UNIVERSITY OF CALIFORNIA SAN DIEGO

Disparity between Maternal and Paternal Contributions to Inherited Risk for Autism

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

in

Biomedical Sciences

by

Danny Antaki

Committee in charge:

Professor Jonathan Sebat, Chair
Professor Joseph Gleeson
Professor Alysson Muotri
Professor Abraham Palmer
Professor Nicholas Schork

2018

The dissertation of Danny Antaki is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2018

DEDICATION

To my parents for sparking my interest in natural sciences.

EPIGRAPH

Festina lente

— Augustus

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to warmly thank my thesis advisor and mentor, Jonathan Sebat, for his invaluable guidance, support, and opportunities he provided throughout my graduate studies. I would also like to thank past and present members of the Sebat lab for their technical support, camaraderie, and stimulating discussions. I would like to especially thank William Brandler for training me and for offering great advice and friendship during my stay in the Sebat lab. Additionally, I would like to thank Madhusudan Gujral for his friendship and for his contributions to the research outlined in this work. I would also like to thank Lilia Iakoucheva and the members of her lab for their valuable feedback and contributions. A special thanks is warranted to all the people that make San Diego Supercomputer Center possible, for all the technical support and patience they exhibit when I clog up the queues. I would also like to thank my thesis committee members Joseph Gleeson, Alysson Muotri, Abraham Palmer, and Nicholas Schork for their valuable insights, advice, and guidance they have provided me throughout my years as a graduate student.

I would like to acknowledge and thank the Biomedical Sciences Graduate Program at UCSD for providing a stimulating environment and community. I would also like to thank the funding programs that have supported me throughout my tenure as a graduate student, in particular the Genetics Training Program and its director Bruce Hamilton for creating an invigorating community and for nurturing healthy skepticism.

Lastly, I wold like to thank my family and friends for their boundless support and encouragement over the years. My parents, Tamim Antaki and Mirfat Hariri deserve special notice for nurturing my scholarly interests throughout my life, and for allowing me to venture out to explore my own potential as they did 30 years ago. I would like to remember the memories of my grandfathers, who have been inspirational models for academic achievement to me. I would also like to thank my friends, Tina Wang, Matthew Duprie, Stephen Ceto, Navarre Gutierrez-Reed, Kevin Ross, Dustin Hicks, and many more for their support and companionship. Finally, I would

like to thank my wife and best friend, Genevieve, for her unyielding love to me and for her companionship and intellectual input she has given me throughout the years. I also cannot forget my cats, Alyosha and Evgeny for their love and solace they have given me.

Chapter 2 has been previously published in *Bioinformatics* (Danny Antaki, William M. Brandler, Jonathan Sebat. 2018. SV^2: accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformatics* 34(10):1774-1777). The dissertation author is the primary author of this material. William M. Brandler provided technical advice and aided in generating genotype likelihood filters. Jonathan Sebat supervised the project and provided advice.

Chapter 3 has been previously published in *Science* (William M. Brandler, Danny Antaki, Madhusudan Gujral, Morgan L. Kleiber, Joe Whitney, Michelle S. Maile, Oanh Hong, Timothy R. Chapman, Shirley Tan, Prateek Tandon, Timothy Pang, Shih C. Tang, Keith K. Vaux, Yan Yang, Eoghan Harrington, Sissel Juul, Daniel J. Turner, Bhooma Thiruvahindrapuram, Gaganjot Kaur, Zhuozhi Wang, Stephen F. Kingsmore, Joseph G. Gleeson, Denis Bisson, Boyko Kakaradov, Amalio Telenti, J Craig Venter, Roser Corominas, Claudio Toma, Bru Cormand, Isabel Rueda, Silvina Guijarro, Karen S. Messer, Caroline M. Nievergelt, Maria J. Arranz, Eric Courchesne, Karen Pierce, Alysson R. Muotri, Lilia M. Iakoucheva, Amaia Hervas, Stephen W. Scherer, Christina Corsello, and Jonathan Sebat. 2018. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* 360(6386):327-331). The dissertation author shares primary authorship with William M. Brandler and Madhusudan Gujral. William M. Brandler assisted with the conception, variant filtering and merging, and genetic analysis. Madhusudan Gujral assisted with sequence alignment and variant calling. The dissertation author was responsible for providing input on methodologies, software development, variant genotyping and filtering, and formal analysis. Full author contributions are detailed in the publication[10]. Jonathan Sebat supervised the project and provided invaluable advice.

Chapter 4, in part, is currently being perpared for submissions for publication of this

material. Danny Antaki, Madhusudan Gujral, Jonathan Sebat. Madhusudan Gujral assisted with variant calling and data processing. The disseration author is the primary investigator on this material, while Jonathan Sebat supervised the project and provided advice.

VITA

2013	B. S. in General Biology, Purdue University
2018	Ph. D. in Biomedical Sciences, University of California, San Diego

PUBLICATIONS

William M. Brandler, Danny Antaki, Madhusudan Gujral, Morgan L. Kleiber, Joe Whitney, Michelle S. Maile, Oanh Hong, Timothy R. Chapman, Shirley Tan, Prateek Tandon, Timothy Pang, Shih C. Tang, Keith K. Vaux, Yan Yang, Eoghan Harrington, Sissel Juul, Daniel J. Turner, Bhooma Thiruvahindrapuram, Gaganjot Kaur, Zhuozhi Wang, Stephen F. Kingsmore, Joseph G. Gleeson, Denis Bisson, Boyko Kakaradov, Amalio Telenti, J. Craig Venter, Roser Corominas, Claudio Toma, Bru Cormand, Isabel Rueda, Silvina Guijarro, Karen S. Messer, Caroline M. Nievergelt, Maria J. Arranz, Eric Courchesne, Karen Pierce, Alysson R. Muotri, Lilia M. Iakoucheva, Amaia Hervas, Stephen W. Scherer, Christina Corsello, Jonathan Sebat; Paternally inherited cis-regulatory structural variants are associated with autism. *Science*, 20 Apr 2018 : 327-331.

Danny Antaki, William M. Brandler, Jonathan Sebat; SV^2: accurate structural variation genotyping and de novo mutation detection from whole genomes, *Bioinformatics*, Volume 34, Issue 10, 15 May 2018, Pages 1774–1777.

William M. Brandler, Danny Antaki, Madhusudan Gujral, Amina Noor, Gabriel Rosanio, Timothy R. Chapman, Daniel J. Barrera, Guan Ning Lin, Dheeraj Malhotra, Amanda C. Watts, Lawrence C. Wong, Jasper A. Estabillo, Therese E. Gadomski, Oanh Hong, Karin V. Fuentes Fajardo, Abhishek Bhandari, Renius Owen, Michael Baughn, Jeffrey Yuan, Terry Solomon, Alexandra G. Moyzis, Michelle S. Maile, Stephan J. Sanders, Gail E. Reiner, Keith K. Vaux, Charles M. Strom, Kang Zhang, Alysson R. Muotri, Natacha Akshoomoff, Suzanne M. Leal, Karen Pierce, Eric Courchesne, Lilia M. Iakoucheva, Christina Corsello, Jonathan Sebat; Frequency and Complexity of de novo Structural Mutation in Autism. *The American Journal of Human Genetics*, Volume 98, Issue 4, 7 April 2016, Pages 667-679.

ABSTRACT OF THE DISSERTATION

Disparity between Maternal and Paternal Contributions to Inherited Risk for Autism

by

Danny Antaki

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2018

Professor Jonathan Sebat, Chair

The genetic basis of autism is known to consist of de novo and inherited loss of function mutations in haploinsufficient genes. It is thought that inherited risk primarily derives from mothers, believed to be due to an increased tolerance for risk alleles. However, the distinct contributions of each parent to inherited risk for autism has not been explored in depth. We investigated paternal and maternal contributions to autism by analyzing the transmission of private deletions in coding and cis-regulatory (CRE-SVs) regions of functionally constrained genes in whole genomes of 10,015 individuals (2650 families). We then extended our transmission distortion analysis to encompass of loss of function single nucleotide variants (SNVs) and insertion/deletion (INDELs), as well as private potentially pathogenic missense mutations. Our

goal is to untangle distinct modes of inheritance for autism risk, hypothesizing that fathers and mothers carry distinct risk contributions. We report that mothers and fathers over-transmit loss of function variants within functionally constrained coding regions. However, fathers but not mothers tended to over-transmit damaging CRE-SVs and missense variants to affected offspring. When we test the segregation of loss of function variants stratified by sex of the offspring, we find that most of the genetic risk to sons is derived from the father, which is not consistent with the previous female protective effect model. Our work demonstrates that inherited damaging variants comprise a significant component of missing heritability for autism with fathers contributing a substantial amount of risk.

Chapter 1

Introduction

Autism is a psychiatric disorder of impaired social interaction and restricted behavior. Autism is heterogeneous with a wide range of phenotype and intellectual capability. There is also a strong male bias of 3 to 4[22]. The incidence of autism has been reported to be 1 in 68 individuals[22]. However, the recent change of diagnostic criteria may alter this observation[52]. It has been established that autism has a strong genetic component. Decades of measuring concordance of the disease in monozygotic twins have determined genetic liability to fall within a range of 30-99%[64, 65, 7], suggesting both environmental and genetic factors at work. Commonly, patients with autism are diagnosed with other disorders such as epilepsy, cardiac deficiencies, and gastrointestinal maladies[4]. Taken altogether, the cause for autism is complex in nature, but the genetic contribution is significant with many loci of varying risk, penetrance, and pleiotropy contributing to the disorder.

The very first genetic locus attributed to autism was the Fragile-X locus in 1969[47]. Linkage mapping associated regions on 20 chromosomes to autism[64, 71, 77], but these studies require large pedigrees, effectively excluding sporadic cases. Later, genotyping arrays provided increased resolution and allowed for more samples to be processed; the first studies attributed submicroscopic de novo copy number variants in autism[71]. However, genome wide association

studies, which implement common variants, into autism with this technology failed to associate and replicate more than 3 confident regions[64]. In contrast to genome wide association studies, numerous risk loci for autism were associated with de novo and inherited structural variants such as the 16p11.2 deletion[25], 1q21.1 deletion[53]. Likewise, many genes localized to postsynaptic densities[50] and with functional roles such as neuronal cell adhesion[77] or ubiqutin[26] were found to be implicated in autism. Microarray and comparative genomic hybridization methods were fruitful, implicating damaging copy number variants to around 10% of cases[49]. However these methods are limited in scope; the mutant alleles are not directly sequenced. Therefore, many researchers turned to sequencing the entire exome to search for missing heritability in autism. In over 2500 families with one affected child and a sibling control, Iossifov and others found that coding de novo mutations contribute to about 30% of simplex cases and 45% of female affected offspring[33]. Since the initial genetic discovery of autism, many regions and genes have been implicated and characterized, with some studies reporting a range of 400–1000 genes that predispose risk for autism[24]. However, this success is met with the looming missing heritability problem where over 40% of cases have unexplained diagnoses with 49% of the remaining missing heritability thought to exist in common variants with additive effects[64, 23]. Thus creative approaches in leveraging newer technologies such as whole genome sequencing and associating risk from rare inherited variants should be considered.

