gene expression data are deposited at Gene Expression Omnibus (GEO accession: GSE125020).

3.2.6 Gene ontology analysis

Gene ontology (GO) analysis was performed to determine which biological pathways were associated with differentially expressed genes. The statistical overrepresentation tool from PantherDB was used to perform Fisher's exact tests for a list of uncorrected differentially expressed genes and each GO term in the database [Mi et al., 2016, Thomas et al., 2003]. P-values were adjusted for multiple comparisons using false discovery rate correction. Biological processes with a $p_{FDR} \leq 0.05$ were reported and plotted.

3.2.7 Statistical tests for gene list overlap and clinical correlations

Statistical analyses for gene list overlap and clinical correlations were performed using R version 3.51 (Vienna, Austria) [Team et al., 2013]. Clinical correlation was examined between MED day 15 average values and subject clinical scores. To further explore the differential expression results, overlap was measured with ASD genes, brain expressed genes, and a negative control of liver-expressed genes [Abrahams et al., 2013, Krishnan et al., 2016, Uhlén et al., 2015]. Tissue-specific gene sets were obtained from the Human Protein Atlas (available from www.proteinatlas.org). For each list of interest, overlap was tested using the binomial test, and significance was assigned to $p \leq 0.05$. Supplemental File 2, Table 4 further describes the gene lists used in this study. Clinical correlations between MED and various clinical endpoints were calculated using the Pearson correlation test. Presented correlations are significant at $p \leq 0.05$

level.

3.2.8 Spatiotemporal enrichment analysis

Spatiotemporal enrichment of the MED differentially expressed genes was assessed using gene expression data from the BrainSpan Atlas of the Developing Human Brain [Sunkin et al., 2012]. Normalized gene transcript counts were acquired for region-stage pairs. Twelve developmental stages and 16 discrete brain structures were included. In a process similar to one described by Krishnan et al. (2016), representative genes sets were chosen for each region-stage pair by calculating the modified z-score of a given gene in the distribution of counts for all region-stage pairs [Krishnan et al., 2016]. The modified z-score was calculated using the median, and the median absolute deviation (MAD) of each gene across all pairs was calculated for the *i*th gene and *j*th region-stage pair:

$$z_{i,j} = \frac{0.645 \times (count_{i,j} - median_i)}{MAD_i} \tag{3.5}$$

For the *j*th region-stage, all genes for which $z_{i,j} \geq 2$ were selected as representative genes. For each region-stage, the representative genes are listed in the Supplemental File 2, Table 7.

Enrichment was calculated by performing the Fisher's exact test using the list of differentially expressed MED genes and each representative region-stage set of genes. Enrichment scores were corrected for multiple comparisons using false discovery rate controlling, and pFDR values are plotted in the heatmap in Figure 3.14. Significance was assigned to region-stage pairs with pFDR less than 0.05. The table detailing all spatiotemporal enrichments is listed in the Supplemental File 2, Table 8.

3.3 Results

3.3.1 Study design and MED analysis technique

This study was designed to identify a new dynamical electrophysiological endophenotype associated with idiopathic ASD, as well as the gene expression correlates of this signal. To accomplish this, iPSC lines derived from 8 ASD patients with early brain overgrowth and 7 neurotypical controls were obtained and differentiated into neurons using a pan-cortical protocol. A schematic of the study design is depicted in Figure 3.4, and a description of the samples is given in Table 3.1.

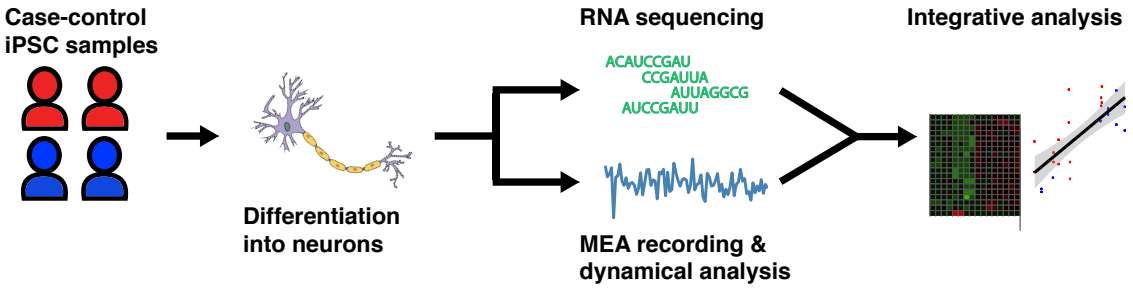


Figure 3.4: Study design for dynamical analysis of neuronal lines. iPSCs from 7 neurotypical controls and 8 ASD cases were differentiated into neurons. Neurons were studied using RNA sequencing and MEA recordings to identify disease signatures that integrate across both levels of analysis.

Additional clinical metadata for these samples is presented in Supplemental File 2, Table 1. From the neural progenitor phase, longitudinal electrical recordings were gathered with MEA over the course of 47 days. For each recording, standard spike related variables, as well as MED, a measure of dynamical complexity, were computed. On day 15 post-differentiation, the developing cells were sorted for neuronal identity using the PSA-NCAM marker and then RNA sequenced. The subsequent analysis focused on linking the observed dynamical endophenotype with gene expression changes that relate to ASD biology. This was done by using the average day 15 MED values for each subject as a variable in the gene expression models fit to the RNAseq data.

Figure 3.5 depicts how dynamical complexity may vary across samples, even when

Table 3.1: Samples for neuronal analysis. A description of subjects, from whom iPSC cell lines were generated and subsequently differentiated into neurons for analysis. Abbreviations: IQ (Wechsler Intelligence Quotient), ADOS (Autism Diagnostic Observation Schedule), Vineland ABC (Vineland Adaptive Behavioral Composite, second edition). Age, IQ, ADOS, and Vineland ABC scores were recorded at the time of biopsy. Brain volume was computed via MRI during early childhood. All samples were derived from males. ASD cases met the diagnostic criteria as defined by ADOS total score cutoffs or DSM-IV guidelines. Where relevant, quantities are shown as averages one standard deviation.

	iPSC Lines (N=15)	
	Control	**ASD**
Samples	7	8
Brain Volume (cm^3)	$1,237.2 \pm 86.6$	$1,372.9 \pm 87.8$
Age (years)	25.0 ± 27.7	13.3 ± 5.6
IQ	118.6 ± 8.6	67.9 ± 14.9
ADOS	–	16.75 ± 2.8
Vineland ABC	99 ± 6.2	57.13 ± 15.1

standard measures like firing rate remain constant. In such cases, nonlinear dynamical differences may uncover new patterns of disruption in gene expression. Further intuition for MED analysis is provided by a simulated example of the Hénon map, a two-dimensional dynamical system, in Figure 3.2.

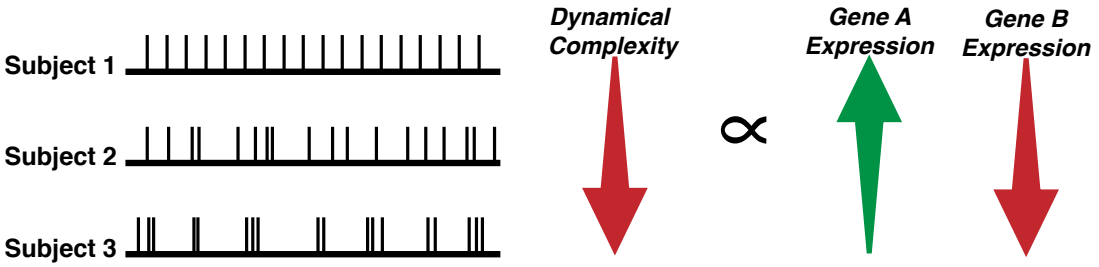


Figure 3.5: Visualizing dynamical complexity. Dynamical complexity may vary even when standard spiking variables are constant. For these pedagogical examples, the firing rate is constant at 20 spiking events in the interval. Nevertheless, differences in the organization of the spiking events may result in large differences in complexity, as measured by dynamical analysis. These differences may distinguish groups of interest, such as ASD, and correlate to gene expression changes related to neuronal dysfunction.

3.3.2 Neuronal electrical recordings and dynamical analysis

To glean new knowledge from this in vitro model of ASD, both standard analytical tools and nonlinear dynamical methods were applied to the electrical recordings of the neurons. This approach is novel but also appropriate, as neuronal activity is a nonlinear dynamical system that arises from the firing of nascent neuronal networks. Representative electrical recordings and spiking events are shown in Figure 3.6.

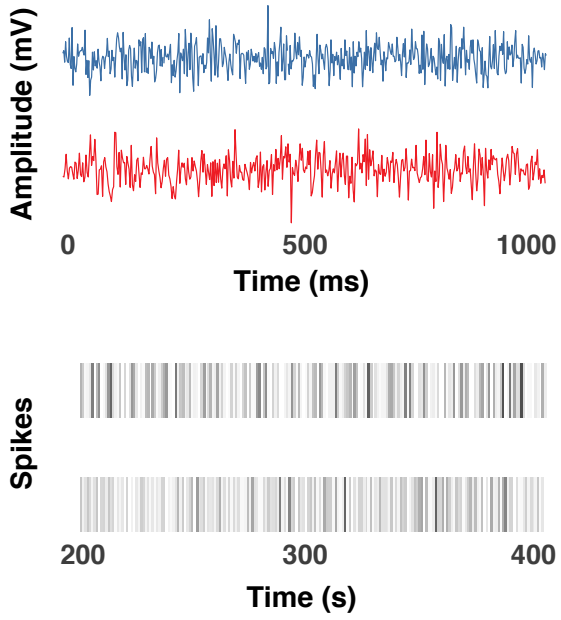


Figure 3.6: Electrical recordings from neuronal lines. Raw electrode waveforms (top) are used for spike detection (bottom) and downstream analyses.

Traditional metrics, such as the mean firing rate, fail to show statistical separation between the ASD and control group (Figure 3.7). However, more complex measures, such as bursting and the variation of interspike intervals (ISI) within bursts, are able to show a difference between groups, suggesting that dynamical analysis may also capture features related to bursting that fundamentally distinguish ASD and control electrical activity (Figure 3.7, 3.8, and Appendix Figure B.1).

MED was computed to quantify the dynamical complexity of the electrical recordings.

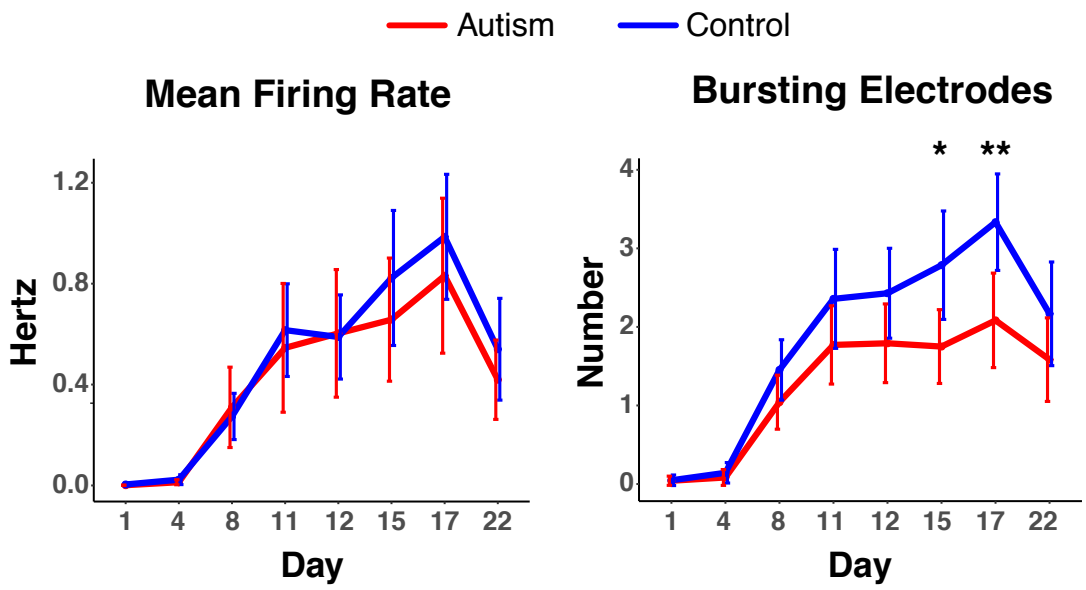


Figure 3.7: Bursting variables highlight electrophysiological differences. Group average values with 95% confidence intervals are given over several timepoints for two variables. Mean firing rate fails to distinguish control (blue) and ASD (red) groups over any day of the recording period. However, number of bursting electrodes does show significant group-based differences for day 15 and 17. For the relevant figures in this chapter, significance was as tested with Welch's two-sided t-test and indicated by asterisks. * corresponds to $p \leq 5 \times 10^{-2}$, ** corresponds to $p \leq 5 \times 10^{-3}$, *** corresponds to $p \leq 5 \times 10^{-5}$, and **** corresponds to $p \leq 5 \times 10^{-7}$.

A previous study provided evidence that iPSC-derived ASD neurons were characterized by diminished synaptogenesis, thus compromising the process of forming connected networks [Marchetto et al., 2016]. It was expected that MED would capture this ASD-related deficit of electrophysiological complexity, thereby uncovering a novel endophenotype for this disorder.

A difference between ASD and control MED was found to emerge after 11 days of differentiation and persist with statistical significance through 1.5 more weeks of recording, as shown in Figure 3.9. Correlation analysis between MED and bursting variables revealed a significant association with the number of bursting electrodes and the ISI coefficient of variation, strengthening the relationship between dynamical complexity and coordinated spiking activity and variation. Despite this relationship, the variance associated with MED was less than that of bursting variables, suggesting improved precision in the characterization of electrical activity.

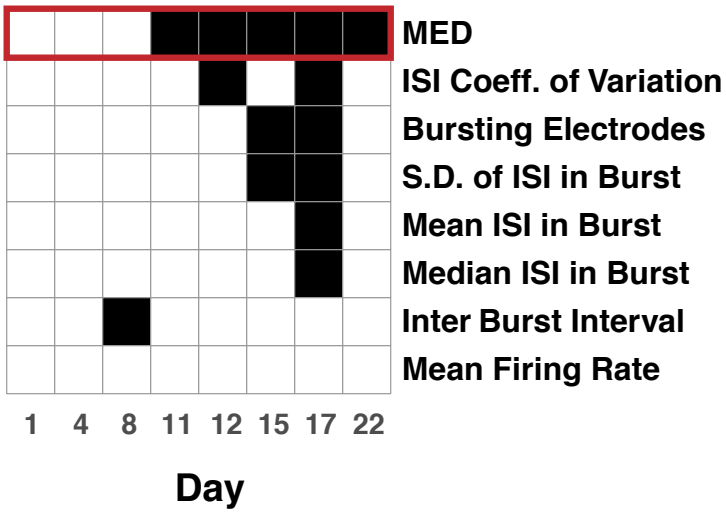


Figure 3.8: Overview of group-wise differences across measures. This binary matrix indicates at which timepoints a measure was able to detect a significant difference between cases and controls (black squares). The MED outperforms all measures in distinguishing groups, but it does overlap with some bursting variables, as further explored in Appendix Figure B.1.