This study largely focuses on structural variants (SV) since SVs are much larger than single nucleotide variants (SNV) and insertion/deletions (INDEL) they are more likely to elicit a functional change. Additionally SV are largely responsible as a mechanism for discontinuous evolution and speciation[32] through the means of large structural changes typically involving large chromosomal rearrangements that create reproductive barriers and the formation of new species. However, the potential to elicit a functional change can also be deleterious. This is our rationale to first search for genetic association of complex traits with structural variants. However, detection of SV in whole genome sequence data carries a higher rate of false positives than

SNVs and INDELs[11]. This is due to many factors such as repeat context; genomic regions that are repetitive are different to align sequences from short-reads uniquely. Hence these regions appear to have diminished coverage and breakpoints, suggesting structural changes. Additionally, there is not a standard set of guidelines for the analysis of SV, unlike for SNVs and INDELs[51]. With this in mind, we sought to create a genotyping algorithm for SVs detected using short read paired-end whole genomes. We implement four distinct features of SVs in paired-end libraries: coverage, discordant paired-end, split-reads, and heterozygous allele ratio. Coverage is simply the average number of reads that align to a locus; deletions will have depleted coverage while duplications with have enriched coverage. Discordant paired-ends are sequenced fragments that span a breakpoint but the breakpoint resides in the unsequenced insert. Thus the paired-ends will align in a discordant fashion (for deletions, the insert size approximates the length of the deletion). Split-reads are those with breakpoints sequenced within the read; aligners will split the read into two alignments on either side of the breakpoint. Heterozygous allele depth is used for genotyping duplications and is similar to B-allele frequency on microarrays. It's simply, the ratio of minor allele reads to major allele reads for heterozygous SNVs within the duplication. We have trained this model using a gold standard dataset from the 1000 Genomes project[76] and tested our models using two orthogonal platforms: genotyping microarrays and single molecule reads. We show that our algorithm, SV^2, performs better than probabilistic methods[5], and therefore we implemented SV^2 for genotyping all subsequent SVs in this work.

With the problem of false positive calls for SVs resolved, we then asked if there was a class of mutation that previous exome studies neglected. Due to the design of targeted exome sequencing, cis-regulatory elements such as transcription start sites, UTRs, and promoters were not assayed in these studies. To this end, we collected whole genome sequence data on a discovery cohort of 829 families and asked if there was an association of mutated cis-regulatory elements to autism. Previous studies on inherited risk to autism have implemented scores of haploinsufficiency[42] with success. Fortuitously, Exome Aggregation Consortium (ExAC)[45]

released probability scores of functional constraint for nearly every gene in the genome. These scores were calculated from observed and expected counts of loss of function (LoF) mutations in over 60,000 control exomes. The probability for likely to be intolerant (pLI) provides a simple way to score genes according to functional constraint; for these studies we used a score $\geq$0.9 recommended by ExAC[45]. Given scores of functional constraint, we then asked if deletions of cis-regulatory elements of haploinsufficient genes are associated with autism. We chose to limit ourselves to deletions since the functional impact of deletions is easily interpretable, and that the ExAC pLI scores were derived from loss of function mutations. We find in our discovery cohort that fathers but not mothers over-transmitted deletions over cis-regulatory regions (hence dubbed CRE-SVs)[10]. When we tested this finding in a replication cohort, we confirmed our initial result[10]. Our finding was noteworthy in the fact that fathers also contribute inherited risk to autism, a finding that goes into the face of accepted theory of inherited risk for autism.

The female protective effect model[86] is widely accepted among researchers. This model suggests that autism risk can be explained by two distinct genetic modes: highly penetrant de novo mutations, and maternally transmitted risk. Severely affected cases and females tend to carry large de novo SVs or LoF mutations in extremely conserved genes[12]. Large de novo mutations typically act in a dominant fashion with high penetrance, and are unlikely to be transmitted given the low fecundity of autism[62] (0.25 children for males, 0.48 children for females, relative to general population). Affected females tend to have increased burden of de novo LoF mutations compared to affected males[86, 12]. This observation implies that females have increased tolerance for risk variants and require a greater genetic load of these mutations to become affected. This implication is clear when considered the extreme male bias autism exhibits[64], suggesting that males have decreased tolerance for genetic risk. Therefore, if a moderately penetrant de novo mutation occurred in a female, she may not develop autism. However, that risk variant can be transmitted to male offspring, severely increasing the risk for autism. The female protective effect also assumes that inherited risk for autism should solely

derive from the maternal lineage, since the mothers can tolerate the mutations. In fact, studies of inherited LoF mutations in autism show a clear maternal but not paternal bias[42]. However, our previous study has found that CRE-SVs from fathers not mothers are associated with autism[10]. Therefore, it is not outside the realm of possibility that autism risk in the form of rare inherited mutations can derive from the paternal lineage. However, such a hypothesis suggests either one or a combination of two considerations given the female protective effect: (1) fathers that carry risk variants are slightly affected and that risk acts in an additive fashion, (2) if fathers are not affected by any degree, then imprinting might offer a solution as to how these fathers can carry a mutation without become affected. In all, autism is a very diverse disorder phenotypically and genetically. Investigations into the burden of de novo mutations have been successful, attributing about 30% of cases to either LoF or missense de novo mutations[11, 33]. However, inquiry into the inherited risk of autism is largely unexplored and has the potential to explain a large component of missing heritability for autism.

Chapter 2

in silico Genotyping of Structural Variation with Machine Learning

2.1 Abstract

Structural Variation (SV) are more likely to cause functional change compared to Single Nucleotide Variants (SNV) and Insertion/Deletions (INDEL), making SV an alluring class of mutation for attributing risk to complex disorders. However, detection of SV in short-read paired-end Illumina libraries carries a high burden of error. Sensitive SV detection is further complicated with the need for multiple calling algorithms, making integration of the genotypes and likelihoods challenging. Machine learning genotyping of SNVs has been successful in the past, achieving false discovery rates (FDR) well below other probabilistic models [55]. Levering common germline SV from the 1000 Genomes Project (1KGP) [76] and high coverage (COV) paired-end whole genomes (27 individuals), we developed a machine learning software tool dubbed, SV2 (Support Vector Structural Variant genotyper) in order to genotype deletions and duplications. We test the validity of our models with orthogonal data sets, using both genotyping arrays (N=57) and single molecule long reads (N=9), resulting in FDRs of 0.85% for deletions and

0% for duplications. We then compare SV^2 to two other probabilistic models for SV genotyping, SVTyper[17] and Manta[16], and found that SV^2 outperformed both models with areas under the receiver operating characteristic curves of 0.92 for deletions and 0.8 for duplications.

2.2 Introduction

Structural Variation (SV) is a change of the structure of a chromosome larger than 50bp. SV is a major contributor to human genetic variation with 13% of the human genome defined as structurally variable [76]. SV is also implicated in a variety of human diseases including cancer, heart disease, and psychiatric disorders [18, 85, 11]. In many congenital and sporadic disorders such as idiopathic autism and intellectual disability, de novo germline SVs, those that are novel in offspring (i.e. not found in parents), are known to contribute risk [12]. Due to the fact SV comprise more genomic real estate than SNVs or INDELs, SV are more likely to elicit a functional change [76]. Hence, for many complex disorders with substantial missing heritability, SV has become an attractive class of mutation to attribute risk. However, SV detection for paired-end Illumina sequencing libraries carries a high burden of false positives [5]. This burden of false positives complicates pedigree analysis of inherited and de novo variants. False positives and false negatives distort transmission rates and complicate interpretation of de novo mutations.

Given the diversity of structural variation and their wide range of sizes (50bp to 50Mb) [76], typically multiple algorithms are required for comprehensive variant calling. SV calling algorithms each operate as a standalone solution relying on either read depth [54, 1] or discordant paired-ends and split-reads [44, 16]. Each of these features have distinct properties from each other in paired-end data. Read depth is associated with copy number, but GC context and repetitive regions can skew coverage [1]. Discordant paired-ends, sequencing fragments with insert sizes outside the expected distribution, and split-reads, reads that span breakpoints that have two non-overlapping alignments on opposite ends of the SV, are also correlated with copy

number but are limited to SVs without repetitive breakpoints. Such SVs lack breakpoint features (discordant paired-ends and split-reads) because the reads spanning the repeats are too short to be confidently aligned to a unique locus. Thus, other features like coverage and heterozygous allele ratio (the ratio of coverage of the reference and derived alleles) would need to be used to characterize these SVs. Given the wide characteristics of SVs and the singular approach current methods apply, no solution exists that can integrate different variant calls and features of structural changes in paired-end reads in one step. Hence, we present SV^2 (support-vector structural-variant genotyper), a turn-key solution for unifying SV predictions into an integrated set of genotypes and likelihoods. SV^2 (https://www.github.com/dantaki/SV2) is an open source software written in Python that exploits read depth, discordant paired-ends, and split-reads in a supervised support vector machine classifier.

2.3 Methods

2.3.1 Software Availability

The most recent version of SV^2 source code and documentation is hosted on GitHub (https://github.com/dantaki/SV2)

2.3.2 SV^2 Workflow

SV^2 is a high-throughput SV genotyper that requires BAM alignments with chimeric reads labeled with SA tags, a bgzipped and tabix indexed SNV VCF with allele depth, and a BED or VCF file of deletion and duplication positions to be genotyped (Figure 2.1). SV^2 first performs a preprocessing step that records basic statistics of each chromosome such as median coverage, insert size, and read length. Then SV^2 operates on each variant extracting informative features for genotyping with six support vector machine classifiers as described below:

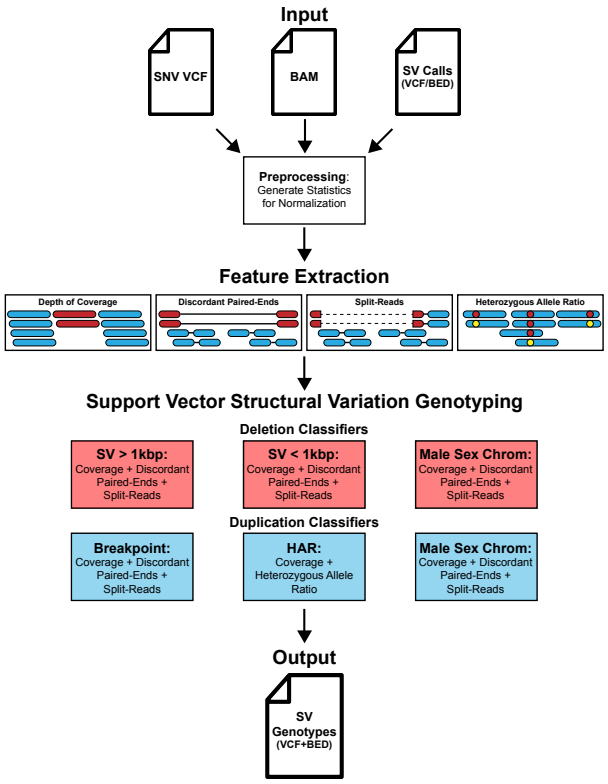

Figure 2.1: **SV^2 Workflow.** SV^2 requires a VCF file of SNVs, a BAM file, and a set of SVs to genotype as input. Before genotyping, preprocessing is performed where the median coverage, insert size, and read length is recorded for feature normalization. Features for genotyping, which include depth of coverage, discordant paired-ends, split-reads, and heterozygous allele ratio (HAR), are measured for each SV. SVs are then genotyped with an ensemble of support vector machine classifiers. SV^2 produces two output files, a BED file and a VCF, containing annotations for RefSeq genic elements, RepeatMasker repeats, Segmental Duplications, Short Tandem Repeats, and common SVs from the 1000 Genomes phase 3 call set.

1. Deletion SV $>$ 1000bp classifier: genotypes putative deletions on autosomes, sex chromosomes in females, and pseudoautosomal regions on sex chromosomes. The features implemented in this classifier are depth of coverage estimated via read count, discordant paired-ends, and split-reads. This model classifies with three states: homozygous reference, heterozygous, and homozygous alternate corresponding to copy numbers 2, 1, and 0 respectively.

2. Deletion SV $\leq$ 1000bp classifier: genotypes variants on autosomes, sex chromosomes in females, and pseudoautosomal regions on sex chromosomes. Features for classification

include depth of coverage estimated as the median per-base pair coverage, discordant paired-ends, and split-reads. Similar to the deletion $>$ 1000bp classifier, this model genotypes with three genotype states.