Appendix Figure B.2 details average subject-wise MED boxplots across the key timepoints of the study, including the day when RNA sequencing was performed.

To assess the behavioral relevance of MED, the correlation between subject MED values and cognitive scores was examined. This revealed a significant correlation in two relevant endpoints, the Vineland ABC Score and Nonverbal IQ, as shown in Figure 3.10. This suggests that cellular dynamic state may meaningfully impact cognitive and behavioral states relevant to ASD. Therefore, MED is a novel cellular electrophysiological endophenotype that is related to both ASD status and some of its relevant clinical features.

3.3.3 Differential expression models for ASD and MED

RNA sequencing was performed at 15 days post-differentiation in order to explore the gene expression correlates of MED and ASD. Given that MED captures both individual and ASD-related electrical dynamics, I expected that it would be associated with robust and biologically relevant transcriptomic alterations (Appendix Figure B.2). To test this rationale, gene expression

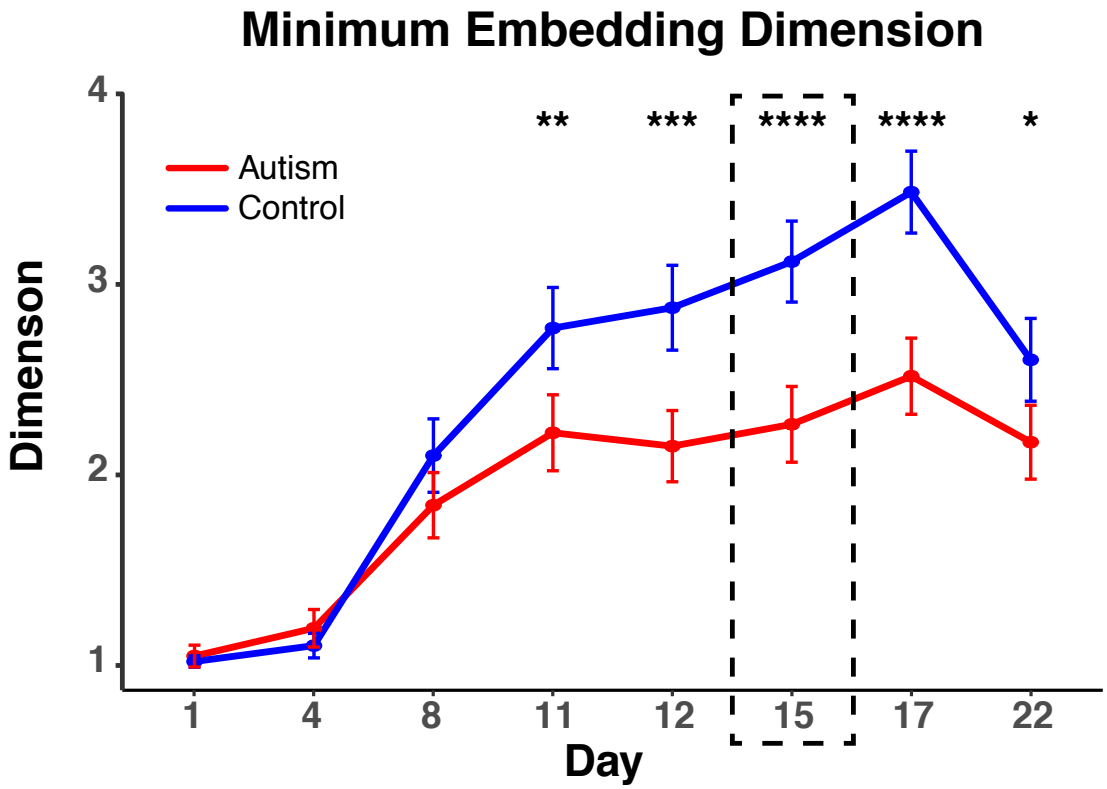


Figure 3.9: MED offers a sustained and statistically significant separation of groups. Average MED values and 95% confidence intervals for both groups are depicted for the first eight timepoints. The controls subjects (blue) are associated with higher MED complexity score, in comparison to the ASD subjects (red). The dotted rectangle indicates values of the MED during day 15 of the MEA recordings, when RNA sequencing was performed.

was examined with respect to 1) MED and 2) ASD interacting with MED in separate models. Two subject outliers were excluded based on principal component analysis (Appendix Figure B.3). As shown in Figure 3.11, before FDR correction, the MED model identified 1,423 differentially expressed genes, and the ASD interaction model was associated with 761 genes. After controlling for multiple comparisons, MED was associated with 53 genes in comparison to 7 for ASD within each respective model.

The comprehensive lists of differentially expressed genes are given in Supplemental File 2, Tables 2 and 3, and examples of genes differentially expressed with respect to MED value are shown in Appendix Figure B.4. Of note, a model that only included the ASD status without MED

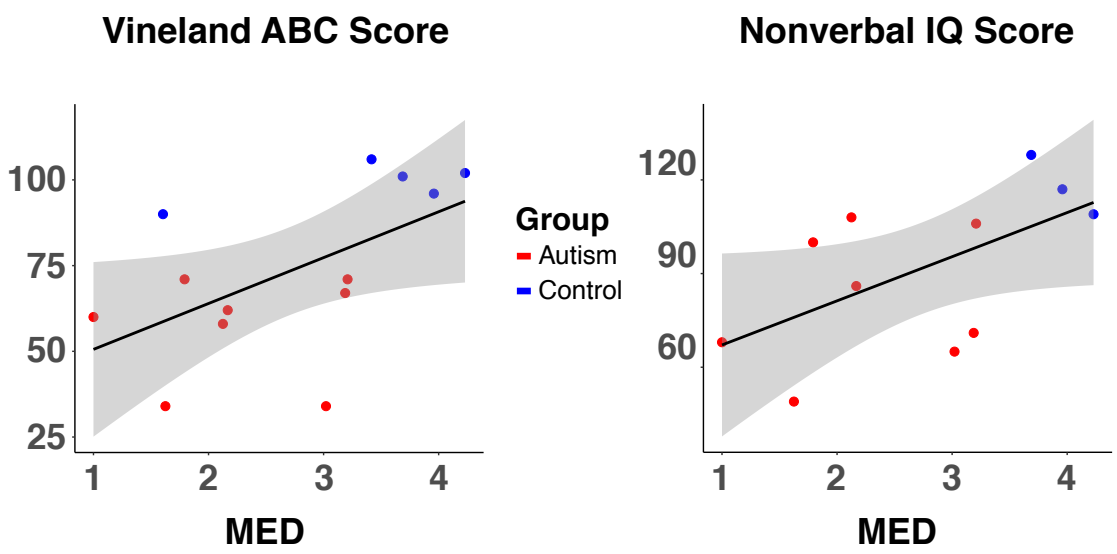


Figure 3.10: MED is correlated to clinical endpoints of interest. Low MED scores, which are associated with ASD diagnosis, are correlated to lower Vineland adaptive behavior score (left) and lower early nonverbal IQ (right). The grey shading represents a 95% confidence interval around the fitted curve, as estimated by a linear model.

was unable to find any differentially expressed genes after FDR correction. This result suggests that cell-derived endophenotypes could represent a path forward in studying complex genetic disorders, such as idiopathic ASD. Binary ASD-control labels do not account for the tremendous intersubject heterogeneity present in such samples, and cellular markers may better model both individual and group-related variance for gene expression analyses.

3.3.4 Interrogation of the MED signature

The gene expression signature of the MED was examined in relation to known ASD genes and pathways, as well as brain regions and stages of development. For these comparisons, the full set of differentially expressed MED and ASD genes, prior to FDR correction, was considered. Figure 3.13A shows that both ASD and MED genes are overrepresented in brain-expressed genes and putative ASD genes. However, the enrichment effect sizes were markedly larger for MED genes. This finding provides evidence that MED is related to relevant physiological and disease-relevant gene lists. Next, biological pathways were analyzed through gene ontology analysis,

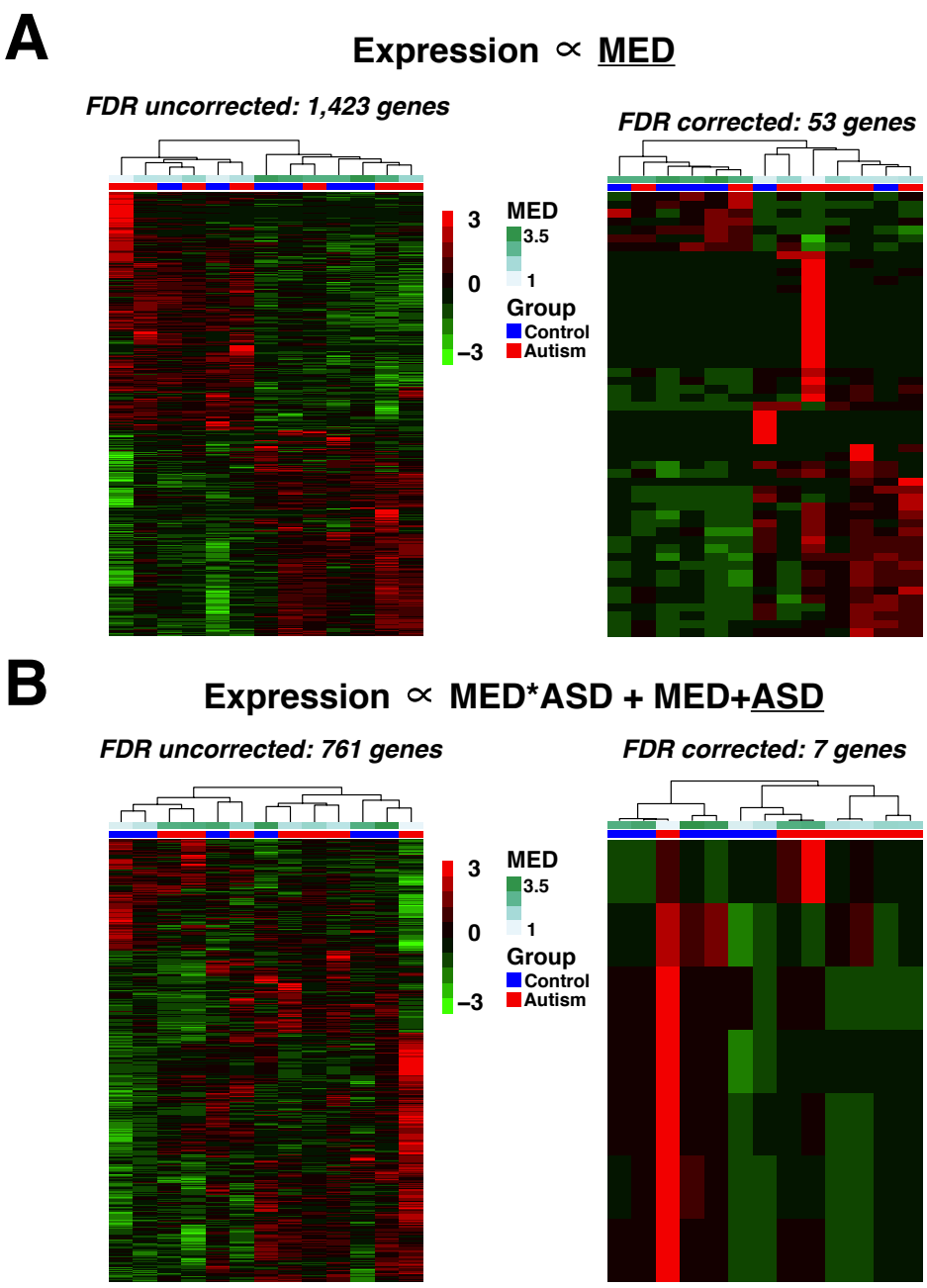


Figure 3.11: Gene expression signatures of MED and ASD. A. MED is associated with differential expression in 1,423 genes before multiple comparison correction and 53 genes after correction, as depicted in the heatmaps. Hierarchical clustering of the samples based on gene expression separates low and high complexity samples, as indicated by the MED color bar. B. Examination of ASD genes in a combined MED and ASD interaction model identifies 761 differentially expressed genes before correction and 7 differentially expressed genes after multiple comparison correction.

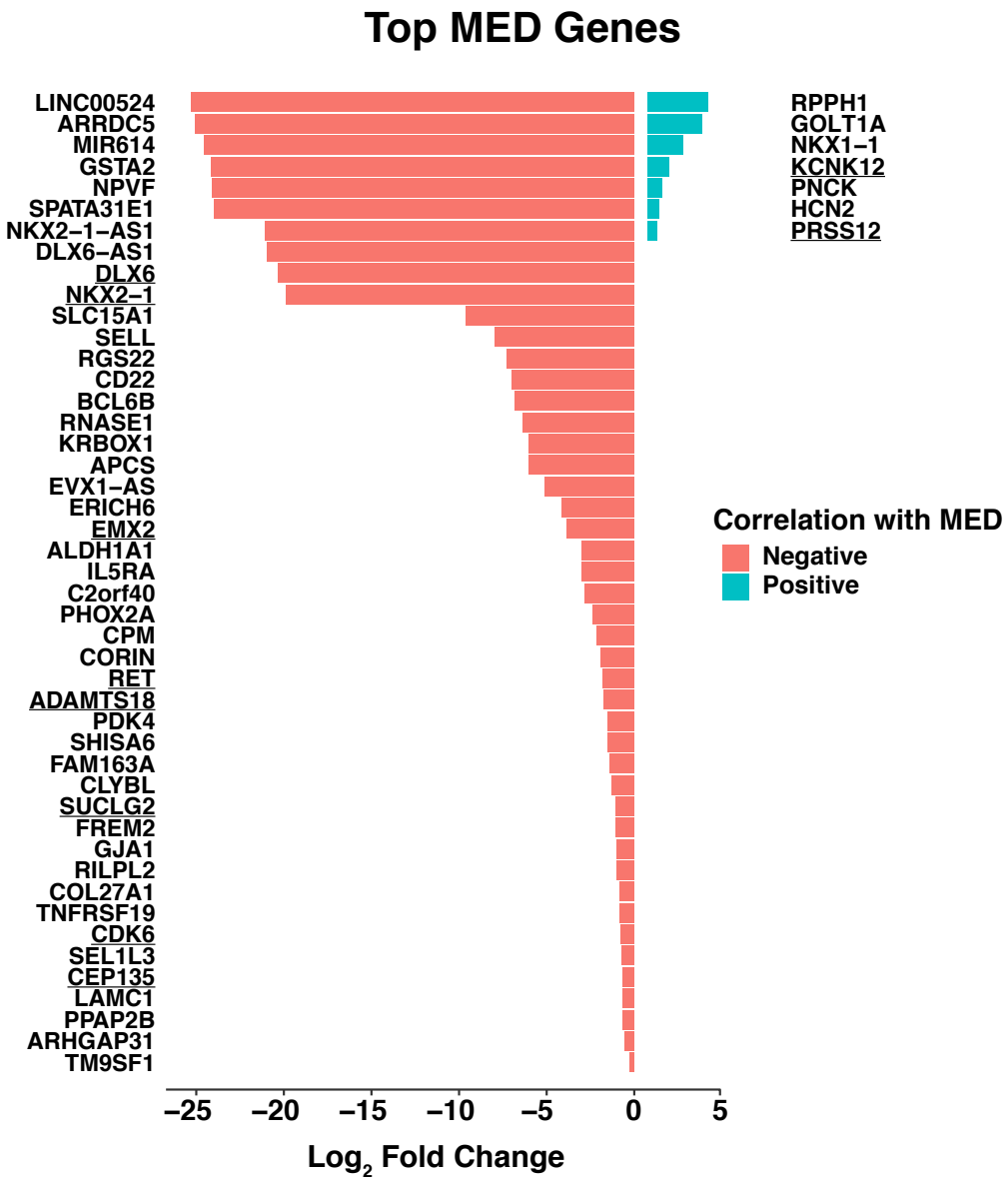


Figure 3.12: Differentially expressed MED genes. The $\log_2$ fold changes of the 53 differentially expressed genes after multiple comparison correction for MED are shown. The red bars correspond to genes that are negatively correlated to MED, and the blue bars correspond to genes that are positively correlated to MED. Bolded genes are also implicated in ASD risk or brain function in previous studies. Most differentially expressed genes are negatively correlated to MED, and positively correlated genes have less variability in their fold changes.

as shown in Figure 3.13B. MED genes were enriched in neurodevelopmental, cell migration, junction assembly, and regulatory terms. Additional exploration of cellular components related

to the MED gene signature showed a significant overlap with neuronal components, such as the synapse, postsynaptic density, axon, and dendritic spines. Full results for gene ontology analysis testing of biological processes and cellular components are given in Supplemental File 2, Tables 5 and 6, respectively. These findings suggest that MED encodes a measure of neuronal maturation and participation in circuit formation.

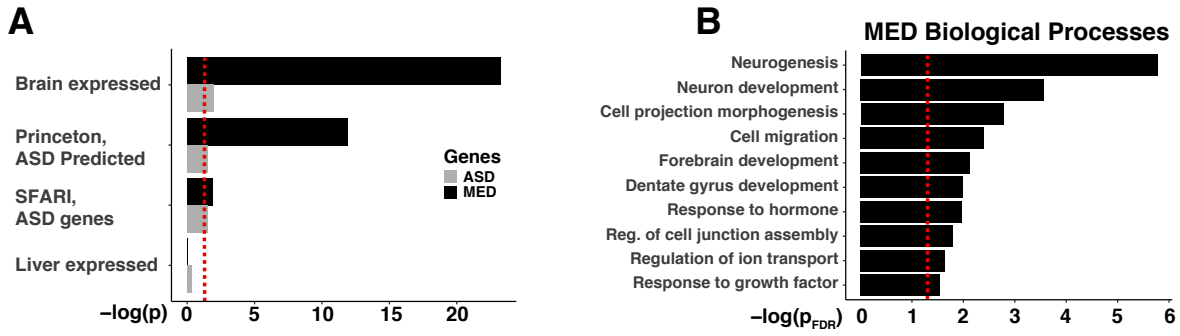


Figure 3.13: Enrichment of MED gene expression signature with ASD-related genes and biological pathways. To interpret the broad biological relevance of MED and ASD expression signatures, differentially expressed genes prior FDR correction were further analyzed. A. The MEA (black) and ASD (gray) differentially expressed genes were tested for enrichment in 5 lists: 1) highly brain-expressed genes, 2) putative ASD risk genes, 3) gold standard ASD risk genes, and 4) highly liver-expressed genes (negative control). Significance was testing using the binomial test with a cutoff of $p = 0.05$ (red line). Both ASD and MED genes were enriched in brain, Princeton, and SFARI lists, though the effect size of the MED signature was notably stronger. Neither set overlapped with liver-expressed genes, a negative control. B. Biological processes implicated with MED. The gene expression signature of MED impacts neurodevelopmental, cell migration, and growth-related ontologies. Significance of enrichment was calculated using the Fisher's exact test with false discovery rate control and a cutoff of $p_{FDR} = 0.05$ (red line).

Finally, to localize the set of MED genes to developmental time points and brain regions, data from the BrainSpan Atlas were used to identify enrichments in specific region-stage pairs. The enrichments were visualized as a matrix in Figure 3.14A for all region-stage pairs, and regional involvement during the highly enriched, late prenatal timepoint was mapped onto a representation of the brain in Figure 3.14B. Supplemental File 2, Table 8 lists the exact p-values obtained for the spatiotemporal analysis. This analysis uncovered a strong prenatal temporal enrichment in a mixture of cortical and deep structures (e.g., the visual cortex, inferior temporal

cortex, and thalamus), as well as a secondary involvement of late childhood cortical regions, suggesting that MED-related ASD pathology might play out at various stages of development with diverse structural involvement.

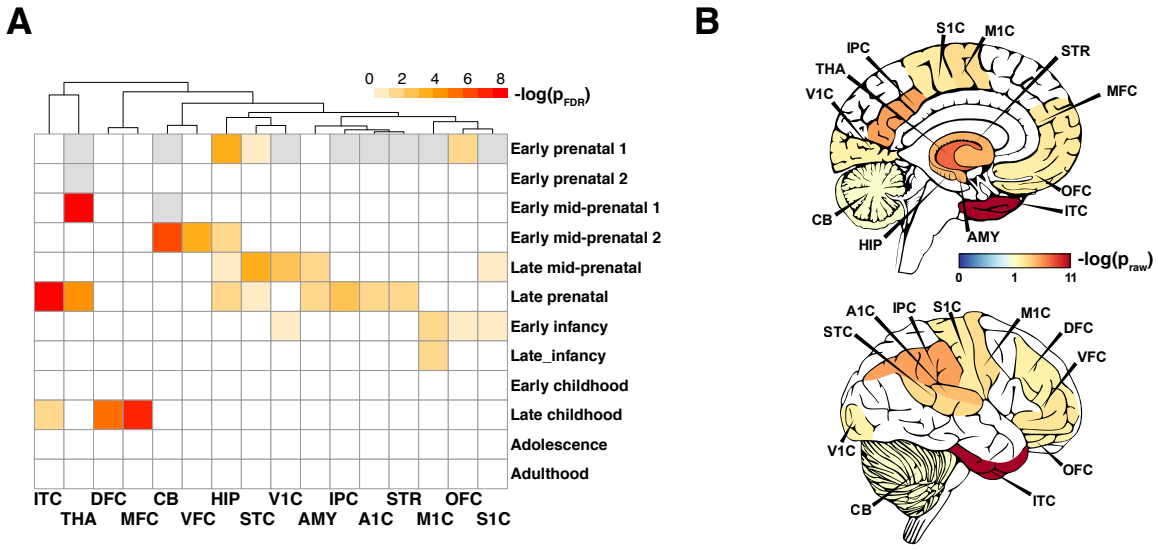


Figure 3.14: Spatiotemporal analysis of MED gene expression signature. A. Gene lists from 16 neuroanatomical regions and 13 developmental stages were tested for enrichment with the MED genes. The neuroanatomical regions include the inferior temporal cortex (ITC), thalamus (THA), dorsal frontal cortex (DFC), medial frontal cortex (MFC), cerebellar cortex (CB), ventral frontal cortex (VFC), hippocampus (HIP), superior temporal cortex (STC), primary visual cortex (V1C), amygdala (AMY), inferior parietal cortex (IPC), primary auditory cortex (A1C), striatum (STR), primary motor cortex (M1C), olfactory cortex (OFC), and primary somatosensory cortex (S1C). Fisher's exact test was performed for each region-stage pair and plotted on a heatmap after false discovery rate correction. A variety of cortical and deeper structures are implicated in MED, primarily at prenatal timepoints. A strong enrichment of the DFC and MFC are also indicated during late childhood. The grayed-out regions represent structures that are not present in the early prenatal brain. B. Visualization of cortical and interior MED-associated regions during late prenatal development (25-28 pcw). Raw enrichment p-values were plotted to show the range of structural involvement at this timepoint. Cortical regions such as the A1C, S1C, and ITC are enriched for the MED signature; however, deeper lying structures such as the thalamus and striatum are also revealed.

3.4 Discussion

In this chapter, I propose a new endophenotype for idiopathic ASD, called MED, which is derived from the dynamical complexity of neuronal electrical recordings. Previous in vitro studies examining ASD neurons have demonstrated both single unit recording abnormalities and alterations in spiking and bursting properties using MEA [Liu et al., 2017, Marchetto et al., 2016, Mariani et al., 2015]. Patient EEG studies have shown that recordings taken from people diagnosed with ASD exhibit reduced dynamical complexity compared to controls [Bosl et al., 2011]. For the first time, I demonstrate ASD-related nonlinear dynamical electrophysiologic alterations using iPSC derived neurons. Traditional spiking variables, such as firing rate, failed to capture statistically significant differences between the ASD and control group at day 15-21 days post-differentiation. Other electrophysiological variables related to bursting did successfully separate the case and control groups, but with less precision than MED. This finding suggests that dynamical complexity may be related to the integrated and synchronized firing of neuronal networks in culture rather than the isolated behavior of single units.

The divergence in dynamical complexity occurs early on in differentiation from the neural progenitor phase and peaks after 2 weeks of maturation. A recent study using the same ASD samples revealed that transient pathological priming of expression networks occurs around this timepoint, supporting the relevance of this early developmental period [Schafer et al., 2019]. Mirroring the findings in this chapter, MED deficits in the ASD group begin to lessen after 2 weeks of recording, along with a general trend of decreasing activity noted for across measures in both groups, perhaps due to the pruning of weakly connected neurons and the establishment of more mature circuits. The convergence of MED at later timepoints suggests that the electrical complexity gap may be a transient phenomenon in development that nevertheless may disrupt the early formation of neural connections and contribute to ASD pathology at later timepoints. The mechanism by which neuronal complexity deficits may propagate to the circuit and brain-wide

levels to drive pathology remains an interesting avenue for further study.

The clear and sustained reduction in complexity found in the ASD group may arise from deficiencies in synaptogenesis, which were previously reported in this cohort [Marchetto et al., 2016]. Diffuclty in forming synaptic connections in vitro likely diminishes the underlying complexity of their electrical outputs, as measured by MED. Other iPSC studies report an excess in ASD synapse formation, potentially highlighting the heterogeneity of cellular mechanisms in this disorder [Mariani et al., 2015]. Additionally, compared to simulated Hénon map data, real electrical recordings contain unpredictable and uninterpretable noise that leads to a non-zero FNN percentage and imprecision in MED estimation. As such, the estimated MEDs in our analysis may not exactly represent the true dimensionality of these neuronal systems. However, MED still provides value, because it reflects the relative dimensional complexity of the electrical signals, which is sufficient to distinguish the ASD and control group. A key area of investigation is the generalizability of MED as an endophenotype in other cohorts of ASD, particularly those without an early brain overgrowth phenotype.

The value of a continuous endophenotype of ASD is that it more robustly captures individual and group-wise variance than binary labels for follow-up analyses, such as RNA sequencing. The detection of molecular changes related to idiopathic ASD is complicated by the immense genetic heterogeneity associated with the disorder [Devlin and Scherer, 2012, **?**]. Though this study sample is much smaller than those typically recruited for genomic analyses, this iPSC-based approach offers a major advantage over other methods in that cellular variables relevant to disease may be directly recorded and studied in relation to gene expression changes [Vadodaria et al., 2018]. In this manner, I are able to find molecular changes associated with idiopathic ASD by tapping into the fundamental electrical dynamics that lie at the core of the disease. MED was associated with a broader differential gene expression signature than ASD status (53 vs 7 differentially expressed giants, after FDR correction), and these genes were associated with ASD-relevant genes and pathways such as neurogenesis, cell migration, and

synaptic components. These findings agree with sequencing studies that have shown ASD impacts a variety of neurodevelopmental processes such as neuronal proliferation, migration, and synaptogenesis [Courchesne et al., 2018, Krumm et al., 2014, Marchetto et al., 2016, Mariani et al., 2015, Parikshak et al., 2013, Pinto et al., 2014, **?**, Voineagu et al., 2011, Willsey et al., 2013]. Therefore, the validity of MED as a cellular endophenotype for ASD is demonstrated both directly, through electrical recordings, and indirectly, through the examination of gene expression trends related to MED.

ASD disrupts the normal formation of brain circuits during the early stages of life. In concordance with the literature, this study shows that the MED gene signature is most relevant during fetal development [Courchesne et al., 2018, Parikshak et al., 2013, Willsey et al., 2013]. However, some of the results in this study differ from the findings in related work. First, unlike other studies, which primarily localize gene expression signatures to cortical regions, I find that the expression changes correlated to MED impact both cortical structures (e.g., frontal and temporal regions) and deeper structures (e.g., hippocampus and thalamus) [Parikshak et al., 2013, Willsey et al., 2013]. However, recent studies using functional imaging have linked ASD to reductions in hippocampal activity and disruptions in thalamocortical connectivity [Cooper et al., 2017, Nair et al., 2013]. Second, the MED expression signature was also implicated in the late childhood timepoint, suggesting either mechanistic heterogeneity within the idiopathic ASD population or complexity in longitudinal disease mechanisms. The meaning of this finding is unclear at present, due to a lack of appropriate longitudinal genetic or imaging datasets for ASD. Regardless, these spatiotemporal findings indicate a mixture of previously demonstrated and novel associations uncovered through the integration electrical endophenotypes for ASD and gene expression analysis.

In summary, I demonstrate that nonlinear dynamical analysis can be applied to electrical recordings from patient-derived neurons to reveal differences in MED complexity between ASD and control subjects. The MED is associated with gene expression changes that both validate

existing genetic and mechanistic studies and present new knowledge, such as the implication of fetal basal brain regions and late childhood cortical development in ASD mechanisms. iPSC-derived models are a rich platform to study neurodevelopmental disorders because they allow for the examination of cellular disease mechanisms on otherwise inaccessible tissue. Further investigation of MED as an endophenotype for ASD will shed light on the utility of this novel approach. I believe that this integration of methods is a good strategy for tackling the immense heterogeneity associated with idiopathic ASD and hope that it may facilitate future studies in autism and related disorders.