3. Deletion Male Sex Chromosomes classifier: genotypes variants on male sex chromosomes and includes depth of coverage, discordant paired-ends, and split-reads as features. This model genotypes with two states: reference and alternate representing copy number 1 and 0.

4. Duplication Breakpoint classifier: genotypes variants on autosomes, sex chromosomes in females, and pseudoautosomal regions on sex chromosomes. Features include depth of coverage, discordant paired-ends, and split-reads. This model genotypes with three genotype states: homozygous reference (copy number 2), heterozygous (copy number 3), and homozygous alternate (copy number 4).

5. Duplication SNV classifier: genotypes variants on autosomes, sex chromosomes in females, and pseduoautosomal regions on sex chromosomes. Features for this classifier are depth of coverage and heterozygous allele ratio taken from SNV calls. Like the duplication breakpoint model, the SNV classifier emits three genotypes states.

6. Duplication: Male Sex Chromosomes classifier: genotypes variants on male sex chromosomes and includes depth of coverage, discordant paired-ends, and split-reads. This model genotypes with two genotypes states: reference (copy number 1) and alternate (copy number 2).

The output of SV^2 includes a BED file and a VCF file. The VCF output file contains both standard and stringent filters for de novo mutation discovery. Additionally, each variant is annotated for genes, repeatMasker elements, and 1000 Genomes phase 3 deletions and duplications.

2.3.3 Machine Learning Features of SV^2

We sought to leverage SV genotyping with four orthogonal features: depth of coverage, discordant paired-ends, split reads, and heterozygous allele ratio (HAR). Coverage was defined as either the number of reads spanning a locus or as the median base-pair depth for lengths $\leq$ 1kbp. Reads were excluded if they aligned within our genome mask comprising of segmental duplications, short tandem repeats, assembly gaps, telomeres, and centromeres. Raw coverage values were normalized according the chromosome average, and then adjusted based on GC content with respect to PCR or PCR-free chemistries, adapted from CNVnator [1]. We defined discordant paired-ends to have insert sizes greater than the chromosome median plus five times the median absolute deviation. To reduce noise, we limited the search for discordant paired-ends and split-reads to +/-500bp of the start and end positions of the SV. Likewise, only discordant paired-ends and split-reads were included if the mate-pair or the supplementary alignment mapped to the opposite side of the breakpoint. The resulting number of discordant paired-ends and split-reads was then normalized to the number of concordant reads within the locus. Akin to B-allele frequency on SNV microarrays, HAR was defined as the median ratio of coverage of the minor allele to the major allele for all heterozygous variants encompassing the SV.

2.3.4 SV^2 Training Set

Features were obtained from 27 PCR-free high coverage whole genomes (48x, 250bp read length) and 2,494 low coverage whole genomes (7x, 100bp read length) provided by 1KGP [76]. SV positions were obtained from the 1KGP phase three structural variation call set[76], retaining alleles with at least one alternate variant in the cohort. Training features are shown in Figure 2.2 and a tabulated summary of the features and number of examples used in training is described in Supplementary Table S2 in the publication[5]. Due to the larger number of low coverage samples used in the breakpoint duplication classifier, we randomly selected 100,000

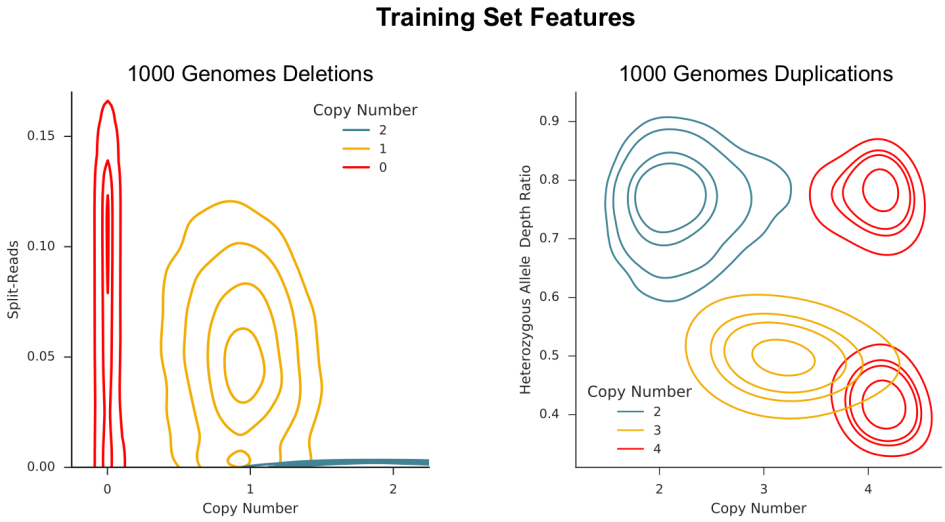

Figure 2.2: **SV^2 training set of 1000 Genomes phase 3 SVs.**Deletions less than 1000bp (N=65,808) (left) and duplications with heterozygous allele depth features (N=8,772) (right) in 27 high coverage samples. Contour lines are derived via Gaussian kernel density estimation and colors are representative of gold standard genotype. Copy number on the X axis is a function of depth of coverage. The implementation of a radial basis function kernel for classification is able to distinguish classes among nonlinear distributions.

homozygous reference examples for the final training set. In assembling the sex chromosome training data, we discovered unresolvable errors in how deletion genotypes were encoded for males. Therefore, the training set (for classifier 3. above) was replaced with an earlier release containing unphased genotypes from the 1KGP (T. Rausch personal communication). Features for every classifier were then excluded if the estimated copy number was greater than 10. Sample weights were generated to control for noise via an inverse distance weighting scheme (Equation 2.1). The inverse distance weights were calculated with respect to expected coverage of the phase 3 genotype. We defined the expected normalized coverage for homozygous reference was 1.0. The remaining expected coverages either added or subtracted 0.5 from 1.0 according to the number of copies gained or lost. Training samples for the SNV duplication classifier were weighted according to the inverse Euclidian distance of expected coverage and mean HAR value of each copy number class (Equation 2.2).

$$w_i = \frac{1}{|x - coverage_i| + 0.01} \tag{2.1}$$

Equation 2.1: Inverse distance weights for deletion classifiers and duplication breakpoint and male Sex chromosome classifiers. Weights for each training sample (w_i) are derived from the inverse distance of expected coverage (x) to the normalized coverage of the training sample ($coverage_i$). Expected coverage was defined as 1.0 for homozygous reference (copy number 2 for autosomes), and 0.5 was added or subtracted for each copy number gain or loss respectively. A coefficient of 0.01 was added to the absolute difference between expected and sample coverage to ensure the value was not zero.

$$w_i = \frac{1}{\sqrt{(x - coverage_i)^2 + (y - HAR_i)^2} + 0.01} \tag{2.2}$$

Equation 2.2: Inverse distance weights for duplication SNV classifier. Weights for each training sample (w_i) are derived from the inverse Euclidean distance of expected coverage (x) to the normalized coverage of the training sample ($coverage_i$) and to the expected heterozygous allele depth (y) to the normalized heterozygous allele ratio of the sample (HAR_i). As explained in 2.1, expected coverage was defined to be 1.0 for homozygous reference genotypes. The expected normalized heterozygous allele ratio (HAR) ratio for homozygous reference genotypes is 1.0 (equal proportions of the major and minor allele). For three copies, the expected ratio is 0.5, and four copies can either be 1.0 or depending on the number of copies present on each haplotype. A coefficient of 0.01 was added to the Euclidean distance to ensure the difference was not zero.

2.3.5 SV^2 Classifier Parameter Selection

SV^2 genotypes SV with a support vector machine model with a radial basis function kernel from scikit-learn [59]. Support vector machine classifiers are governed by the parameters C and gamma, which represent the error of classification and the influence of training samples

respectively. Parameter sweeps of varying C and gamma values were performed with balanced class weights, with the exception of the paired-end duplication classifier which used heuristic class weights. Parameters were chosen by optimizing false discovery rate with SVToolkit in a previously published cohort [11].

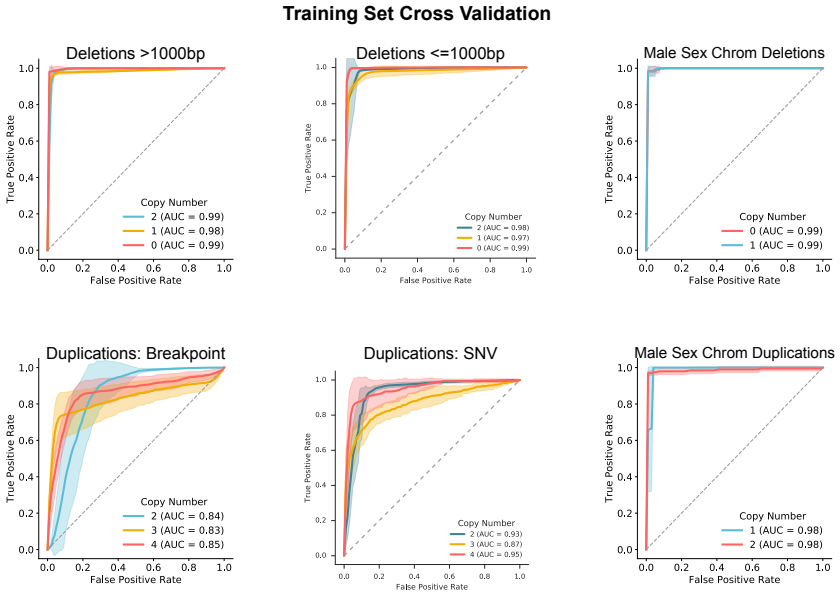

Figure 2.3: **Training Set Cross Validation Performance.** Genotyping accuracy was estimated through cross-validation of the 1000 Genomes training sets plotting the average ROC curve of 7-folds with shaded areas representing the 95% confidence interval. The number of training examples used in each classifier are tabulated in Supplementary Table S2 in the publication[5]. The average AUC across all classifiers and copy numbers for deletions was 0.98 and for duplications 0.88.

2.3.6 Genotyping Accuracy with Cross Validation

We assessed genotyping accuracy through seven-fold cross validation of the training sets, where each fold maintained the proportion of copy number classes in the full training set. Using the 1KGP phase 3 SV genotypes as truth, the mean ROC and area under the curve was determined for each genotype class (Figure 2.3).

2.3.7 SV Genotyping Performance with SNV Arrays

We evaluated false discovery rates at varying genotype likelihood cutoffs using Illumina 2.5M SNV microarrays and SV calls from high coverage, paired-end whole genomes were obtained from 57 samples described previously[11] LUMPY[44] and Manta[16] were used to call SV in high coverage alignments of the 57 samples. Genotypes and likelihoods were produced with SV^2 and the resulting variants were merged according to 50% reciprocal overlap, removing any call that overlapped 50% of its length to regions in the genome mask. False discovery rates were obtained for the resulting call set using the IRS test from SVToolkit.

2.3.8 SV Genotyping Performance with PacBio Single Molecule Real-Time Sequencing

We chose 9 individuals sequenced using PacBio Single Molecule Real-Time from the 1KGP: Human Genome Structural Variation Consortium (HGSVC) [14]. Raw reads (mean length = 8,345.2bp) were aligned to GRCh38 with bwa mem with the –x pacbio option. We then restricted our analysis to chromosome 1 to comply with HGSVC data release policy for these samples. SV calls from LUMPY and Manta were genotyped with SV^2 using complementary Illumina paired-end whole genomes sequenced to deep depths (74.2X,125bp reads). The resulting SV calls were merged according to 50% reciprocal overlap and filtered if either start or end positions overlapped to to elements in the genome mask. Then we defined supporting reads as PacBio split-reads with breakpoints that reciprocally overlap 50% to SVs genotyped in the paired-end alignments. Additionally, PacBio reads with a CIGAR string insertion or deletion with positions that reciprocally overlap 50% to the paired-end SV call were considered supporting. We omitted loci if the coverage of PacBio reads over a 1000bp flanking span of either the start or end position was less than 3x. False positives were defined as ALT genotypes without supporting PacBio spilt-reads, while true positives required 1 supporting read.

2.3.9 SV Genotyping Performance Leveraging Inheritance

We measured transmission bias using whole genome sequencing (42.6X) from 630 children and parents totaling 1884 individuals. 1,551 of the samples were obtained from the Simons Simplex Collection. SVs were called using ForestSV, LUMPY, and Manta, and then genotyped by SV^2. SVs were merged if the reciprocal overlap was greater than 50% and removed if the overlap to regions in the genome mask was greater than 50%. We measured rates of SV transmission with the group-wise transmission disequilibrium test (gTDT)[15] at varying levels of SV^2 ALT genotype likelihoods.

2.3.10 Construction of Standard and De Novo Mutation Filters

Genotype likelihood filters were determined using the IRS test from SVToolkit on Illumina 2.5M SNV arrays for 57 samples. For stringent de novo filters, we leveraged variants previously validated by PCR and Sanger sequencing[11] as a guide in determining appropriate filters. Additionally, we created a set of conditions that consider feature availability and the length of the SV to determine appropriate cutoffs, which can be found in Supplementary Table S1.

2.3.11 Comparison of Genotype Performance

We compared SV^2 genotyping performance to two other methods: SVTyper and Manta. SVs $>$1000bp were called in 57 samples using LUMPY and Manta. Genotypes and corresponding likelihood scores for LUMPY calls were provided by SVTyper, since both methods are packaged into SpeedSeq[17]. Manta supplies genotypes and confidence scores for variants it predicts. SV^2 genotyped and scored the union of the two call sets. Variants were omitted if either breakpoint intersected elements in the genome mask. Likewise, we limited this analysis to rare variants defined as those with less than 1% allele frequency in parents. We then generated receiver operating characteristic (ROC) curves for the three genotyping methods with a truth set of

genotypes determined by SVToolkit IRS test on Illumina 2.5M arrays.

For SVs $\leq$1000bp, we evaluated genotyping performance for the three methods above with PacBio long-reads. For 9 individuals, SVs were called in complementary deep coverage Illumina WGS (74x) using LUMPY and Manta with SV^2 genotyping the union of SV calls. Variants were omitted if either start or end position intersected with elements in the genome mask. Additionally, we restricted the analysis to variants with a median coverage greater than or equal to 3X of 1000bp flanking regions. Supporting PacBio reads contained either at least one split-read or CIGAR string insertion/deletion with positions that reciprocally overlap at least 50% to the SV in question.

2.4 Results

SV^2 (support-vector structural-variant genotyper) is an open source application written in Python that requires a BAM file, a single nucleotide variant (SNV) VCF file, and either a BED or VCF file of deletions and duplications as input. SV^2 operates in three stages: preprocessing, feature extraction, followed by genotyping. Genotyping utilizes four informative metrics: read depth, discordant paired-ends, split-reads, and heterozygous allele depth in a supervised support vector machine classifier trained on whole genome sequences (WGS) from the 1000 Genomes Project (1KGP). The output VCF file contains genotypes and annotations for genes, repeats, and variant identifiers from the 1KGP (Figure 2.1).

2.4.1 Training Set Cross Validation

We initially sought determination of SV^2's genotyping performance with cross-validation. We calculated the mean receiver operating characteristic (ROC) curve of 7 folds, maintaining the proportion of classes in the full training set. We found the average area under the curve (AUC) for deletions as 0.98 and for tandem duplications as 0.88. ROC curves for the remaining classifiers

produced similar AUCs (Figure 2.2).

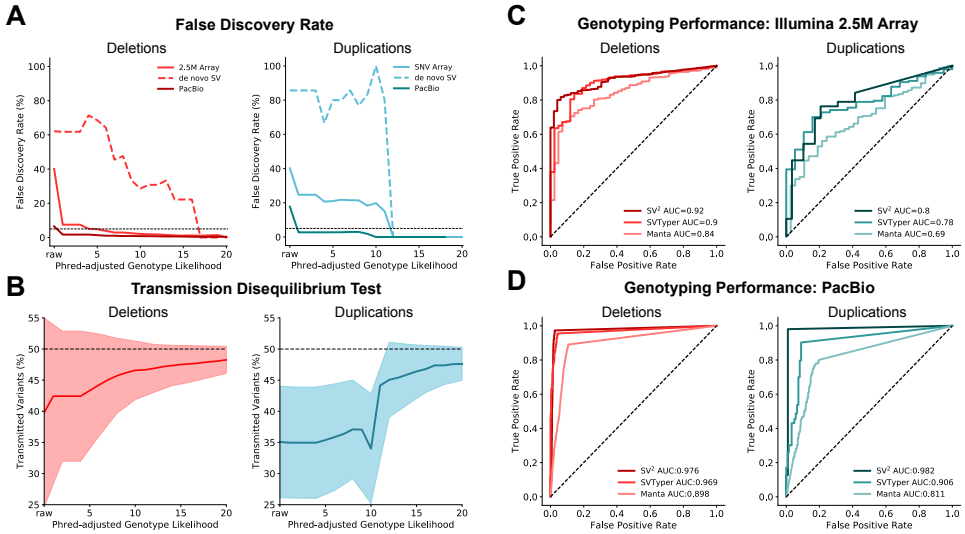

Figure 2.4: **SV^2 genotyping performance.** (A) False discovery rate across SV^2 genotype likelihoods estimated from Illumina 2.5M arrays (N=57) and PacBio long reads (N=9). Black dotted line indicates 5% FDR. (B) Group-wise transmission disequilibrium tests across SV^2 genotype likelihoods in 630 offspring with shaded regions representing one standard deviation. (C) Receiver operating characteristic (ROC) curves of WGS genotyping calculated from Illumina 2.5M arrays for SV^2, SVTyper, and Manta in 57 individuals. (D) ROC curves of WGS genotyping calculated from supporting PacBio SMRT reads for SV^2, SVTyper, and Manta for SVs in 9 individuals.

2.4.2 False Discovery Rates and Filters

In order to measure the performance of our genotypers, we applied orthogonal assays of structural variant detection such as SNV genotyping arrays and single molecule long reads. False discovery rates (FDR) of SV^2 genotypes from WGS in 57 subjects[11] were determined using Illumina 2.5M SNV arrays with SVToolkit(sourceforge.net/projects/svtoolkit/) (Figure 2.4A). SVToolkit validates variants by performing a rank-sum test across the cohort for each variant. Since this method uses the input cohort as a population mean, polymorphic CNVs are not reliably validated. Therefore, we only assessed the validity of rare variants at or below an allele frequency of 1%. FDR was 40% for unfiltered deletions (N=5,344) and duplications (N=776). Next, we formulated genotype likelihood filters at a "standard" level of stringency and a higher ("de novo")

stringency for de novo variant discovery (Supplementary Table S1). At the standard level of stringency, the FDR was 1.24% for deletions and 4.41% for duplications (Figure 2.5). Likewise, unfiltered de novo variants carry a FDR of 60% for deletions and 86% for duplications with strict de novo filters reducing the FDR to 0.54% for deletions and 0% for duplications (Figure 2.5).

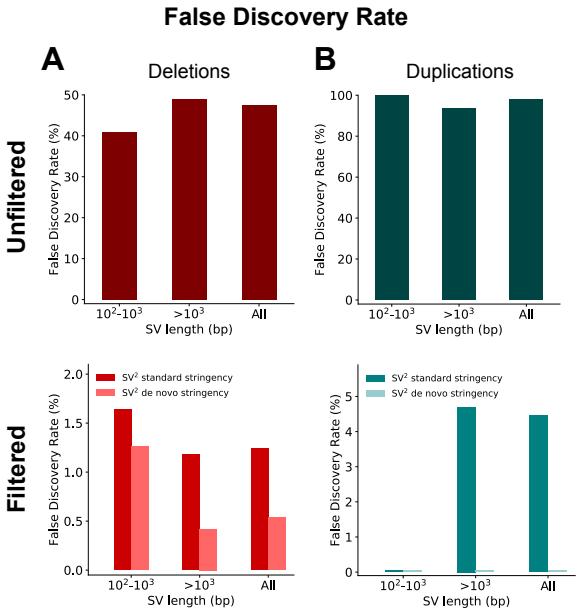

Figure 2.5: **False discovery rates estimated on SNV arrays in 57 samples of unfiltered (top) and SV^2 filtered (bottom).** (A) deletions at varying size bins (100-1000bp, >1000bp) with variants were called by ForestSV, LUMPY, and Manta and genotyped by SV^2. Unfiltered variants had high rates of false positives at all size bins. However, SV^2 filtering controlled for false positives resulting in a 1.24% FDR with standard filters and 0.54% FDR with stringent filters for de novo mutation discovery. (B) FDR of duplications estimated on SNV arrays in the same samples in panel A where a FDR of 4.41% and 0% were recorded for standard filters and de novo filters respectively.

2.4.3 Transmission Disequilibrium Test

False positives and false negatives distort the expected transmission rate (50%) of inherited variants. Thus measuring the transmission rate of variants can provide a means to test genotype accuracy. We applied a group-wise transmission disequilibrium test[15], which tests the transmission of a group of variants instead of each variant separately. This method is robust

since it increases the statistical power by lowering the number of statistical tests performed. Assuming each variant is independent from each other, transmission rates follow a binomial distribution. Therefore, we can calculate the probability that a group of variants are within the expected distribution of transmission.

We tested the performance of SV^2 using the group-wise transmission disequilibrium test on SVs called in 630 children and their parents, totaling 1884 individuals. As expected, unfiltered SV calls exhibited a significant under-transmission bias: transmission rates of 39.8% for deletions and 35.08% for duplications (deletions: P=9.61x10-51, N=105,023; duplications P=7.8x10-18, N=346,173) (Figure 2.4B). Applying standard genotype likelihood filters reduced the transmission bias to 48.2% (P=1.32x10-2, N=40,587) for deletions and 47.3% (P=3.39x10-3, N=3,863) for duplications. Applying more stringent de novo filters further reduced under-transmission bias to 49.1% (P=1.32x10-2, N=21,772) for deletions and 49.3% (P=1.0, N=2,847) for duplications (Figure 2.6).