Chapter 3, in full, is a reprint of the material as it has been written in a manuscript that has been submitted for publication. The authors of this study are Debha Amatya, Sara Linker, Ana Mendes, Renata Santos, Galina Erikson, Maxim Shokhirev, Yuansheng Zhou, Tatyana Sharpee, Fred Gage, Maria Marchetto, and Yeni Kim. The dissertation author was the primary investigator of this paper.

Chapter 4

Conclusions

Neuropsychiatry is in need of biologically derived markers for the classification and characterization of mental disorders. This dissertation covers two innovative approaches that quantitatively describe the genomic and electrical patterns of ASD. Both methods show promise in their ability to resolve ASD-related changes from normal function and concordance with existing findings in the literature.

In Chapter 2, I explore the question of how machine learning can be applied to massive whole genome sequencing datasets to train predictive classification models of ASD. I present genome vectorization as a method that facilitates this process by holistically summarizing variant burden per gene in the genome in a useful vector format. By training classification models and testing both within and across datasets, I provide convincing evidence that this method enables genomic classification of ASD subjects. Further, I argue that these models detect disease-specific risk in expected genes, pathways, neurodevelopmental time points, brain regions, and cell types related to ASD. Thus, I provide a preliminary demonstration of the potential utility of genome vectorization and machine learning for the molecular classification of disease risk in ASD, warranting further study of clinical utility and applicability in other traits or disorders of interest.

Outside of the applications discussed in this dissertation, the uses of a vectorized genomic representation are manifold. In essence, this novel representation for whole genome sequencing data is a new means for interrogating the link between genotype and phenotype. Apart from binary classification, potential uses for variant burden vectors include correlation model training to prognosticate continuous clinical variables (e.g. symptom severity scores or age of onset), clustering analysis to identify previously unidentified genomic subtypes within a complex disease, and multinomial classification models to differentiate phenotypically similar disorders (e.g. bipolar and major depression). With the mounting need for biomarkers for diagnosis and prognosis in neuropsychiatry, techniques such as genome vectorization offer hope for families that desperately seek the benefits of precise, quantitative tools that already exist in other branches of medicine.

In Chapter 3, I applied dynamical analysis techniques to present a novel marker of neuronal dysfunction in ASD. Though the molecular determinants of ASD are heterogeneous, it is likely that mechanistic convergence can be detected in the electrical activity of neurons. By nature, this activity is highly nonlinear, resulting from the complex interactions between numerous units. Therefore, the dynamical analysis framework was well-suited for studying these signals. Particularly, my analysis suggests that genetic alterations in ASD neurons manifest as defective neuronal network formation and reduced dynamical complexity in early development, as measured by MED. The gene expression correlates of MED overlap with known autism risk genes, as well as neurodevelopmental pathways and relevant timepoints in brain regions. The combination of dynamical analysis with iPSC-derived cellular models is a powerful new way to tap into disease-related electrical alterations that may be missed by standard electrical measures.

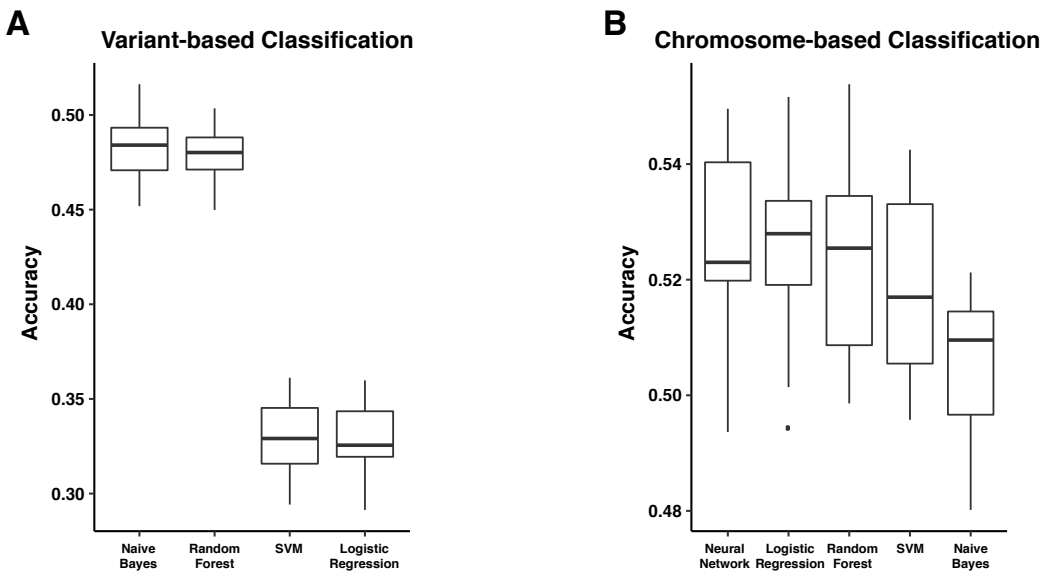
Autism exists manifests as a diverse spectrum, and measures like MED allow for patients to be more informatively described, in comparison to binary case-control labels. In order to further understand dynamical complexity, I believe that it is important to forge ahead with studying both the underlying causes and emergent effects of this phenomenon. Specifically, rigorous biological experimentation is needed to isolate the link between synaptic disruption and MED. Additionally,

in vivo electrical and behavioral recordings are likely required to trace the effects of reduced dynamical complexity to circuit level and behavioral phenotypes associated with ASD. Till then, value may be added by replicating this finding in other ASD cohorts, namely those without macrencephaly, to judge the reproducibility of the result.

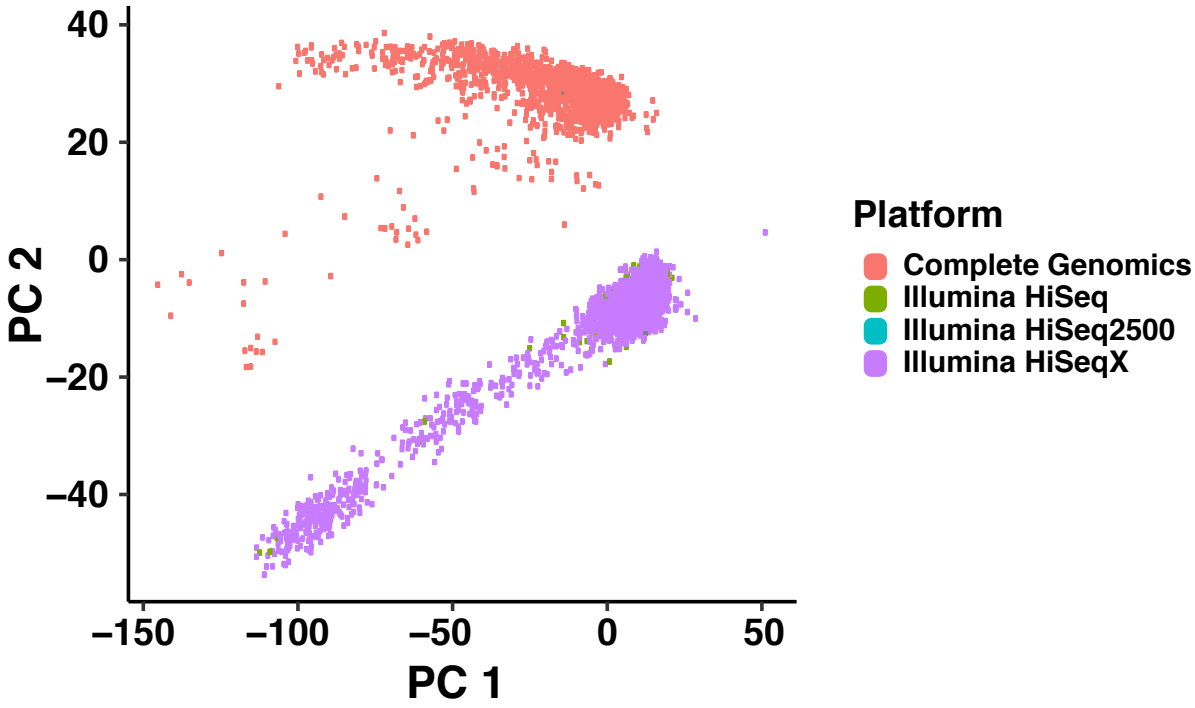
Both genome vectorization and dynamical analysis were performed with the intetion that the right technique should be selected to solve the right problem. Genomic risk prediction can be cast as a binary classification problem, which unlocks the power of machine learning to detect genomic patterns of risk in big data. Dynamical analysis is capable of modeling the complexity and nonlinearity of neuronal electrical signals to accurately resolve differences in neurodevelopment in ASD and control subjects. Though differing approaches, dynamical analysis and machine learning did lead to convergent findings related to ASD. First, genes implicated in both analyses overlap with autism risk genes implicated in GWAS and brain expressed genes. Second, cortical regions and prenatal timepoints are critical in the pathogenesis of ASD. Third, some of the pathways most relevant to ASD involve the synapse and ion channels. Therefore, molecular heterogeneity in ASD does impact stereotyped pathways, regions, and timepoints in development. Dynamical analysis and machine learning help us hone into ASD signatures in powerful and new ways, and the application of these tools to autism, or other complex disorders, may improve both our biological understanding and medical management of such diseases.

Appendix A

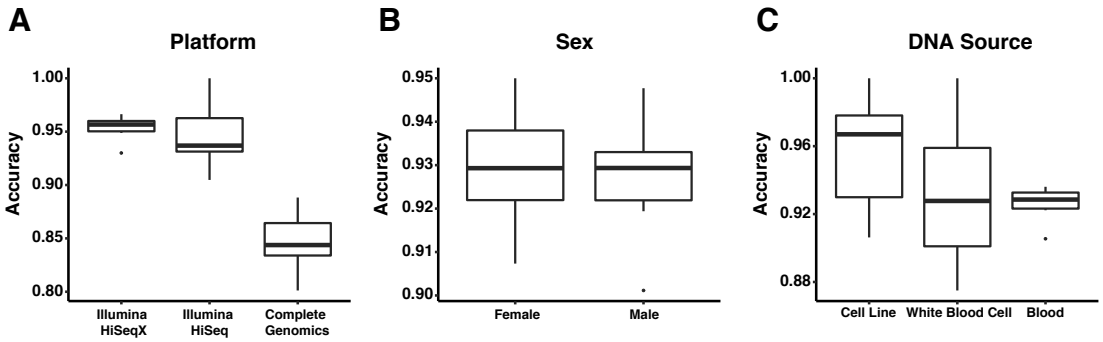
Genome Vectorization for Machine Learning Applications to ASD



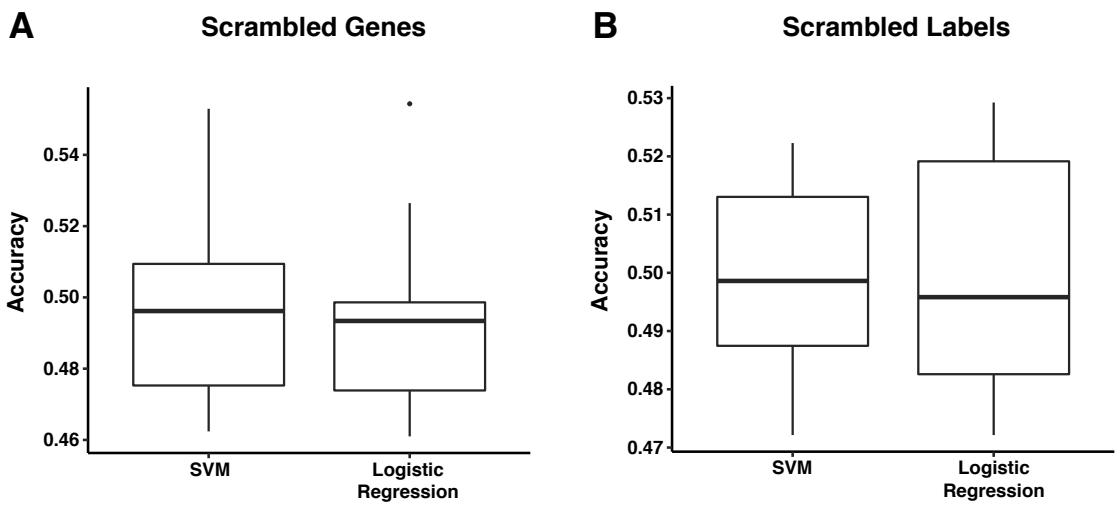
Appendix Figure A.1: Vectorization scale impacts classification performance. These alternative feature scales were attempted to motivate the choice of a gene-based vectorization. A. Variant-based vectors were constructed by forming binary presence/absence arrays from approximately 100,000 of the most commonly occurring variants in the MSSNG data. Of the four classifiers tested, none exceeded random accuracy. The artificial neural network model was unable to be tested due to insufficient computational resources. B. Variant burden was binned across autosomal chromosomes to test the performance of a chromosome-based vectorization. This too resulted in close to random classification accuracy.



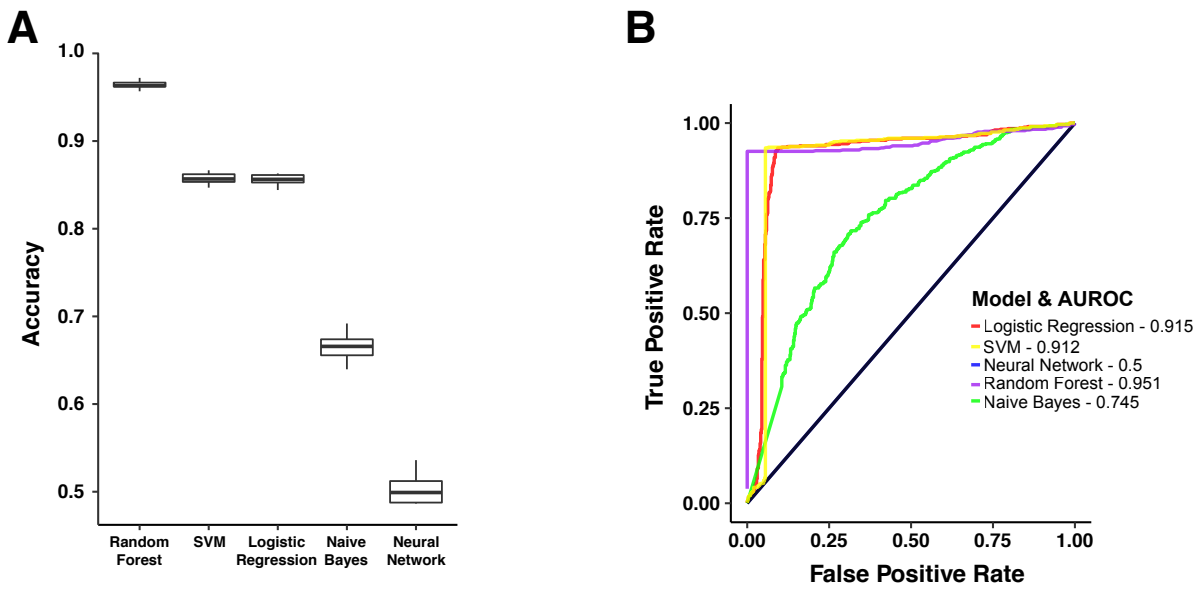
Appendix Figure A.2: Coloring the principal component (PC) analysis plot of the variant burden matrix by sequencing platform reveals that Illumina and Complete Genomics samples segregate. Therefore, sequencing platform is a notable covariate for variant burden vectors.