2.4.4 Validation Cohort

PacBio long-read WGS provided additional SV^2 genotyping validation. Single molecule long reads provide an excellent source for validation since longer reads allow for better alignments to repeats. Thus, breakpoints in repetitive regions that are near impossible for short read methods to align, become resolvable in long read libraries. SVs can be gleaned from long read alignments from either split-reads or CIGAR strings. Further assessment of SV^2's genotype likelihood filters leveraged PacBio long-read WGS (x26) on 9 subjects from the 1KGP[14]. This validation set is independent from the training set since SV^2 genotypes were generated using a separate deep (x72) Illumina WGS library with SV predictions from LUMPY and Manta, both of which were not implemented in the training call set[76]. To comply with the data release requirements for these data, only variants on chromosome 1 were analyzed. As a precaution for overfitting, we excluded SVs that overlapped with $\geq$80% reciprocal overlap to SVs in our training set. Additionally,

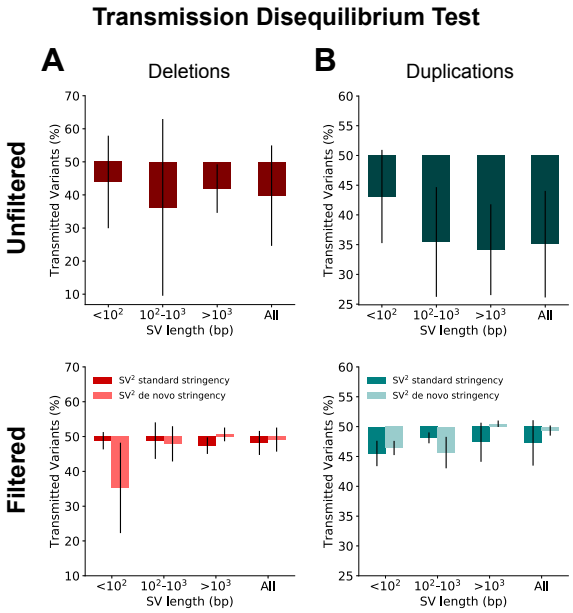

Figure 2.6: **Rate of transmitted variants in 630 children** SVs called by ForestSV, LUMPY, and Manta with error bars as one standard deviation. (A) Unfiltered deletions (top) were biased towards under-transmission with an average of 39.8% variants transmitted to children (P=9.61x10-51, N=105,023). SV^2 standard filters (bottom) resulted in 48.2% of variants transmitted (P=1.32x10-2, N=40,587). Stringent de novo filters produced a transmission rate of 49.1% (P=1.32x10-2, N=21,772). Similarly, duplications (B) had high rates of under-transmission for unfiltered variants (top) with a transmission rate of 35.1% (P=7.8x10-18, N=346,173). Filtered variants (bottom) had a duplication transmission rate of 47.3% (P=3.39x10-3, N=3,863) with a standard filter and 49.3% (P=1.0, N=2,847) for stringent filters.

we omitted variants with less than 3 PacBio reads within 1000bp flanking regions. Valid WGS genotypes required at least one supporting breakpoint with 50% reciprocal overlap to a PacBio split-read or CIGAR string. The FDR was 6.53% (N=3,121) and 17.72% (N=413) for unfiltered deletions and duplications respectively (Figure 2.4A). SV^2 standard filters, lowered the FDR for deletions to 0.85% (de novo filters: 0.62%) and for duplications to 0% (de novo filters: 0%).

2.4.5 Comparison of SV^2 to Other Models

Performance of SV^2 was then compared to that of two widely used SV genotyping software SVTyper[17] and Manta[16]. For this comparison, SVTyper genotyped SV predictions using the companion tool LUMPY[44], Manta produced genotypes for its predictions, and SV^2

genotyped the union of LUMPY and Manta calls for the previous evaluation set of 57 subjects with Illumina 2.5M arrays. Receiver operating characteristic (ROC) curves for each genotyping method were generated, specifying true and false positives with SVToolkit. SV^2 achieved the best genotyping accuracy with an AUC of 0.92 for deletions and 0.8 for duplications, in contrast to Manta (deletion AUC=0.84, duplication AUC=0.69) and SVTyper (deletion AUC=0.9, duplications AUC=0.78) (Figure 2.4C).

We then compared SV^2 genotyping performance using the 9 PacBio long read libraries. Likewise, we found that SV^2 produced the optimal performance with AUCs of 0.98 for deletions and duplications. Conversely, Manta performance resulted in an AUC of 0.9 for deletions and 0.81 for duplications, and SVTyper producing AUCs of 0.97 for deletions and 0.91 for duplications (Figure 2.4D).

2.5 Discussion

SV^2 compared to other SV genotyping software is noteworthy because of its exploitation of machine learning to reliably genotype and score deletion and tandem duplication predictions without compromising sensitivity. One of SV^2's advantages to comparable SV genotyping solutions is the ability to genotype breakpoints overlapping repetitive elements using read depth. Additionally, SV^2's incorporation of heterozygous allelic depth is better able to genotype tandem duplications, which are more prone to false positive genotypes due to fluctuations in read depth. However, relying on the presence of SNVs limits more accurate genotyping to events larger than 3kb. A caveat of SV^2 is that it cannot assign a copy number greater than 4, but this can be addressed with the addition of more gold standard examples. Ultimately, SV^2's strength is harmonizing genotypes and likelihoods from multiple callers and genotypers, simplifying analysis of SV and providing a well needed tool for accurately resolving de novo mutations.

SV^2 is streamlined compared to other machine learning classifiers, since it relies on

three features, at most, for genotyping. Additional features that would be informative for SV genotyping can be applied for future machine learning classifiers. Implementing a Random Forest model on a wide array of features could result in better performance for SV genotyping. Such a model would leverage additional features such as GC content, overlap to repeats, confidence intervals of the breakpoints, strand orientation of paired-ends, mapping quality, a variance metric for coverage, loss of heterozygosity, and clipped reads. Building a training set for a such a model can include the current high coverage dataset from the 1000 Genomes Project, used for SV^2.

Since the publication of SV^2, novel methods for genotyping variants such as Google's Deep Variant[61], which uses a convoluted neural network and images of reads to genotype SNVs and INDELs. Training a neural network on images of SVs is possible, and I have developed prototypes (github.com/dantaki/SVanGogh) that pixelate alignments, coloring them by features such as strand orientation and mapping quality. However, producing such a model is currently limited by a lack of training examples. Deep learning achieves its famed performance by training on hundreds of thousands to millions of examples. Simulating SVs is a viable approach but would lack certain nuances real data has that can't be simulated. In either case, when resources are available, training a neural network on images of SVs would be ideal for two main reasons: the model can extrapolate the structure of complex rearrangements and images can be generated by any sequencing platform. Convoluted neural networks can combine predictions when given input of two or more classes. For example, a model trained to predict if an image contains a person or a horse will report both classes when given an image of a person riding a horse. This powerful characteristic of convoluted neural networks suggests that a classifier trained on images of simple SVs (deletions, duplications, inversions, and insertions), would be able to resolve complex SVs such as a duplication-inversion-duplication. The prototype I have constructed creates images that are matrices of pixels, where each row corresponds to a sequencing fragment and each column a base pair. The image is a composite of two images, taken from the start and end breakpoints. Differences in strand orientation within sequencing reads are indicated by

one of the RGB channels. By pixelating reads in this fashion, SVs from both paired-end and single molecule reads can be displayed, meaning that a finished genotyper can be applied to any sequencing platform. At this time, there is a need for a tool that can predict and genotype SVs with precision adequate for clinical diagnostics and that can be applied to many different platforms. Genotyping SVs with images, seems to be a promising avenue for such a model.

Chapter 2 has been previously published in *Bioinformatics* (Danny Antaki, William M. Brandler, Jonathan Sebat. 2018. SV^2: accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformatics* 34(10):1774-1777). The dissertation author is the primary author of this material. William M. Brandler provided technical advice and aided in generating genotype likelihood filters. Jonathan Sebat supervised the project and provided advice.

Chapter 3

Paternally Inherited Cis-Regulatory Deletions Confer Risk to Autism

3.1 Abstract

Previous studies on the genetic basis of autism have associated de novo variants and maternally derived loss of function variants to the disorder. Currently about 15-25% of cases can be explained by a causal variant, leaving the majority of diagnoses unexplained. We hypothesize that a portion of missing heritability can be explained by rare inherited structural variants in cis-regulatory elements (CRE-SVs) of haploinsufficient genes. Our rationale for this rests with the fact that previous studies utilized targeted exome screens, thus omitting cis-regulatory elements. We investigated this prospect by testing transmission distortion of CRE-SVs in whole genomes of 3169 individuals from 829 families affected by autism. We report that structural variants were depleted within promoters and untranslated regions, similar to functionally constrained coding regions. Additionally, paternally inherited CRE-SVs were preferentially transmitted to affected children and not to their unaffected siblings. We then replicated this initial discovery in an independent cohort of 3016 families, where the association of paternal CRE-SVs, but

not maternal CRE-SVs, was observed. Taken altogether, our results suggest that rare inherited noncoding variants in dosage sensitive regions of the genome carry risk for complex disorders. Likewise for autism, there exists a striking paternal origin effect which further implies that genetic risk for the disease is more complex than previously thought with varying parent of origin effects.

3.2 Introduction

It has been established that autism has a strong genetic component. Decades of measuring concordance of the disease in monozygotic twins have determined a range of 30-99% in the heritability of autism[64]. Large copy number variants have been associated to around 10% of cases[49]. However, these studies relied on methods that do not directly sequence the mutant allele. In constract, exome sequencing of large cohorts have implicated that coding de novo mutations contribute to about 22% of diagnoses can be attributed to a loss of function or de novo missense mutation[33]. However, this success is met with the looming missing heritability problem where over 40% of cases have unexplained diagnoses with 49% of the remaining missing heritability thought to exist in common variants with additive effects[64, 23]. Thus creative approaches in leveraging newer technologies such as whole genome sequencing and associating risk from rare inherited variants should be considered.

Whole genome sequencing offers many prospects, such as being able to call many different classes of mutation such as single nucleotide variants (SNVs), insertion/deletions (INDELs), and structural variants (SVs). Likewise, sequencing the entire genome allows for interrogation of noncoding functional elements that exome sequencing omitted, such as untranslated regions of genes (UTRs), promoters, and enhancers. Microarrays can provide whole genome inquiry but at the cost of lower resolution, thus SVs are limited to $>$20kbp in length. Therefore a SV that alters the function of only cis-regulatory elements must be small enough to not affect any coding regions. Whole genome sequencing provides the means to interrogate variants that affect

cis-regulatory elements, but at the price that most variants are noncoding, making risk attribution difficult.

Finding noncoding variants that are likely to confer genetic risk to a disease like autism can be aided with a set of haploinsufficient genes. Such genes are functionally constrained by dosage, meaning that a loss of one copy would be detrimental to fitness and overall health. Empirical measures of gene haploinsufficiency has been carried out by the Exome Aggregation Consortium (ExAC)[45]. Lek and collaborators measured the frequency of loss of function mutations, such as a gain of a stop codon or a frame shift mutation, in the exomes of over 60,000 individuals; given the expected number of loss of function mutations and the observed count, a probability of loss of function (pLI) score was assigned to each gene[45]. Authors report that genes with pLI scores greater than 0.9 are likely to confer risk to disease[45]. However, using a set of genes slightly limits inquiries into noncoding association to variants that are genic, such as overlapping UTRs, promoters, or transcription start sites. Even with this limitation, we sought to ask if missing heritability in autism can be attributed to cis-regulatory deletions.

3.3 Methods

3.3.1 Study Design

Our discovery data set consisted of families with at least one diagnosed case of autism, which was derived from two data sets: the Relating genes to Adolescent and Child Health (REACH) study and the Simons Simplex Collection Phase 1 (SSC1) study.

The REACH cohort consists of 309 families comprising of 1095 individuals. In these families there were 112 control offspring and 362 affected cases with diagnoses ranging from autism (285 offspring), pervasive developmental disorder - not otherwise specified (PDD-NOS, 10 offspring), 10 with attention deficit hyperactivity disorder, and 24 with speech delay, epilepsy, anxiety, or other related developmental disorders that could be considered as affected.

The SSC1 cohort provides our study with 518 families, each with one affected offspring and an unaffected sibling control. The SSC1 cohort was preselected on basis of screening negative for any plausibly causal de novo or inherited variant, including SNVs, INDELs, or SVs taken from genotyping arrays or exome sequencing. Thus, it is plausible to consider that the SSC1 cohort is more likely to contain noncoding risk signal than other datasets that include samples with high risk variants.