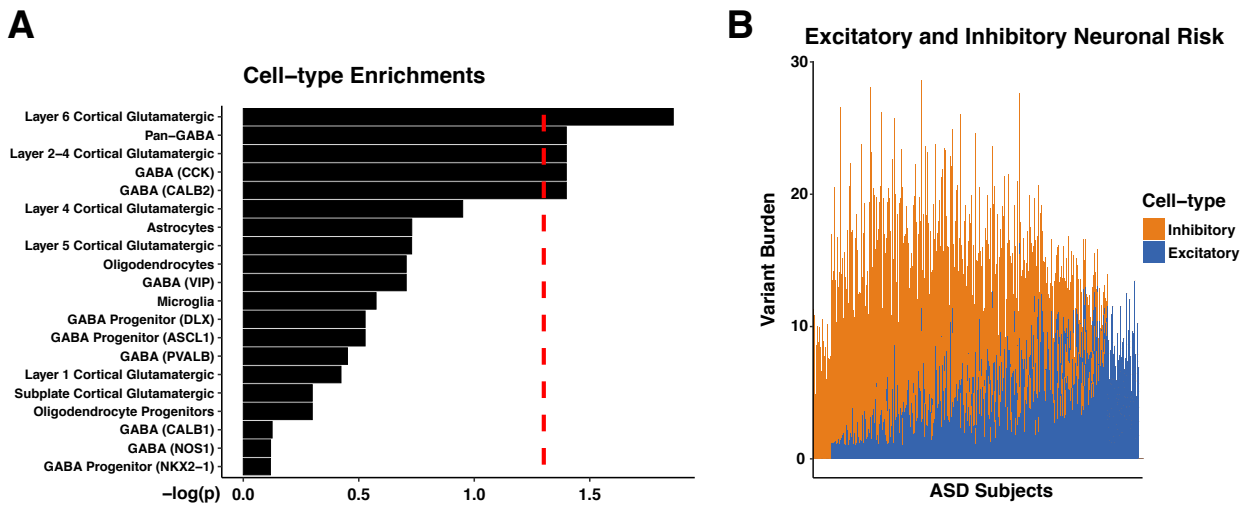
Appendix Figure A.3: Three important covariates were further examined by stratified analysis of performance: sequencing platform, sex, and DNA source. Accuracy was measured for the SVM classifier with 10-fold cross-validation. A. Illumina samples (HiSeqX and HiSeq) show better performance than the Complete Genomics samples, but both platforms show accuracy greater than 85%. B. Stratification by sex shows similarly high classification accuracy across gender. C. Stratification by DNA source shows similarly high performance across whole blood, white blood cell, and cell line samples.



Appendix Figure A.4: Classification performance is abolished by scrambling either case-control labels or gene columns per subject. Accuracy was measured for both the SVM and logistic regression (LR) classifiers with 10-fold cross-validation. A. Scrambling each subject's variant burden vector gene columns disrupts common patterns of variant burden across ASD vectors. Classification performance for this scenario is essentially random. B. Scrambling the case-control labels for the variant burden vectors also results in close to random performance. Therefore, the labels encode meaningful differences between the vectors of each group.



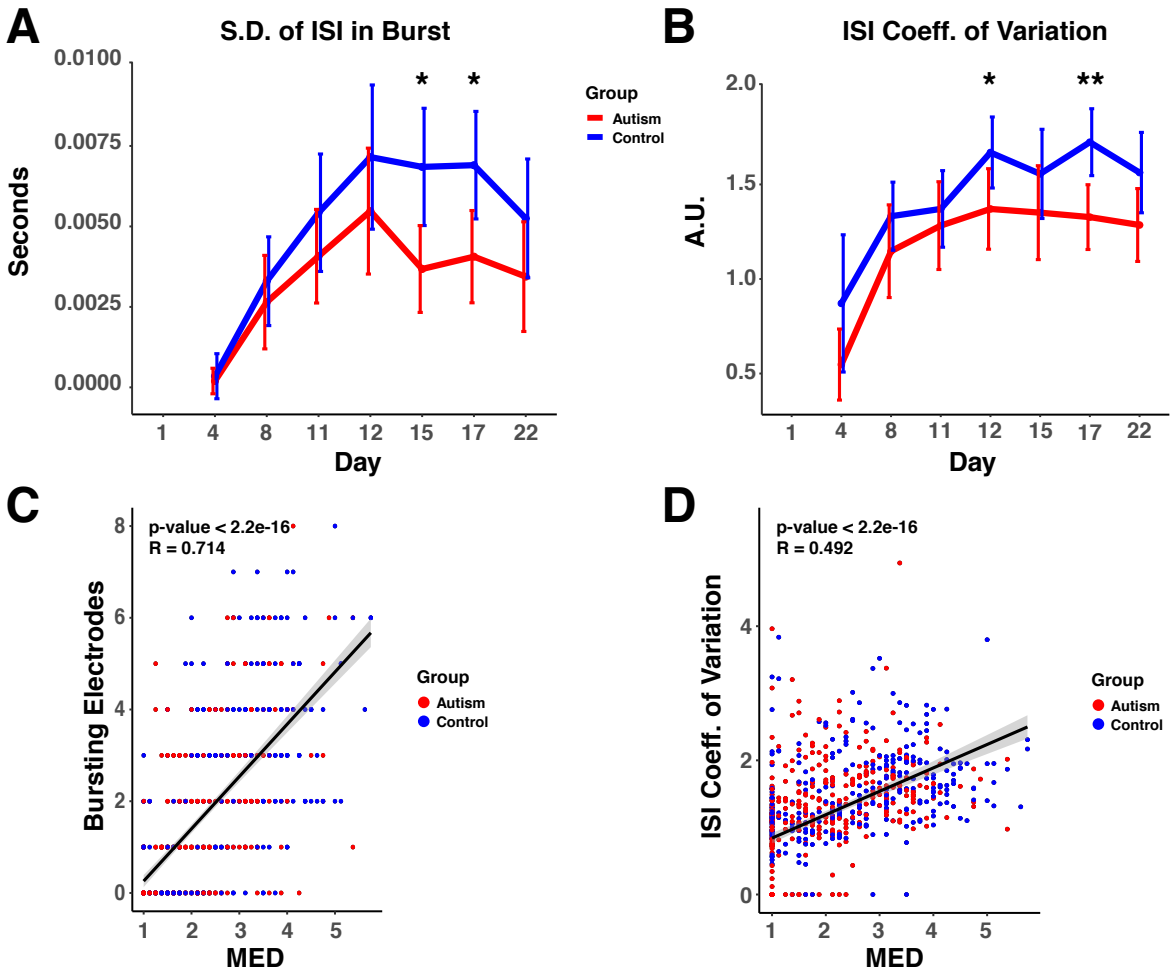
Appendix Figure A.5: Classification performance is reproducible in an independent ASD genomics data. The SFARI Simons Simplex Collection (SSC) was used to validate the genome vectorization and classification methodology. Unlike the primary MSSNG data, the SSC includes healthy sibling controls, which increase the complex of the learning problem and impact performance. A. High accuracy is obtained by three of the classification models (random forest, SVM, and logistic regression) but not by the neural network and naive Bayes classifiers. B. Similarly, the logistic regression, SVM, and random forest models are able to achieve a high degree of sensitivity and specificity for this classification task. The classifier curve colors are specified in the legend, and the black curve represents a random classifier. Area under the receiving operating characteristic curve (AUROC), a performance measure of binary classification, is also listed in the legend for each model.



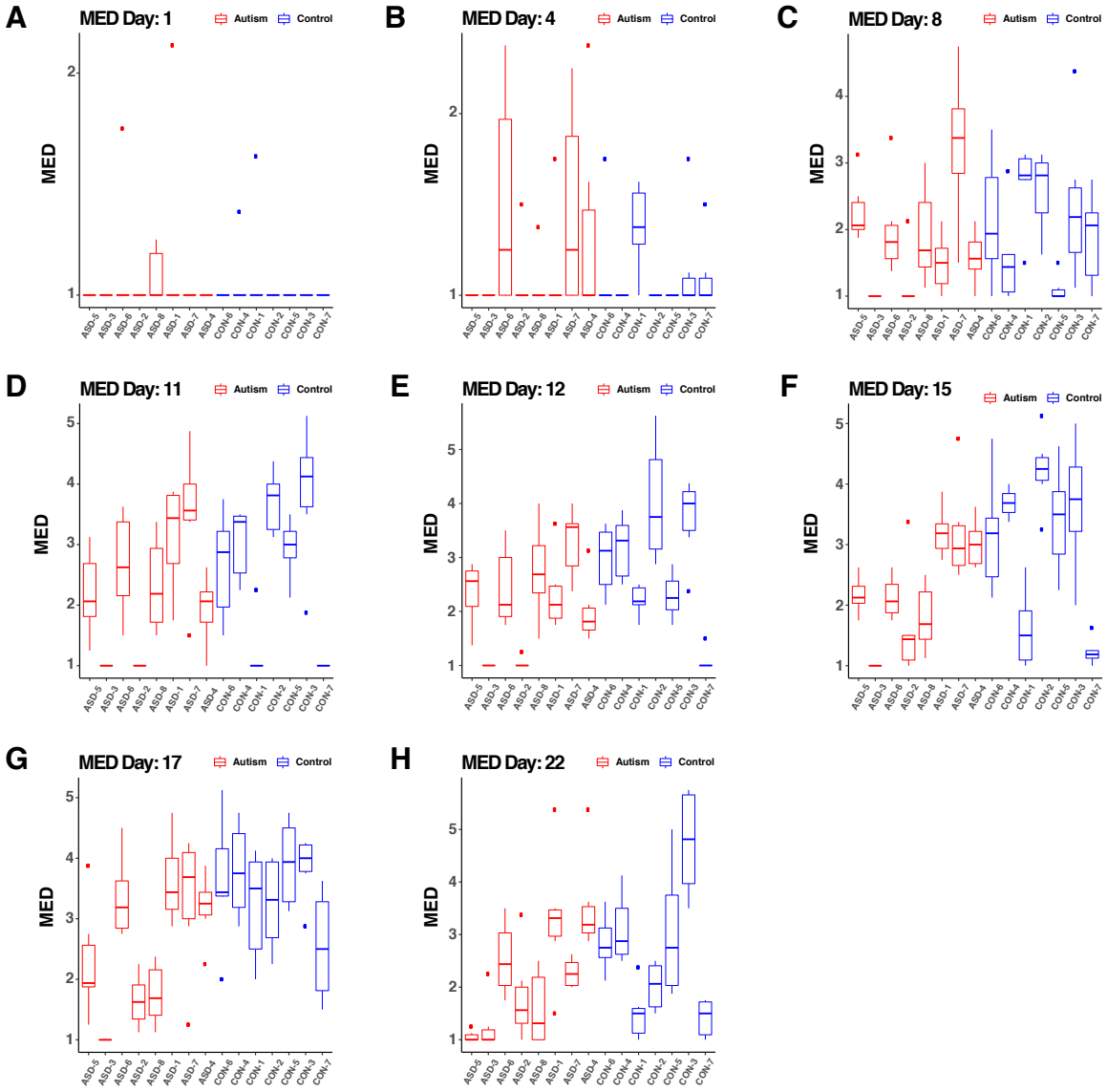
Appendix Figure A.6: A dataset containing genetic correlates for 20 neural cell types was used to examine enrichment in the genome-wide rankings predicted by the SVM classification model [Kang et al., 2011]. A. Highly ranked genes are enriched for both cortical glutamatergic neurons (layer 2-4 and layer 6) and GABAergic neurons (CCK and CALB2), as well as pan GABA cell types. B. ASD subjects vary with respect to both relative and absolute risk of excitatory (layer 2-4 and layer 6 cortical glutamatergic neurons) versus inhibitory (CCK and CALB2 cortical interneurons) cell-types. The vast majority of subjects possess nontrivial variant risk (low, moderate, and high damage) in a mixture of excitatory and inhibitory neuronal genes.

Appendix B

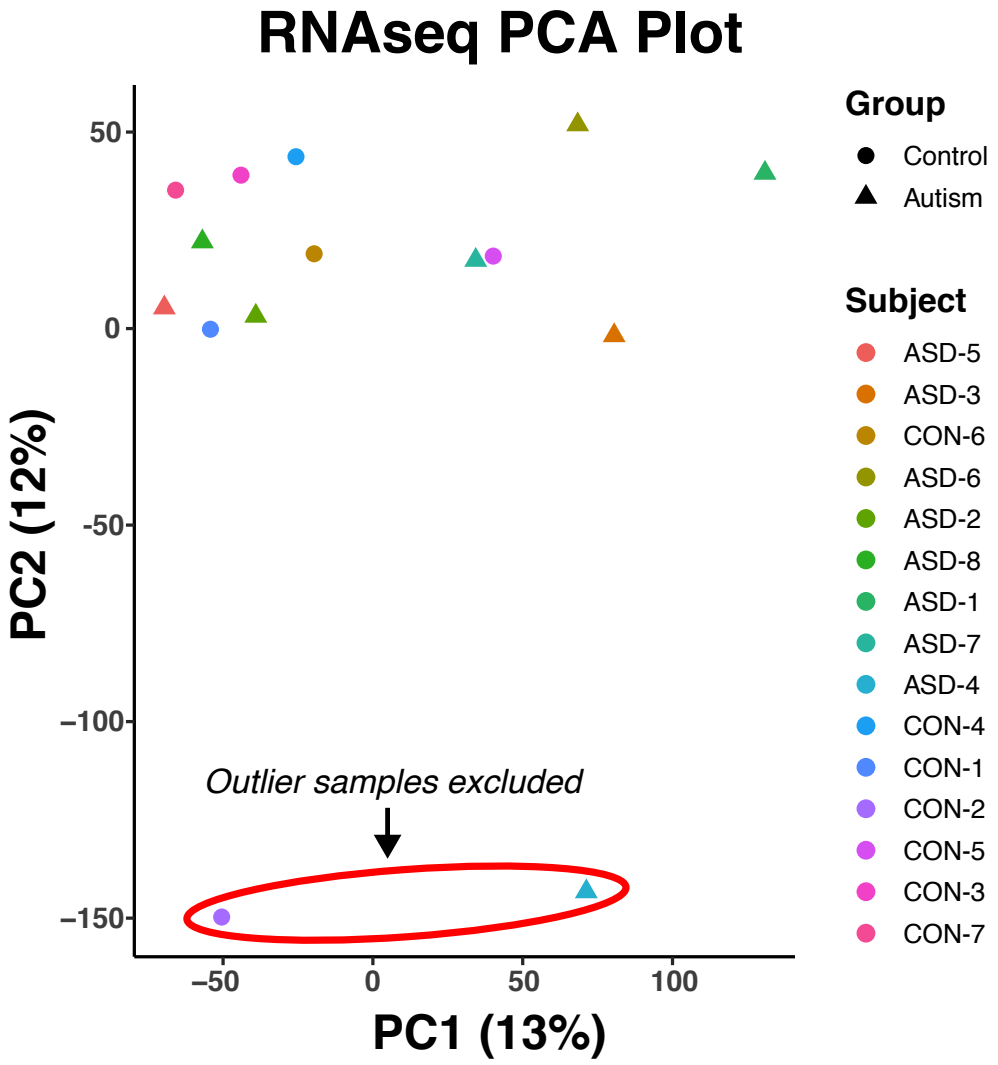
Dynamical Analysis of Neuronal Electrical Activity in ASD