The replication dataset consists of two cohorts: the Simons Simplex Collection Phases 2,3, and 4 (SSC2-4) and the MSSNG cohort (principal investigator: Stephen Scherer). The SSC2-4 cohort has not been preselected like the SSC1, evident by the presence of previously published samples in the cohort[33]. The MSSNG cohort includes 1395 autism cases from 1187 families, while the SSC2-4 cohort contains 1621 families with an affected child (6047 individuals). Full counts of individuals in each cohort can be found in Table 3.1

Table 3.1: Sample Counts by Cohort

	Cohort	**Families**	**Individuals**	**Cases**	**Controls**
Discovery	REACH	309	1095	362	112
	SSC1	518	2072	518	518
Replication	SSC2	598	2392	598	598
	SSC3	783	3125	783	776
	SSC4	442	1331	442	5
	MSSNG	1187	3769	1395	0
Total		**3837**	**13784**	**4098**	**2009**

3.3.2 Whole Genome Sequencing

Samples from the REACH cohort were sequenced at Human Longevity Inc. (HLI) on Illumina HiSeq X10 machines with 150bp long paired-ends at a mean coverage of 50X. 204 individuals were sequenced on the older Illumina HiSeq 2500 platform that have been previously described[11]. The SSC1 and SSC2-4 cohorts were sequenced at the New York Genome Center on an Illumina HiSeq X10 (150bp paired-end reads, mean coverage 40X). The replication MSSNG

cohort from the Autism Speaks initiative were sequenced on Illumina HiSeq X10 machines at a mean depth of 30X. All DNA samples were derived from blood. Standard quality control steps were performed to ensure relatedness, paternity, and gender for each sample.

For REACH and SSC1, genomes were aligned to the human hg19 (GRCh37) reference with bwa-mem[46]. The SSC2-4 cohorts were aligned with bwa-mem[46] to the human reference hg38 (GRCh38). Duplicate reads were flagged and removed from each sample. Additionally, base quality scores were recalibrated with GATK[51] BaseQualityScoreRecalibration. SNPs and INDELs for the REACH and SSC1 cohort were called using GATK[51] HaplotypeCaller and recalibrated using the default parameters for VariantQualityScoreRecalibration. For the SSC2-4 cohort, SNPs and INDELs were called using GATK[51] HaplotypeCaller and recalibrated with VariantQualityScoreRecalibration, but were genotyped jointly in sub-cohorts of roughly 50 families. Since for SV genotyping, SNVs and INDELs are used in a way that is similar to microarray probes, the differences between pedigree-wise variant calling and joint genotyping do not affect the SV genotyping.

3.3.3 SV Detection, Filtering, and Genotyping

Structural Variants were called in each sample using ForestSV[54], LUMPY[44], and Manta[16]. ForestSV implements a Random Forest classifier to predict the copy number of a given genomic region, because of the windowing approach, SVs called by forestSV lack precise breakpoints. LUMPY and Manta call SVs with discordant paired-ends and split-reads, allowing for better resolution of smaller variants ($<$500bp). Since LUMPY and Manta rely on aligned reads to find SVs, it is restricted to breakpoints that fall outside repetitive elements. For short 150bp reads, alignments are rarely confidently aligned to repeat elements such as segmental duplications and short tandem repeats. However, we are not limited to SVs that are mediated by repeats, such as those that derive from non-allelic homologous recombination, because of forestSV's windowing approach. ForestSV also implements coverage in predicting SVs, a feature

that LUMPY and Manta do not use. Hence, by creating a consensus call set from the three methods, we are able to sensitively assay the genome for SVs.

We first removed variants that overlapped more than 66% with centromeres, segmental duplications, regions with low mappability with 100bp reads, regions subjects to somatic V(D)J recombination such as antibody regions and T-cell receptors (data obtained from UCSC Table Browser[39]. These regions either lack reads due to their repetitive context, such is the case with telomeres, or appear as false positive germline de novo variants, since they are somatic in origin (as the case is with T-cell receptors). LUMPY and Manta SVs were omitted if either breakpoint overlapped with one of these regions. Additionally, LUMPY/Manta SVs were omitted if the breakpoint overlapped a short tandem repeat. We then generated a set of uniformly called genotypes for the combined set of deletions and duplications with SV^2[5]. We then filtered variants according to SV^2 filters that were determined using an orthogonal cohort[5]. Variants that passed the standard filters of SV^2 were retained. Variants that overlapped within samples were collapsed to the variant with the highest median ALT genotype likelihood. A consensus call set was generated by merging the calls from the three algorithms. ForestSV calls were collapsed if variants across samples overlapped $\geq$50% reciprocally. For LUMPY and Manta, SVs were collapsed if the confidence intervals for both the start and end positions overlapped. After genotyping variant allele frequencies were calculated with plink[63] with the resulting VCF files.

3.3.4 SV Validation by Microarrays

We estimated the false discovery rate (FDR) of deletions and duplications with Illumina 2.5M SNP array data on a subset of 205 genomes using the Intensity Rank Sum test implemented using SVToolkit's IntensityRankSum test. Briefly, the software performs a rank sum test of probe intensities across a cohort, thus for rare variants (allele frequency $<$1%) this method is quite robust at validating structural variants. However, due to the limitations of microarrays, SVs can be validated if they are over 10kbp.

3.3.5 SV Validation by Nanopore Sequencing

We validated our SV detection and genotyping approach in three unrelated individuals with Oxford Nanopore (ONP) long read sequencing. ONP reads were aligned to the human hg19 (GRCh37) reference with bwa-mem[46] and ngmlr[72]. The average coverage was 7.4X and average read length was 2574bp for bwa-mem alignments and 7.3X and 2525bp for ngmlr alignments. We restricted validation to variants with less than 50% overlap to elements in our genome mask. Additionally, we ensured that the median base-pair depth of coverage was greater than 0X in 1000bp regions flanking the breakpoints, totaling 3252 deletion and 62 duplication candidates for validation. We then searched for supporting reads in bwa-mem and ngmlr alignments, defined as supplementary alignments or CIGAR string deletions and insertions with breakpoints that overlap at least 50% reciprocally to the SV in question. Short-read SV predictions were considered validated if at least 1 supporting read was detected in either bwa-mem or ngmlr alignments.

3.3.6 Prioritization of Functional Elements

Each variant was prioritized according to overlap to functional elements. We created a hierarchy of annotations that classifies each variant into one of four groups: exonic, cis-regulatory, intronic, or noncoding. For this analysis only the exonic, cis-regulatory, and intronic variants were analyzed. Variants that intersected at least one exon were labeled as ”exonic”. We defined cis-regulatory as variants that do not intersect coding regions, but intersect either a 3’UTR, transcription start site, and/or promoter. Intronic variants do not overlap coding regions or cis-regulatory regions, but are in introns, which are not as constrained as coding regions.

We then determined the gene and the corresponding pLI score each exonic, cis-regulatory, and intronic variant. If a variant intersected two or more genes, the gene with the highest (most functionally constrained) score was reported. Our target list were generated using the hg19

reference, for the SSC2-4 variants that were called on a hg38 reference, we transformed the coordinates to hg19 with liftOver[38], with a minimum of 75% matching bases to the destination locus.

3.3.7 Group-Wise Transmission Disequilibrium Test

For this analysis, we only tested private variants, those that are unique to one parent in each family. Our rationale for this restriction is that private variants are evolutionarily young and haven't had enough time for natural selection to remove deleterious alleles from the population. As described earlier in this work, the transmission disequilibrium test accounts for deviations from the expected transmission rate of 50%. Typically the transmission disequilibrium test is performed on a per-variant basis, but this approach would fail due to our requirement for private variants. Hence, we sought to group variants according to functional constraint and sum the number of transmitted and not transmitted variants in each group to determine the overall transmission rate. This approach is robust, since it lowers the number of statistical tests that need to be performed[15]. Additionally by limiting to rare variants, the variants become effectively unlinked, thus distortions that arise from common variants due to population stratification is removed.

We then grouped variants according to functional constraint (pLI >0.9) and tested if exonic or cis-regulatory deletions were over-transmitted to affected offspring, relative to controls. An over-transmission of putative risk variants would suggest disease association, since those variants are unlikely to occur in healthy offspring. Since the total number of independent transmissions follow a binomial distribution, we defined statistical significance with the binomial p-value.

3.4 Results

3.4.1 Landscape of Deletions in Human Whole Genomes

Our discovery data set consisted of whole-genome sequencing (mean coverage = 42.6X) of 829 families, comprising 880 affected individuals, 630 unaffected individuals, and their parents. This discovery cohort consisted of two data sets, the REACH study containing 309 families, and the Simons Simplex Collection Phase 1 (SSC1) study containing 518 families. The samples in the SSC1 cohort were selected on the basis that they had previously screened negative for de novo loss of function mutations or large copy number variants from exome sequencing[33] and microarray studies[69]. The ascertainment of this sample was therefore designed to eliminate the well-established categories of genetic risk and thereby to enrich for novel inherited and noncoding risk variants.

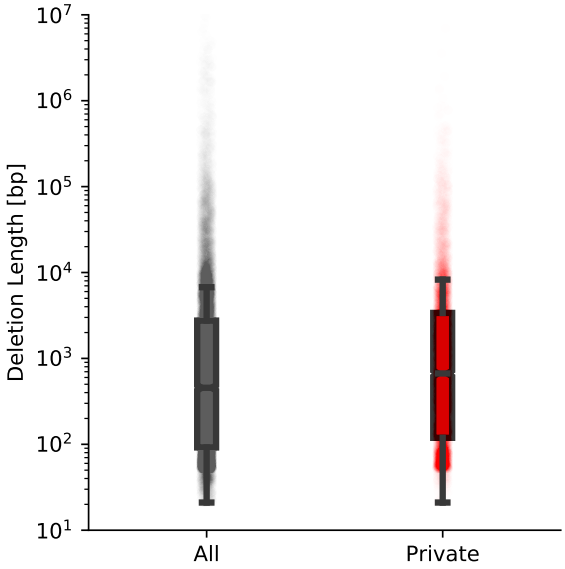

Figure 3.1: **Size Distribution of Deletions.** Boxplots for all deletions (left) and private deletions (right). Private deletions are those that exist in one founder (parent). The mean deletion length for all deletions was 21,472bp (25pc-75pc: 92-2755bp; IQR= 2663bp). For private deletions the mean length was 9171bp (25pc-75pc: 119-3384bp; IQR= 3265)

We developed a pipeline for genome-wide analysis of SV that consisted of three structural

variant callers, forestSV[54], LUMPY[44], and Manta[16]. We then generated a consensus call set by collapsing overlapping variants and removing variants that intersected problematic regions for sequencing by more than 66% of the lenght of the SV. A key innovation was the development of SV^2[5], a support-vector machine based software for accurately estimating genotype likelihoods from short-read WGS data, which enabled accurate genotyping of SVs in families with a detection limit of $\geq$100bp[5]. The mean length of all deletions was 21,472bp (IQR = 2663bp), while for private deletions the mean was 9171bp (IQR = 3265bp; Figure 3.1)

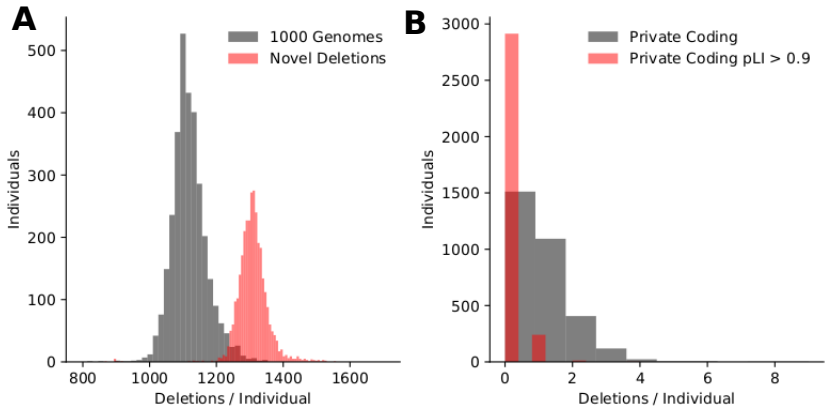

Figure 3.2: **Burden of Deletions in 3169 Individuals.** (A) Distribution of deletions per individual for variants that intersect common SVs found in the 1000 Genomes phase 3 call set[76] (shown in black), and variants that were novel to the 1000 Genomes call set (shown in red). The average number of deletions per individual was 2427 variants (IQR = 87). (B) Distribution of deletions in coding regions. Variants that intersect exons with pLI scores $\geq$0.9 are shown in red. The average number of private deletions per sample was 14 variants (IQR = 7); the average number of private coding deletions was 0.774 (IQR = 1), while the number of private coding deletions in haploinsufficient genes was 0.086. 255 samples had 1 or more private deletion in a functionally constrained gene, while 14 had two or more.