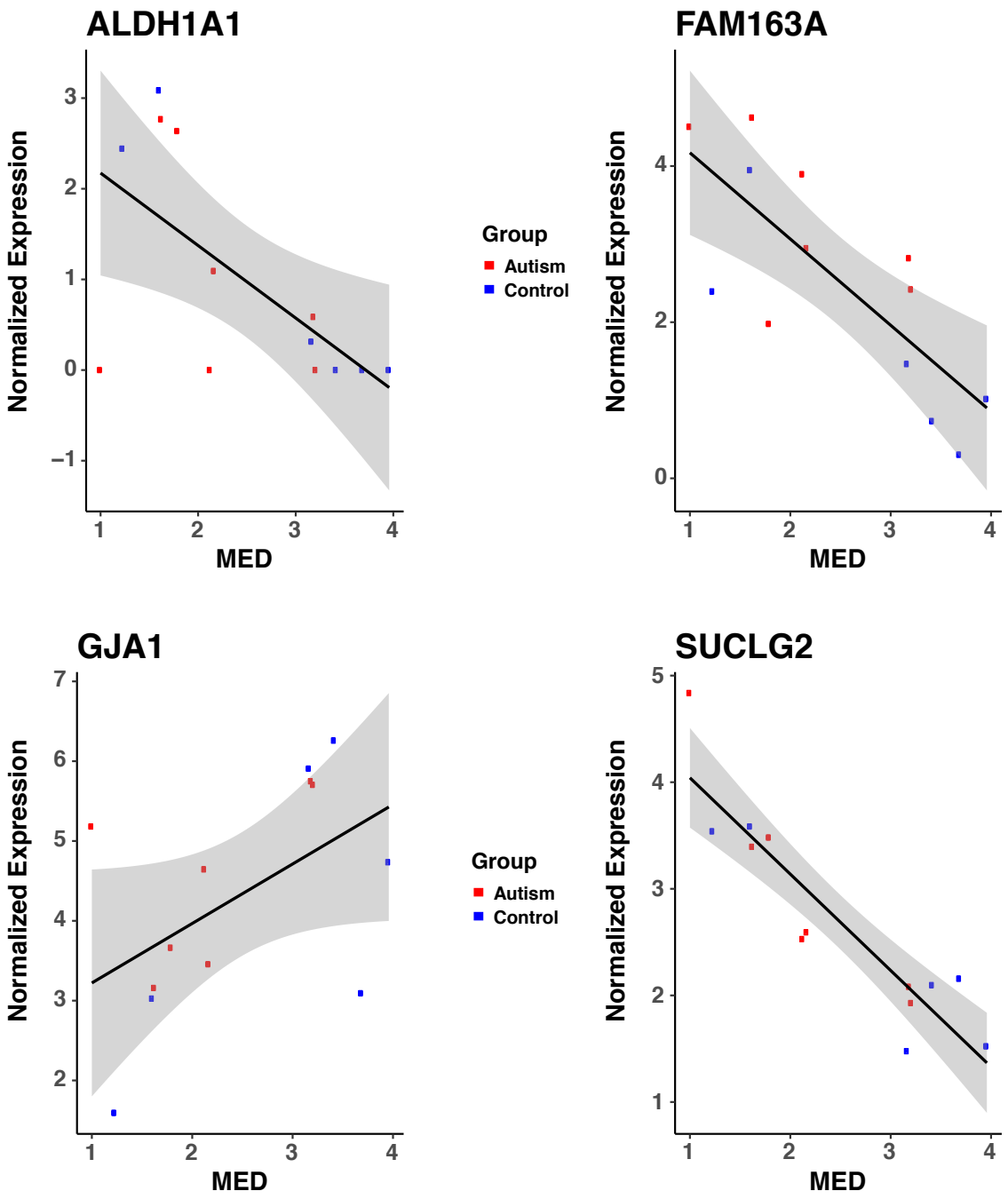
Appendix Figure B.1: Additional MEA Spiking Variables and Correlation to the MED. Time-series plots of the ASD (red) and control (blue) groups for A. standard deviation (S.D.) of interspike interval (ISI) in bursts, B. ISI coefficient of variation show statistically significant trends across ASD and control samples. The concordance of these trends with the MED suggests a relationship between bursting and spiking variance and dynamical complexity. The plots show average values at each timepoint with a 95% confidence interval. Significance was tested using a Welch's two-sided t-test. Correlation plots between MED and spiking variables show significant associations between MED and C. number of bursting electrodes and D. ISI coefficient of variation. For these panels, the line represents a fitted linear with a 95% confidence interval for the model in grey shading.



Appendix Figure B.2: MED Intersubject Variability. A-H show boxplots for 7 control (blue) and 8 ASD (red) subject MED values averaged over replicates for eight timepoints of MEA recording. The diminished dynamical complexity of the ASD subjects becomes apparent in the second week timepoints, through the course of the remaining days. Despite the group differences, these panels reveal the individual subject variation in the measurement. F. Of note, this panel shows the MED subject values for day 15 of the recordings, which was the day that cells were extracted for RNAseq. The average subject MED values on this day were used as the variable of interest for gene expression analysis.



Appendix Figure B.3: RNA Sequencing Counts Matrix PCA Plot. Principal components analysis plot of the RNA sequencing counts matrix reveals two outlier subjects, CON-2 and ASD-4, which were removed for differential expression analysis. In parentheses, the percentage of variance explained by each principal component is given.



Appendix Figure B.4: Examples of Genes Differentially Expressed for MED. Individual gene plots for genes that are differentially expressed for the MED value reveal linear correlations between gene expression and dynamical complexity. These relationships capture transcriptomic variance more comprehensively than ASD-control group labels. For all panels, the grey shading represents a 95% confidence interval around the mean values, as estimated by a linear model.

Bibliography

[Abadi et al., 2015] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[Abrahams et al., 2013] Abrahams, B. S., Arking, D. E., Campbell, D. B., Mefford, H. C., Morrow, E. M., Weiss, L. A., Menashe, I., Wadkins, T., Banerjee-Basu, S., and Packer, A. (2013). Sfari gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (asds). *Molecular Autism*, 4(1):36.

[Akar et al., 2015] Akar, S. A., Kara, S., Agambayev, S., and Bilgiç, V. (2015). Nonlinear analysis of eegs of patients with major depression during different emotional states. *Computers in Biology and Medicine*, 67:49–60.

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

[Andrews, 2010] Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

[Asperger, 1944] Asperger, H. (1944). Die autistischen psychopathen im kindesalter. *Archiv fur Psychiatrie und Nervenkrankheiten*, 117(1):76–136.

[Baio, 2014] Baio, J. (2014). Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, United States,

2010. *MMWR. Surveillance summaries : Morbidity and mortality weekly report. Surveillance summaries*, 66(2).

[Baudry and Taketani, 2006] Baudry, M. and Taketani, M. (2006). *Advances in network electrophysiology: Using multi-electrode arrays*. Springer.

[Bick et al., 2017] Bick, D., Fraser, P. C., Gutzeit, M. F., Harris, J. M., Hambuch, T. M., Helbling, D. C., Jacob, H. J., Kersten, J. N., Leuthner, S. R., May, T., North, P. E., Prisco, S. Z., Schuler, B. A., Shimoyama, M., Strong, K. A., Van Why, S. K., Veith, R., Verbsky, J., Weborg, A. M., Wilk, B. M., Willoughby, R. E., Worthey, E. A., and Dimmock, D. P. (2017). Successful application of whole genome sequencing in a medical genetics clinic. *Journal of Pediatric Genetics*, 06(02):061–076.

[Bosl et al., 2011] Bosl, W., Tierney, A., Tager-Flusberg, H., and Nelson, C. (2011). Eeg complexity as a biomarker for autism spectrum disorder risk. *BMC Medicine*, 9(1):18.

[Boyle et al., 2017] Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186.

[Butter et al., 2003] Butter, E. M., Wynn, J., and Mulick, J. A. (2003). Early intervention critical to autism treatment. *Pediatric Annals*, 32(10):677–684.

[Bzdok and Meyer-Lindenberg, 2018] Bzdok, D. and Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(3):223–230.

[C Yuen et al., 2017] C Yuen, R. K., Merico, D., Bookman, M., L Howe, J., Thiruvahindrapuram, B., Patel, R. V., Whitney, J., Deflaux, N., Bingham, J., Wang, Z., Pellecchia, G., Buchanan, J. A., Walker, S., Marshall, C. R., Uddin, M., Zarrei, M., Deneault, E., D'Abate, L., Chan, A. J. S., Koyanagi, S., Paton, T., Pereira, S. L., Hoang, N., Engchuan, W., Higginbotham, E. J., Ho, K., Lamoureux, S., Li, W., MacDonald, J. R., Nalpathamkalam, T., Sung, W. W. L., Tsoi, F. J., Wei, J., Xu, L., Tasse, A.-M., Kirby, E., Van Etten, W., Twigger, S., Roberts, W., Drmic, I., Jilderda, S., Modi, B. M., Kellam, B., Szego, M., Cytrynbaum, C., Weksberg, R., Zwaigenbaum, L., Woodbury-Smith, M., Brian, J., Senman, L., Iaboni, A., Doyle-Thomas, K., Thompson, A., Chrysler, C., Leef, J., Savion-Lemieux, T., Smith, I. M., Liu, X., Nicolson, R., Seifer, V., Fedele, A., Cook, E. H., Dager, S., Estes, A., Gallagher, L., Malow, B. A., Parr, J. R., Spence, S. J., Vorstman, J., Frey, B. J., Robinson, J. T., Strug, L. J., Fernandez, B. A., Elsabbagh, M., Carter, M. T., Hallmayer, J., Knoppers, B. M., Anagnostou, E., Szatmari, P., Ring, R. H., Glazer, D., Pletcher, M. T., and Scherer, S. W. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature Neuroscience*, 20:602 EP –.

[Castelvecchi, 2016] Castelvecchi, D. (2016). Can we open the black box of ai? *Nature News*, 538(7623):20.

[Chao et al., 2010] Chao, H.-T., Chen, H., Samaco, R. C., Xue, M., Chahrour, M., Yoo, J., Neul, J. L., Gong, S., Lu, H.-C., Heintz, N., Ekker, M., Rubenstein, J. L. R., Noebels, J. L., Rosenmund, C., and Zoghbi, H. Y. (2010). Dysfunction in GABA signalling mediates autism-like stereotypies and rett syndrome phenotypes. *Nature*, 468:263.

[Chapman et al., 2011] Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., Maio, M., Hogg, D., Lorigan, P., Lebbe, C., Jouary, T., Schadendorf, D., Ribas, A., O'Day, S. J., Sosman, J. A., Kirkwood, J. M., Eggermont, A. M., Dreno, B., Nolop, K., Li, J., Nelson, B., Hou, J., Lee, R. J., Flaherty, K. T., and McArthur, G. A. (2011). Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. *New England Journal of Medicine*, 364(26):2507–2516. PMID: 21639808.

[Chollet et al., 2015] Chollet, F. et al. (2015). Keras. `https://keras.io`.

[Chow et al., 2012] Chow, M. L., Pramparo, T., Winn, M. E., Barnes, C. C., Li, H.-R., Weiss, L., Fan, J.-B., Murray, S., April, C., Belinson, H., Fu, X.-D., Wynshaw-Boris, A., Schork, N. J., and Courchesne, E. (2012). Age-dependent brain gene expression and copy number anomalies in autism suggest distinct pathological processes at young versus mature ages. *PLOS Genetics*, 8(3):1–14.

[Cirulli and Goldstein, 2010] Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415.

[Clarke et al., 2015] Clarke, T.-K., Lupton, M. K., Fernandez-Pujals, A. M., Starr, J., Davies, G., Cox, S., Pattie, A., Liewald, D. C., Hall, L. S., MacIntyre, D. J., Smith, B. H., Hocking, L. J., Padmanabhan, S., Thomson, P. A., Hayward, C., Hansell, N. K., Montgomery, G. W., Medland, S. E., Martin, N. G., Wright, M. J., Porteous, D. J., Deary, I. J., and McIntosh, A. M. (2015). Common polygenic risk for autism spectrum disorder (asd) is associated with cognitive ability in the general population. *Molecular Psychiatry*, 21:419 EP –.

[Colvert et al., 2015] Colvert, E., Tick, B., McEwen, F., Stewart, C., Curran, S. R., Woodhouse, E., Gillan, N., Hallett, V., Lietz, S., Garnett, T., Ronald, A., Plomin, R., Rijsdijk, F., Happ, F., and Bolton, P. (2015). Heritability of Autism Spectrum Disorder in a UK Population-Based Twin SampleHeritability of Autism Spectrum DisorderHeritability of Autism Spectrum Disorder. *JAMA Psychiatry*, 72(5):415–423.

[Cooper et al., 2017] Cooper, R. A., Richter, F. R., Bays, P. M., Plaisted-Grant, K. C., Baron-Cohen, S., and Simons, J. S. (2017). Reduced hippocampal functional connectivity during episodic memory retrieval in autism. *Cerebral Cortex*, 27(2):888–902.

[Courchesne et al., 2003] Courchesne, E., Carper, R., and Akshoomoff, N. (2003). Evidence of brain overgrowth in the first year of life in autism. *JAMA*, 290(3):337–344.

[Courchesne et al., 2011] Courchesne, E., Mouton, P. R., Calhoun, M. E., Semendeferi, K., Ahrens-Barbeau, C., Hallet, M. J., Barnes, C. C., and Pierce, K. (2011). Neuron number and size in prefrontal cortex of children with autism. *JAMA*, 306(18):2001–2010.

[Courchesne and Pierce, 2005] Courchesne, E. and Pierce, K. (2005). Brain overgrowth in autism during a critical time in development: implications for frontal pyramidal neuron and interneuron development and connectivity. *International Journal of Developmental Neuroscience*, 23(2-3):153–170.

[Courchesne et al., 2018] Courchesne, E., Pramparo, T., Gazestani, V. H., Lombardo, M. V., Pierce, K., and Lewis, N. E. (2018). The ASD living biology: from cell proliferation to clinical phenotype. *Molecular Psychiatry*, page 1.

[Crick, 1970] Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561.

[Crippa et al., 2015] Crippa, A., Salvatore, C., Perego, P., Forti, S., Nobile, M., Molteni, M., and Castiglioni, I. (2015). Use of machine learning to identify children with autism and their motor abnormalities. *Journal of Autism and Developmental Disorders*, 45(7):2146–2156.

[DeRosa et al., 2018] DeRosa, B. A., El Hokayem, J., Artimovich, E., Garcia-Serje, C., Phillips, A. W., Van Booven, D., Nestor, J. E., Wang, L., Cuccaro, M. L., Vance, J. M., Pericak-Vance, M. A., Cukier, H. N., Nestor, M. W., and Dykxhoorn, D. M. (2018). Convergent pathways in idiopathic autism revealed by time course transcriptomic analysis of patient-derived neurons. *Scientific Reports*, 8(1):8423.

[Devlin and Scherer, 2012] Devlin, B. and Scherer, S. W. (2012). Genetic architecture in autism spectrum disorder. *Current Opinion in Genetics & Development*, 22(3):229–237.

[Dobin et al., 2013] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

[Domingos, 2012] Domingos, P. M. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

[Dudbridge, 2013] Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, 9(3):e1003348.

[Eilbeck et al., 2005] Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005). The sequence ontology: a tool for the unification of genome annotations. *Genome biology*, 6(5):R44.

[Emerson et al., 2017] Emerson, R. W., Adams, C., Nishino, T., Hazlett, H. C., Wolff, J. J., Zwaigenbaum, L., Constantino, J. N., Shen, M. D., Swanson, M. R., Elison, J. T., Kandala, S., Estes, A. M., Botteron, K. N., Collins, L., Dager, S. R., Evans, A. C., Gerig, G., Gu, H., McKinstry, R. C., Paterson, S., Schultz, R. T., Styner, M., , Schlaggar, B. L., Pruett, J. R., and Piven, J. (2017). Functional neuroimaging of high-risk 6-month-old infants predicts a diagnosis of autism at 24 months of age. *Science Translational Medicine*, 9(393).

[Fernández et al., 2018] Fernández, A., Al-Timemy, A. H., Ferre, F., Rubio, G., and Escudero, J. (2018). Complexity analysis of spontaneous brain activity in mood disorders: a magnetoencephalography study of bipolar disorder and major depression. *Comprehensive Psychiatry*, 84:112–117.

[Fischbach and Lord, 2010] Fischbach, G. D. and Lord, C. (2010). The simons simplex collection: a resource for identification of autism genetic risk factors. *Neuron*, 68(2):192–195.

[Folstein and Rutter, 1977] Folstein, S. and Rutter, M. (1977). Infantile autism: a genetic study of 21 twin pairs. *Journal of Child psychology and Psychiatry*, 18(4):297–321.

[Fountain et al., 2011] Fountain, C., King, M. D., and Bearman, P. S. (2011). Age of diagnosis for autism: individual and community factors across 10 birth cohorts. *Journal of Epidemiology and Community Health*, 65(6):503–510.

[Friedman et al., 2001] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:.

[Furey et al., 2000] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914.

[Gaffrey et al., 2007] Gaffrey, M. S., Kleinhans, N. M., Haist, F., Akshoomoff, N., Campbell, A., Courchesne, E., and Müller, R.-A. (2007). A typical participation of visual cortex during word processing in autism: An fMRI study of semantic decision. *Neuropsychologia*, 45(8):1672–1684.

[Garbett et al., 2008] Garbett, K., Ebert, P. J., Mitchell, A., Lintas, C., Manzi, B., Mirnics, K., and Persico, A. M. (2008). Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiology of Disease*, 30(3):303 – 311.

[Gervais et al., 2004] Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Sfaello, I., Barthélémy, C., Brunelle, F., Samson, Y., and Zilbovicius, M. (2004). Abnormal cortical voice processing in autism. *Nature Neuroscience*, 7(8):801.

[Geschwind and State, 2015] Geschwind, D. H. and State, M. W. (2015). Gene hunting in autism spectrum disorder: on the path to precision medicine. *The Lancet Neurology*, 14(11):1109–1120.

[Gohlke et al., 2007] Gohlke, J. M., Griffith, W. C., and Faustman, E. M. (2007). Computational models of neocortical neuronogenesis and programmed cell death in the developing mouse, monkey, and human. *Cerebral Cortex*, 17(10):2433–2442.

[Green et al., 2011] Green, E. D., Guyer, M. S., and National Human Genome Research Institute (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470(7333):204.

[Gupta et al., 2014] Gupta, S., Ellis, S. E., Ashar, F. N., Moes, A., Bader, J. S., Zhan, J., West, A. B., and Arking, D. E. (2014). Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nature Communications*, 5:5748.

[Hazlett et al., 2011] Hazlett, H. C., Poe, M. D., Gerig, G., Styner, M., Chappell, C., Smith, R. G., Vachet, C., and Piven, J. (2011). Early brain overgrowth in autism associated with an increase in cortical surface area before age 2 years. *Archives of General Psychiatry*, 68(5):467–476.

[Heinsfeld et al., 2018] Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage: Clinical*, 17:16–23.

[Hénon, 1976] Hénon, M. (1976). A two-dimensional mapping with a strange attractor. In *The Theory of Chaotic Attractors*, pages 94–102. Springer.

[Howsmon et al., 2017] Howsmon, D. P., Kruger, U., Melnyk, S., James, S. J., and Hahn, J. (2017). Classification and adaptive behavior prediction of children with autism spectrum disorder based upon multivariate data analysis of markers of oxidative stress and DNA methylation. *PLoS Computational Biology*, 13(3):e1005385.

[Illumina, 2015] Illumina (2015). iGenomes online.

[Iossifov et al., 2014] Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., Smith, J. D., Paeper, B., Nickerson, D. A., Dea, J., Dong, S., Gonzalez, L. E., Mandell, J. D., Mane, S. M., Murtha, M. T., Sullivan, C. A., Walker, M. F., Waqar, Z., Wei, L., Willsey, A. J., Yamrom, B., Lee, Y.-h., Grabowska, E., Dalkic, E., Wang, Z., Marks, S., Andrews, P., Leotta, A., Kendall, J., Hakker, I., Rosenbaum, J., Ma, B., Rodgers, L., Troge, J., Narzisi, G., Yoon, S., Schatz, M. C., Ye, K., McCombie, W. R., Shendure, J., Eichler, E. E., State, M. W., and Wigler, M. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, 515:216 EP –.

[Jacob et al., 2019] Jacob, S., Wolff, J. J., Steinbach, M. S., Doyle, C. B., Kumar, V., and Elison, J. T. (2019). Neurodevelopmental heterogeneity and computational approaches for understanding autism. *Translational Psychiatry*, 9(1):63.

[Jaganathan et al., 2019] Jaganathan, K., Panagiotopoulou, S. K., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S. J., and Farh, K. K.-H. (2019). Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535 – 548.e24.

[Jeong et al., 1998] Jeong, J., Kim, D.-J., Chae, J.-H., Kim, S. Y., Ko, H.-J., and Paik, I.-H. (1998). Nonlinear analysis of the eeg of schizophrenics with optimal embedding dimension. *Medical Engineering & Physics*, 20(9):669–676.

[Jones et al., 2014] Jones, E., Oliphant, T., and Peterson, P. (2014). {SciPy}: open source scientific tools for {Python}.

[Kang et al., 2011] Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M., Pletikos, M., Meyer, K. A., Sedmak, G., Guennel, T., Shin, Y., Johnson, M. B., Krsnik, Ž., Mayer, S., Fertuzinhos, S., Umlauf, S., Lisgo, S. N., Vortmeyer, A., Weinberger, D. R., Mane, S., Hyde, T. M., Huttner, A., Reimers, M., Kleinman, J. E., and Šestan, N. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, 478:483 EP –.

[Kanner et al., 1943] Kanner, L. et al. (1943). Autistic disturbances of affective contact. *Nervous Child*, 2(3):217–250.

[Kennel et al., 1992] Kennel, M. B., Brown, R., and Abarbanel, H. D. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A*, 45(6):3403.

[Keogh and Mueen, 2011] Keogh, E. and Mueen, A. (2011). Curse of dimensionality. In *Encyclopedia of machine learning*, pages 257–258. Springer.

[Khambata-Ford et al., 2007] Khambata-Ford, S., Garrett, C. R., Meropol, N. J., Basik, M., Harbison, C. T., Wu, S., Wong, T. W., Huang, X., Takimoto, C. H., Godwin, A. K., Tan, B. R., Krishnamurthi, S. S., Burris, H. A., Poplin, E. A., Hidalgo, M., Baselga, J., Clark, E. A., and Mauro, D. J. (2007). Expression of epiregulin and amphiregulin and k-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. *Journal of Clinical Oncology*, 25(22):3230–3237. PMID: 17664471.

[Khan et al., 2015] Khan, S., Michmizos, K., Tommerdahl, M., Ganesan, S., Kitzbichler, M. G., Zetino, M., Garel, K.-L. A., Herbert, M. R., Hämäläinen, M. S., and Kenet, T. (2015). Somatosensory cortex functional connectivity abnormalities in autism show opposite trends, depending on direction and spatial scale. *Brain*, 138(5):1394–1409.

[Kreuz, 2013] Kreuz, T. (2013). Synchronization Measures. In *Principles of Neural Coding*, pages 97–120.

[Krishnan et al., 2016] Krishnan, A., Zhang, R., Yao, V., Theesfeld, C. L., Wong, A. K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., and Troyanskaya, O. G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature Neuroscience*, 19:1454 EP –.

[Krumm et al., 2014] Krumm, N., O'Roak, B. J., Shendure, J., and Eichler, E. E. (2014). A de novo convergence of autism genetics and molecular neuroscience. *Trends in Neurosciences*, 37(2):95–105.

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim,

J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., Morgan, M. J., Consortium, I. H. G. S., Whitehead Institute for Biomedical Research, C. f. G. R., Centre:, T. S., Center, W. U. G. S., Institute:, U. D. J. G., of Medicine Human Genome Sequencing Center:, B. C., Center:, R. G. S., Genoscope, UMR-8030:, C., Department of Genome Analysis, I. o. M. B., Center:, G. S., Center:, B. G. I. G., Multimegabase Sequencing Center, T. I. f. S. B., Center:, S. G. T., of Oklahoma's Advanced Center for Genome Technology:, U., for Molecular Genetics:, M. P. I., Cold Spring Harbor Laboratory, L. A. H. G. C., for Biotechnology:, G.-G. R. C., *Genome Analysis Group (listed in alphabetical order, a. i. i. l. u. o. h., Scientific management: National Human Genome Research Institute, U. N. I. o. H., Center:, S. H. G., of Washington Genome Center:, U., Department of Molecular Biology, K. U. S. o. M., of Texas Southwestern Medical Center at

Dallas:, U., Office of Science, U. D. o. E., and Trust:, T. W. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Lavelle et al., 2014] Lavelle, T. A., Weinstein, M. C., Newhouse, J. P., Munir, K., Kuhlthau, K. A., and Prosser, L. A. (2014). Economic burden of childhood autism spectrum disorders. *Pediatrics*, 133(3):e520–e529.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436.

[Leek and Storey, 2007] Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161.

[Leigh and Du, 2015] Leigh, J. P. and Du, J. (2015). Brief report: Forecasting the economic burden of autism in 2015 and 2025 in the united states. *Journal of Autism and Developmental Disorders*, 45(12):4135–4139.

[Leung et al., 2016] Leung, M. K. K., Delong, A., Alipanahi, B., and Frey, B. J. (2016). Machine learning in genomic medicine: A review of computational problems and data sets. *Proceedings of the IEEE*, 104(1):176–197.

[Libbrecht and Noble, 2015] Libbrecht, M. W. and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321.

[Liu et al., 2017] Liu, X., Campanac, E., Cheung, H.-H., Ziats, M. N., Canterel-Thouennon, L., Raygada, M., Baxendale, V., Pang, A. L.-Y., Yang, L., Swedo, S., Thurm, A., Lee, T.-L., Fung, K.-P., Chan, W.-Y., Hoffman, D. A., and Rennert, O. M. (2017). Idiopathic autism: Cellular and molecular phenotypes in pluripotent stem cell-derived neurons. *Molecular Neurobiology*, 54(6):4507–4523.

[Lord et al., 2000] Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., and Rutter, M. (2000). The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, 30(3):205–223.

[Love et al., 2014] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.

[Luo et al., 2012] Luo, R., Sanders, S., Tian, Y., Voineagu, I., Huang, N., Chu, S. H., Klei, L., Cai, C., Ou, J., Lowe, J., Hurles, M., Devlin, B., State, M., and Geschwind, D. (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.

[Magiati et al., 2007] Magiati, I., Charman, T., and Howlin, P. (2007). A two-year prospective follow-up study of community-based early intensive behavioural intervention and specialist nursery provision for children with autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, 48(8):803–812.

[Manolio et al., 2009] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461:747 EP –.

[Marchetto et al., 2016] Marchetto, M. C., Belinson, H., Tian, Y., Freitas, B. C., Fu, C., Vadodaria, K. C., Beltrao-Braga, P. C., Trujillo, C. A., Mendes, A. P. D., Padmanabhan, K., Nunez, Y., Ou, J., Ghosh, H., Wright, R., Brennand, K. J., Pierce, K., Eichenfield, L., Pramparo, T., Eyler, L. T., Barnes, C. C., Courchesne, E., Geschwind, D. H., Gage, F. H., Wynshaw-Boris, A., and Muotri, A. R. (2016). Altered proliferation and networks in neural cells derived from idiopathic autistic individuals. *Molecular Psychiatry*, 22:820 EP –.

[Mariani et al., 2015] Mariani, J., Coppola, G., Zhang, P., Abyzov, A., Provini, L., Tomasini, L., Amenduni, M., Szekely, A., Palejev, D., Wilson, M., Gerstein, M., Grigorenko, E. L., Chawarska, K., Pelphrey, K. A., Howe, J. R., and Vaccarino, F. M. (2015). Foxg1-dependent dysregulation of gaba/glutamate neuron differentiation in autism spectrum disorders. *Cell*, 162(2):375 – 390.

[MATLAB, 2017] MATLAB (2017). *version 9.3 (R2017b)*. The MathWorks Inc., Natick, Massachusetts.

[Mayes and Calboun, 1999] Mayes, S. D. and Calboun, S. L. (1999). Symptoms of autism in young children and correspondence with the dsm. *Infants & Young Children*, 12(2):90–97.