We then calculated the burden of deletions in each individual (Figure 3.2A) stratifying according to deletions that intersected with the 1000 Genomes phase 3 release[76] and those that were novel to the 1000 Genomes call set. The average number of deletions that were also present in the 1000 Genomes data set was 1118 variants per individual (IQR = 59), while the average burden for novel variants were 1308 variants per individual (IQR = 47). We also measured the burden of private deletions that intersected exons (Figure 3.2B) and found the number of exonic

deletions per sample to be 0.774 variants. For functionally constrained exons, defined as pLI $\geq$0.9, the average number of variants was calculated to be 0.86. There were 255 individuals with 1 or more deletion in a haploinsufficient gene; 14 people contained two or more risk deletions with 4 deletions being the highest number of observed deletions in exons with pLI scores greater than 0.9.

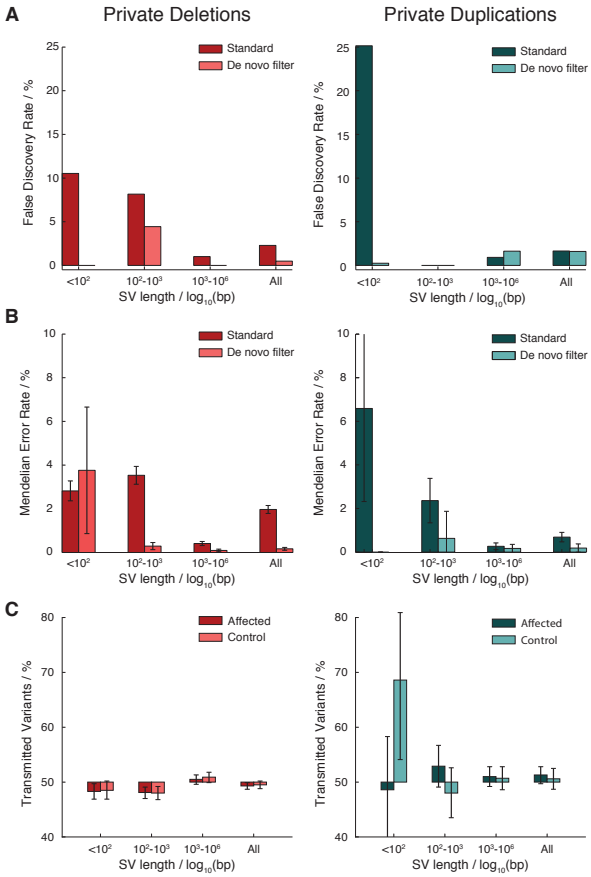

Figure 3.3: **Metrics of genotyping accuracy for deletions and duplications by size.** Bar charts illustrating (A) FDR based on intensity rank sum test from microarray, (B) Mendelian error rates, and (C) variant transmission rates stratified on SV type (deletion and duplication) and SV length bins for private variants. Quality metrics are reported for all private SVs in the callset filtered based on SV^2 genotype likelihood at two levels of stringency (”standard” and ”de novo”). Whiskers represent 95% confidence intervals.

3.4.2 SV Validation by Microarrays and Inheritance-Based Methods

An average of 3746 SVs were detected per individual, including biallelic deletion, tandem duplications, inversions, four classes of complex SV, and four families of mobile element insertion. The overall false discovery rate (FDR) was estimated from Illumina 2.5 M single nucleotide polymorphism (SNP) array data to be 4.2% for deletions and 9.4% for duplications (Figure 3.3A). Private deletions and duplications >100 bp in length displayed low Mendelian error rates (below 5% for deletions across all size ranges and duplications >100bp), with an overall Mendelian error rate of 2.5% for deletions and 1.5% for duplications (Figure 3.3B). We also tested for transmission distortion of all private deletions and duplications and found no apparent bias for deletions (average transmission rate = 0.497) or duplications (average transmission rate = 0.52) (Figure 3.3C).

Table 3.2: False discovery rate of SVs across size ranges and filters.

SV Type	Parent Allele Frequency	Size Range	# SV	# SV Covered with ONP	Failed	Passed	Validation Rate
DEL	All	<100bp	1168	1086	134	952	0.88
DEL	All	100bp-1kb	1583	1486	157	1329	0.89
DEL	All	>1kb	730	680	56	624	0.92
DEL	All	All	3481	3252	347	2905	0.89
DEL	<1%	<100bp	33	33	2	31	0.94
DEL	<1%	100bp-1kb	61	60	6	54	0.90
DEL	<1%	>1kb	51	50	5	45	0.90
DEL	<1%	All	145	143	13	130	0.91
DEL	Private	<100bp	4	4	0	4	1.00
DEL	Private	100bp-1kb	14	14	0	14	1.00
DEL	Private	>1kb	22	22	0	22	1.00
DEL	Private	All	40	40	0	40	1.00
DUP	All	<100bp	0	NA	NA	NA	NA
DUP	All	100bp-1kb	8	8	2	6	0.75
DUP	All	>1kb	55	54	15	39	0.72
DUP	All	All	63	62	17	45	0.73
DUP	<1%	<100bp	0	NA	NA	NA	NA
DUP	<1%	100bp-1kb	0	NA	NA	NA	NA
DUP	<1%	>1kb	11	11	0	11	1.00
DUP	<1%	All	11	11	0	11	1.00
DUP	Private	<100bp	0	NA	NA	NA	NA
DUP	Private	100bp-1kb	0	NA	NA	NA	NA
DUP	Private	>1kb	4	4	0	4	1.00
DUP	Private	All	4	4	0	4	1.00

3.4.3 SV Validation by Nanopore Sequencing

We also validated SV genotyping through Oxford Nanopore whole genome sequencing of 3 unrelated individuals to a mean coverage of 7X to 9X. Due to the novelty of long read technology, we chose to align the samples with two algorithms, bwa-mem[46] and ngmlr[72]. For each alignment, we restricted validation to variants with les than 50% overlap to elements in our genome mask. Additionally, we ensured that the median base-pair coverage was greater than 0X in 1000bp flanking regions. With these filters, we tested the validation of 3252 deletions and 62 duplications in the three samples, calling a SV valid if one or more supporting Nanopore read was present in either alignment. We then calculated the false discovery rate specifying false positives as SVs without supporting reads while binning on allele frequency and SV length. The overall FDR was 10.4% for deletions and 30.6% for duplications; for private variants of SV length 100bp-1000bp the FDR was 0% for deletions (Table 3.2).

3.4.4 Transmission of Private Deletions

Measures of functional constraint that are based on population data are useful metrics for predicting the pathogenicity of rare variants. For example, genes that display strong negative selection against loss-of-function variants in the general population, as assessed by the Exome Aggregation Consortium (ExAC)[45], are highly enriched in de novo mutations in children with ASD[66]. Likewise, it's been reported that the intolerance of genes to exonic deletions is correlated with the SNV-based pLI metric[10]. With this in mind, we can then test for disease association using private (those in one parent) deletions that overlap haploinsufficient genes (pLI $>$0.9), with the rationale that private variants are evolutionarily young and haven't had enough time for natural selection to purify deleterious alleles from the population.

We then classified each private deletion on whether it overlapped coding regions (Loss of Function), cis-regulatory elements, defined as 3'UTRs, transcription start sites, and promoters

(CRE-SV), or introns. This classification scheme was hierarchical, with coding deletions taking precedence to cis-regulatory elements, to introns.

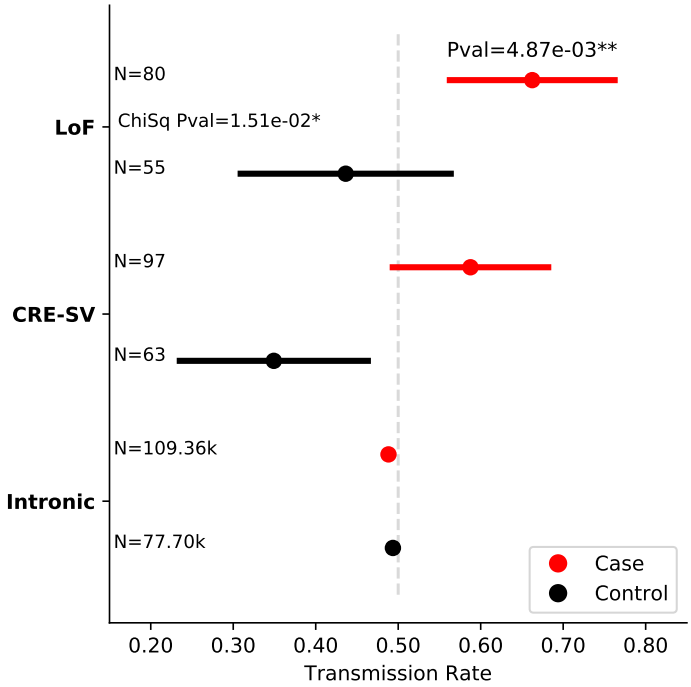

Figure 3.4: **Transmission Disequilibrium of Private Deletions in Functionally Constrained Genes.** Transmission disequilibrium of private deletions in 829 families (880 affected offspring, 680 control siblings). Deletions were categorized into three groups according to their intersection with genic elements (pLI $\geq$0.9): Loss of Function (LoF), CRE-SV, and Intronic. LoF variants were over-transmitted to cases, but not to controls (binomial p-value = 4.87×10^{-3}, Chi-Square p-value = 0.00151). Although the CRE-SVs seem to be over-transmitted to cases and under-transmitted to controls, the results are not significant. Likewise, as expected, we show no transmission distortion of intronic deletions in haploinsufficient genes. P-values over data points represent two-tailed binomial p-values. Error bars indicate 95% confidence intervals defined as the binomial proportion confidence intervals. Numbers for each group signify the number of independent transmission events for cases and controls.

Focusing on the target functional categories above, family-based association was tested using a group-wise transmission disequilibrium test, assuming a dominant model of transmission with additive effects[15]. As expected, since most intronic mutations are more likely to be under neutral selection[13, 29], private intronic deletions exhibited no transmission distortion (case transmission rate = 0.485, control transmission rate = 0.49). Loss of function (LoF) mutations

in functionally constrained genes were over-transmitted to cases but not to controls (Affected transmission rate = 0.676, binomial p-value = 4.87×10^{-3}, Chi-Square p-value = 0.00151. Figure 3.4). CRE-SVs appeared to be slightly over-transmitted in cases and under-transmitted to controls, but the results were not statistically significant (Figure 3.4).