[McKinney et al., 2010] McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

[McLaren et al., 2016] McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. *Genome biology*, 17(1):122.

[Metzker, 2010] Metzker, M. L. (2010). Sequencing technologies: the next generation. *Nature Reviews Genetics*, 11(1):31.

[Mi et al., 2016] Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P. D. (2016). Panther version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1):D183–D189.

[Nair et al., 2013] Nair, A., Treiber, J. M., Shukla, D. K., Shih, P., and Müller, R.-A. (2013). Impaired thalamocortical connectivity in autism spectrum disorder: a study of functional and anatomical connectivity. *Brain*, 136(6):1942–1955.

[Newschaffer et al., 2007] Newschaffer, C. J., Croen, L. A., Daniels, J., Giarelli, E., Grether, J. K., Levy, S. E., Mandell, D. S., Miller, L. A., Pinto-Martin, J., Reaven, J., Reynolds, A. M., Rice, C. E., Schendel, D., and Windham, G. C. (2007). The epidemiology of autism spectrum disorders. *Annual Review of Public Health*, 28(1):235–258. PMID: 17367287.

[Oliphant, 2006] Oliphant, T. E. (2006). *A guide to NumPy*, volume 1. Trelgol Publishing USA.

[Parikshak et al., 2013] Parikshak, N. N., Luo, R., Zhang, A., Won, H., Lowe, J. K., Chandran, V., Horvath, S., and Geschwind, D. H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*, 155(5):1008–1021.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

[Pinto et al., 2014] Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., Vorstman, J., Thompson, A., Regan, R., Pilorge, M., Pellecchia, G., Pagnamenta, A., Oliveira, B., Marshall, C., Magalhaes, T., Lowe, J., Howe, J., Griswold, A., Gilbert, J., Duketis, E., Dombroski, B., DeJonge, M., Cuccaro, M., Crawford, E., Correia, C., Conroy, J., Conceiasao, I., Chiocchetti, A., Casey, J., Cai, G., Cabrol, C., Bolshakova, N., Bacchelli, E., Anney, R., Gallinger, S., Cotterchio, M., Casey, G., Zwaigenbaum, L., Wittemeyer, K., Wing, K., Wallace, S., vanEngeland, H., Tryfon, A., Thomson, S., Soorya, L., Roga, B., Roberts, W., Poustka, F., Mouga, S., Minshew, N., McInnes, L., McGrew, S., Lord, C., Leboyer, M., LeCouteur, A., Kolevzon, A., JimenezGonzalez, P., Jacob, S., Holt, R., Guter, S., Green, J., Green, A., Gillberg, C., Fernandez, B. A., Duque, F., Delorme, R., Dawson, G., Chaste, P., Cafa, C., Brennan, S., Bourgeron, T., Bolton, P., Balte, S., Bernier, R., Baird, G., Bailey, A., Anagnostou, E., Almeida, J., Wijsman, E., Vieland, V., Vicente, A., Schellenberg, G., Pericak-Vance, M., Paterson, A., Parr, J., Oliveira, G., Nurnberger, J., Monaco, A., Maestrini, E., Klauck, S., Hakonarson, H., Haines, J., Geschwind, D., Freitag, C., Folstein, S., Ennis, S., Coon, H., Battaglia, A., Szatmari, P., Sutcliffe, J., Hallmayer, J., Gill, M., Cook, E., Buxbaum, J., Devlin, B., Gallagher, L., Betancur, C., and Scherer, S. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *The American Journal of Human Genetics*, 94(5):677 – 694.

[Poplin et al., 2018] Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., and DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36:983.

[Rabinowicz et al., 1996] Rabinowicz, T., de Courten-Myers, G. M., Petetot, J. M.-C., Xi, G., and de los Reyes, E. (1996). Human cortex development: estimates of neuronal numbers indicate major loss late during gestation. *Journal of Neuropathology and Experimental Neurology*, 55(3):320–328.

[Reichow et al., 2012] Reichow, B., Barton, E. E., Boyd, B. A., and Hume, K. (2012). Early intensive behavioral intervention (EIBI) for young children with autism spectrum disorders (ASD). *The Cochrane Library*.

[Röschke and Başar, 1988] Röschke, J. and Başar, E. (1988). The EEG is not a simple noise: strange attractors in intracranial structures. In *Dynamics of sensory and cognitive processing by the brain*, pages 203–216. Springer.

[Samengo, I., Elijah, D., 2013] Samengo, I., Elijah, D., M. M. (2013). Spike-Train Analysis. In *Principles of Neural Coding*, pages 75–96.

[Sandin et al., 2017] Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H., and Reichenberg, A. (2017). The Heritability of Autism Spectrum DisorderReassessing the Heritability of Autism Spectrum DisordersLetters. *JAMA*, 318(12):1182–1184.

[Sanford et al., 2018] Sanford, E., Watkins, K., Nahas, S., Gottschalk, M., Coufal, N. G., Farnaes, L., Dimmock, D., Kingsmore, S. F., and RCIGM Investigators (2018). Rapid whole-genome sequencing identifies a novel AIRE variant associated with autoimmune polyendocrine syndrome type 1. *Molecular Case Studies*, 4(3):a002485.

[Schafer et al., 2019] Schafer, S. T., Paquola, A. C. M., Stern, S., Gosselin, D., Ku, M., Pena, M., Kuret, T. J. M., Liyanage, M., Mansour, A. A., Jaeger, B. N., Marchetto, M. C., Glass, C. K., Mertens, J., and Gage, F. H. (2019). Pathological priming causes developmental gene network heterochronicity in autistic subject-derived neurons. *Nature Neuroscience*, 22(2):243–255.

[Schultz et al., 2000] Schultz, R. T., Gauthier, I., Klin, A., Fulbright, R. K., Anderson, A. W., Volkmar, F., Skudlarski, P., Lacadie, C., Cohen, D. J., and Gore, J. C. (2000). Abnormal ventral temporal cortical activity during face discrimination among individuals with autism and asperger syndrome. *Archives of General Psychiatry*, 57(4):331–340.

[Sebat et al., 2007] Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.-H., Hicks, J., Spence, S. J., Lee, A. T., Puura, K., Lehtimäki, T., Ledbetter, D., Gregersen, P. K., Bregman, J., Sutcliffe, J. S., Jobanputra, V., Chung, W., Warburton, D., King, M.-C., Skuse, D., Geschwind, D. H., Gilliam, T. C., Ye, K., and Wigler, M. (2007). Strong association of de novo copy number mutations with autism. *Science*, 316(5823):445–449.

[Shen et al., 2013] Shen, M. D., Nordahl, C. W., Young, G. S., Wootton-Gorges, S. L., Lee, A., Liston, S. E., Harrington, K. R., Ozonoff, S., and Amaral, D. G. (2013). Early brain enlargement and elevated extra-axial fluid in infants who develop autism spectrum disorder. *Brain*, 136(9):2825–2835.

[Sherry et al., 2001] Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic Acids Research*, 29(1):308–311.

[Slamon et al., 1987] Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A., and McGuire, W. L. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, 235(4785):177–182.

[Sunkin et al., 2012] Sunkin, S. M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T. L., Thompson, C. L., Hawrylycz, M., and Dang, C. (2012). Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Research*, 41(D1):D996–D1008.

[Tabuchi et al., 2007] Tabuchi, K., Blundell, J., Etherton, M. R., Hammer, R. E., Liu, X., Powell, C. M., and Südhof, T. C. (2007). A neuroligin-3 mutation implicated in autism increases inhibitory synaptic transmission in mice. *Science*, 318(5847):71–76.

[Takens, 1981] Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*, pages 366–381. Springer.

[Team et al., 2013] Team, R. C. et al. (2013). R: A language and environment for statistical computing.

[The Clinical Lung Cancer Genome Project and Network Genomic Medicine, 2013] The Clinical Lung Cancer Genome Project and Network Genomic Medicine (2013). A genomics-based classification of human lung tumors. *Science Translational Medicine*, 5(209):209ra153.

[Thomas et al., 2003] Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). Panther: A library of protein families and subfamilies indexed by function. *Genome Research*, 13(9):2129–2141.

[Thomas et al., 2006] Thomas, P. D., Kejariwal, A., Guo, N., Mi, H., Campbell, M. J., Muruganujan, A., and Lazareva-Ulitsky, B. (2006). Applications for protein sequence?function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Research*, 34(2):645–650.

[Uhlén et al., 2015] Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, 347(6220).

[Vadodaria et al., 2018] Vadodaria, K. C., Amatya, D. N., Marchetto, M. C., and Gage, F. H. (2018). Modeling psychiatric disorders using patient stem cell-derived neurons: a way forward. *Genome Medicine*, 10(1):1.

[Voineagu et al., 2011] Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Horvath, S., Mill, J., Cantor, R. M., Blencowe, B. J., and Geschwind, D. H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, 474(7351):380.

[Vorstman et al., 2017] Vorstman, J. A., Parr, J. R., Moreno-De-Luca, D., Anney, R. J., Nurnberger Jr, J. I., and Hallmayer, J. F. (2017). Autism genetics: opportunities and challenges for clinical translation. *Nature Reviews Genetics*.

[Wall et al., 2012] Wall, D., Kosmicki, J., Deluca, T., Harstad, E., and Fusaro, V. (2012). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational Psychiatry*, 2(4):e100.

[Wang et al., 2005] Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F., and Mewes, H. W. (2005). Gene selection from microarray data for cancer classification: a machine learning approach. *Computational Biology and Chemistry*, 29(1):37–46.

[Weiner et al., 2017] Weiner, D. J., Wigdor, E. M., Ripke, S., Walters, R. K., Kosmicki, J. A., Grove, J., Samocha, K. E., Goldstein, J. I., Okbay, A., Bybjerg-Grauholm, J., Werge, T., Hougaard, D. M., Taylor, J., iPSYCH Broad Autism Group, Bækvad-Hansen, M., Dumont, A., Hansen, C., Hansen, T. F., Howrigan, D., Mattheisen, M., Moran, J., Mors, O., Nordentoft, M., Nørgaard-Pedersen, B., Poterba, T., Poulsen, J., Stevens, C., Group, P. G. C. A., Anttila, V., Holmans, P., Huang, H., Klei, L., Lee, P. H., Medland, S. E., Neale, B., Weiss, L. A., Zwaigenbaum, L., Yu, T. W., Wittemeyer, K., Willsey, A. J., Wijsman, E. M., Wassink, T. H., Waltes, R., Walsh, C. A., Wallace, S., Vorstman, J. A. S., Vieland, V. J., Vicente, A. M., van Engeland, H., Tsang, K., Thompson, A. P., Szatmari, P., Svantesson, O., Steinberg, S., Stefansson, K., Stefansson, H., State, M. W., Soorya, L., Silagadze, T., Scherer, S. W., Schellenberg, G. D., Sandin, S., Saemundsen, E., Rouleau, G. A., Rogé, B., Roeder, K., Roberts, W., Reichert, J., Reichenberg, A., Rehnström, K., Regan, R., Poustka, F., Poultney, C. S., Piven, J., Pinto, D., Pericak-Vance, M. A., Pejovic-Milovancevic, M., Pedersen, M. G., Pedersen, C. B., Paterson, A. D., Parr, J. R., Pagnamenta, A. T., Oliveira, G., Nurnberger, J. I., Murtha, M. T., Mouga, S., Morrow, E. M., De Luca, D. M., Monaco, A. P., Minshew, N., Merikangas, A., McMahon, W. M., McGrew, S. G., Martsenkovsky, I., Martin, D. M., Mane, S. M., Magnusson, P., Magalhaes, T., Maestrini, E., Lowe, J. K., Lord, C., Levitt, P., Martin, C. L., Ledbetter, D. H., Leboyer, M., Le Couteur, A. S., Ladd-Acosta, C., Kolevzon, A., Klauck, S. M., Jacob, S., Iliadou, B., Hultman, C. M., Hertz-Picciotto, I., Hendren, R., Hansen, C. S., Haines, J. L., Guter, S. J., Grice, D. E., Green, J. M., Green, A., Goldberg, A. P., Gillberg, C., Gilbert, J., Gallagher, L., Freitag, C. M., Fombonne, E., Folstein, S. E., Fernandez, B., Fallin, M. D., Ercan-Sencicek, A. G., Ennis, S., Duque, F., Duketis, E., Delorme, R., De Rubeis, S., De Jonge, M. V., Dawson, G., Cuccaro, M. L., Correia, C. T., Conroy, J., Conceição, I. C., Chiocchetti, A. G., Celestino-Soper, P. B. S., Casey, J., Cantor, R. M., Café, C., Brennan, S., Bourgeron, T., Bolton, P. F., Bölte, S., Bolshakova, N., Betancur, C., Bernier, R., Beaudet, A. L., Battaglia, A., Bal, V. H., Baird, G., Bailey, A. J., Bader, J. S., Bacchelli, E., Anagnostou, E., Amaral, D., Almeida, J., Buxbaum, J. D., Chakravarti, A., Cook, E. H., Coon, H., Geschwind, D. H., Gill, M., Hakonarson, H., Hallmayer, J., Palotie, A., Santangelo, S., Sutcliffe, J. S., Arking, D. E., Skuse, D., Devlin, B., Anney, R., Sanders, S. J., Bishop, S., Mortensen, P. B., Børglum, A. D., Smith, G. D., Daly, M. J., and Robinson, E. B. (2017). Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nature Genetics*, 49:978 EP –.

[Willsey et al., 2013] Willsey, A., Sanders, S., Li, M., Dong, S., Tebbenkamp, A., Muhle, R., Reilly, S., Lin, L., Fertuzinhos, S., Miller, J., Murtha, M., Bichsel, C., Niu, W., Cotney, J., Ercan-Sencicek, A., Gockley, J., Gupta, A., Han, W., He, X., Hoffman, E. J., Klei, L., Lei, J., Liu, W., Liu, L., Lu, C., Xu, X., Zhu, Y., Mane, S., Lein, E., Wei, L., Noonan, J., Roeder, K., Devlin, B., Sestan, N., and State, M. (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell*, 155(5):997 – 1007.

[Worthey et al., 2010] Worthey, E. A., Mayer, A. N., Syverson, G. D., Helbling, D., Bonacci, B. B., Decker, B., Serpe, J. M., Dasu, T., Tschannen, M. R., Veith, R. L., Basehore, M. J., Broeckel, U., Tomita-Mitchell, A., Arca, M. J., Casper, J. T., Margolis, D. A., Bick, D. P., Hessner, M. J., Routes, J. M., Verbsky, J. W., Jacob, H. J., and Dimmock, D. P. (2010). Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics In Medicine*, 13:255 EP –.

[Zachariah et al., 2017] Zachariah, S., Oommen, S., and Koshy, B. (2017). Clinical features and diagnosis of autism spectrum disorder in children. *Current Medical Issues*, 15(1):6–16.

[Zhou and Troyanskaya, 2015] Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931.