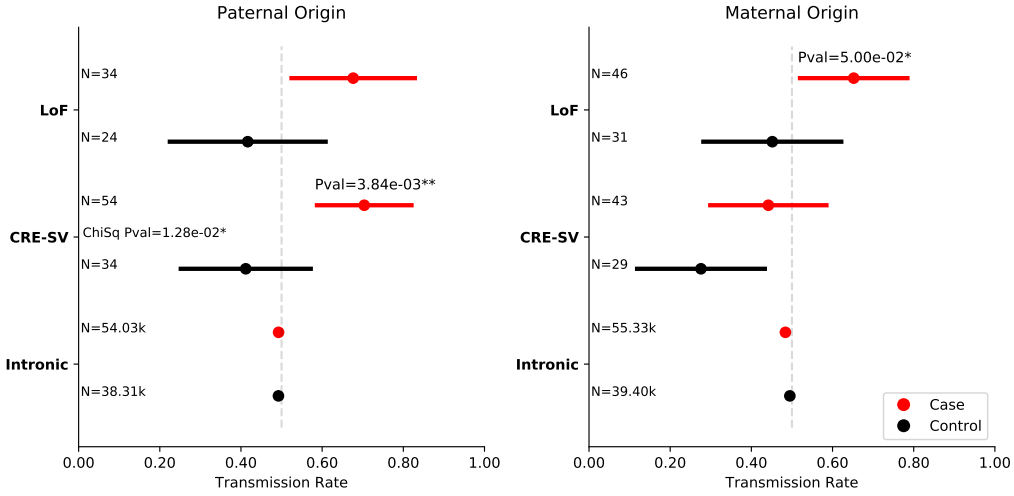

Figure 3.5: **Paternally Derived CRE-SVs are Associated with Autism.** Variant stratification and filtering were performed as outlined in Figure 3.4. While stratifying on the parent of origin (fathers left, mothers right), mothers over-transmitted LoF variants to cases, (binomial p-value = 0.05). Fathers but not mothers over-transmit CRE-SVs in haploinsufficient genes (affected transmission rate = 0.703, binomial p-value = 3.84×10^{-3}, Chi-Square p-value = 0.00128). We report no transmission distortion for intronic variants from either parent. P-values over data points represent two-tailed binomial p-values. Error bars indicate 95% confidence intervals defined as the binomial proportion confidence intervals. Numbers for each group signify the number of independent transmission events for cases and controls.

3.4.5 Paternal Origin Effect of Cis-Regulatory Deletions

We then stratified the transmission disequilibrium test according to parent of origin and performed the same test outlined above. We found that mothers, but not fathers over-transmitted LoF variants but with marginal significance (affected transmission rate = 0.652, binomial p-value = 0.05, Chi-Square p-value = 0.13). This could be partially explained that females tend to have a higher tolerance for autism risk[86]. However, fathers but not mothers over-transmitted

CRE-SVs (Figure 3.5) to affected offspring (transmission rate = 0.703, binomial p-value = 3.84×10^{-3}, Chi-Square p-value = 0.00128), relative to controls (transmission rate=0.412, binomial p-value=0.392).

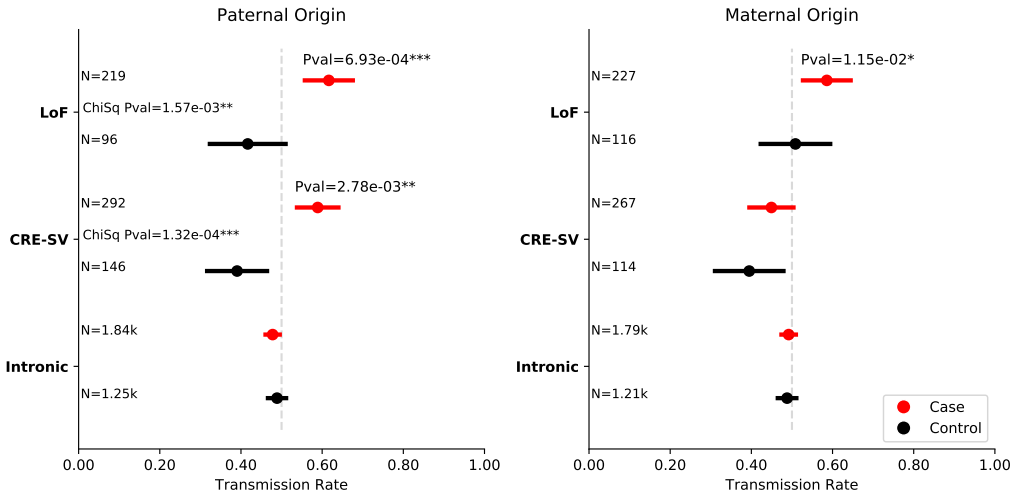

Figure 3.6: **Combined Analysis of Transmission Disequilibrium for Private Deletions in Functionally Constrained Genes.** Variant stratification and filtering were performed as outlined in Figure 3.4. The combined analysis included the REACH, SSC1, MSSNG, and SSC2-4 cohorts, totaling 3837 families (4098 cases, 2009 controls). Transmission counts are stratified according to the parent of origin (fathers left, mothers right). Both mothers and fathers over-transmitted LoF variants to cases. However, fathers but not mothers over-transmit CRE-SVs in haploinsufficient genes (transmission rate = 0.589, binomial p-value= 0.0028, Chi-Square p-value= 1.32×10^{-4}). We report no transmission distortion for intronic variants from either parent. P-values over data points represent two-tailed binomial p-values. Error bars indicate 95% confidence intervals defined as the binomial proportion confidence intervals. Numbers for each group signify the number of independent transmission events for cases and controls.

3.4.6 Replication of the Association of Paternally Inherited CRE-SVs

We then replicated the association by applying our pipeline to an independent sample of 3010 families consisting of the MSSNG initiative and the Simons Simplex Collection Phases 2-4. The association of private paternally transmitted CRE-SVs was significant in the replication sample. Also consistent with our primary results, maternally transmitted CRE-SVs were not associated with autism and inherited coding variants from both parents were associated with

autism. In the combined data set of 3837 families, the association of paternal CRE-SVs was significant. Consistent with a paternal-origin effect, CRE-SVs in cases were inherited more frequently from fathers (transmission rate = 0.589, binomial p-value= 0.0028, Chi-Square p-value= 1.32×10^{-4}). Both fathers and mothers over-transmitted LoF variants to cases relative to controls (Figure 3.6). The median lengths of cis-regulatory and exonic SVs were 2920 bp [interquartile range (IQR) = 396 to 8282 bp] and 17,261 bp (IQR = 4390 to 112,251 bp), respectively (Figure 3.7). The smaller effect size observed in the replication sample (over-transmission of 59.6%, compared with 70.6% in the discovery sample) could be explained by a combination of factors, including chance or true differences in the genetic architecture between samples. Cohorts did not differ dramatically in the numbers of trios and concordant sibling pair (multiplex) families; thus, family structure is unlikely to have an influence. As mentioned above, selection of families for a subset of the discovery sample (SSC1) was designed to enrich for novel inherited and noncoding risk variants. Thus, ascertainment could in part explain why the SSC1 had the largest effect size of all individual cohorts.

3.5 Discussion

Here we demonstrate that rare SVs that disrupt CREs confer risk for autism, and this association is concentrated among genes that are highly dosage sensitive (pLI $\geq$0.9). The contribution of CRE-SVs that we observe consists exclusively of inherited variants that are carried by a parent. This result is consistent with noncoding variants having moderate effects on gene function and disease risk[75]. We find no evidence for a contribution of de novo CRE-SVs, in contrast to anecdotal findings from previous studies[21, 79]. We cannot exclude the possibility that de novo CRE-SVs contribute to autism; however, we can conclude that they are extremely rare. CRE-SVs exhibited a significant paternal-origin effect. This result was unexpected and contrasts with a simpler genetic model[86] in which inherited genetic risk is transmitted predominantly

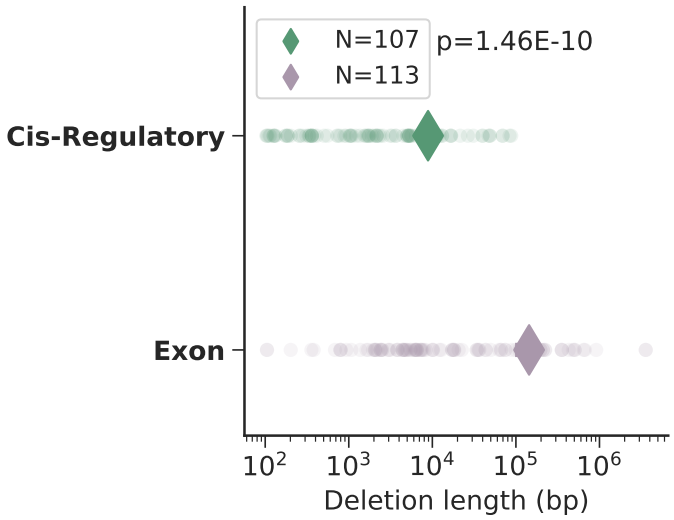

Figure 3.7: **SV Length Distribution of Private LoF and CRE-SVs.** Private deletions that intersect cis-regulatory elements (CRE-SV) or exons (LoF). Diamonds represent the mean. The median lengths of cis-regulatory and exonic SVs were 2920 bp [interquartile range (IQR) = 396 to 8282 bp] and 17,261 bp (IQR = 4390 to 112,251 bp), respectively. CRE-SVs tend to be smaller than LoF SVs (T-test p-value = 1.46×10^{-10}).

from mothers due the reduced vulnerability of females to autism. The rationale behind this is that de novo mutations that confer autism risk is less penetrant in females; hence, when the risk variant is transmitted mother to son, the chance for autism increases drastically due to the decreased tolerance for autisk risk in males[86]. Previous studies have shown a maternal bias for inherited truncating variants in genes that were previously implicated from studies of de novo mutation[33, 42, 81]. In our study, the contribution of exonic variants to risk was similar for paternal and maternal SVs, suggesting that a maternal origin bias might be restricted to genes that have the most extreme dosage sensitivity. Taken together, our findings indicate that parent-of-origin effects on genetic risk for autism are more complex than we previously thought, and the allelic spectrum of variants differs between the maternal and paternal genomes.

We propose three possible mechanisms to explain the observed paternal-origin effect of CRE-SVs, the first is a “bilineal two-hit model”, in which inherited risk is attributable to a combination of two risk variants: a maternally-inherited coding variant of large effect and a

paternally-inherited CRE variant of moderate effect. This bilineal model predicts that a paternal bias might also be evident for other variants of moderate effect including hypomorphic missense alleles or LoF variants in genes with a moderate degree of intolerance. Likewise, a genetic study of common variation in autism families reported suggestive evidence of a paternal bias for variants of modest effect[84], a result that lends support to a bilineal model. One could test such a model by testing the transmission of variants, both LoF and CRE-SVs while conditioning risk from one parent. That is to say, given a mother transmitted a LoF variant in a haploinsufficient gene, what is the remaining autism risk stemming from the father for LoF variants and CRE-SVs.

An alternative explanation for a paternal-origin effect is an epigenetic mechanism. For example, deletion of CREs can lead to de-repression of imprinted genes[78]. However, an epigenetic mechanism could only explain our results if non-canonical imprinting of regulatory elements is widespread. Such a phenomenon has not been described, but we cannot rule out this possibility. In fact, differential imprinting has been characterized throughout human and mouse brain development[27, 40]. This would suggest that potentially, the striking parent of origin effect we observed could be explained by imprinting. However, better maps of imprinting throughout development of the central nervous system is needed before exploring this avenue. A third potential mechanism to explain parent-of-origin effects could be a type of "meiotic drive", in which allele-specific selection occurs differently in paternal and maternal germ cells. This mechanism is interesting and would explain the apparent under-transmission of CRE-SVs in controls. Implying that selective pressures rendered the ancestral locus to be selfish and encourages its transmission in germ cells. However, this mechanism is also unlikely given that there are few known examples of gene drive in humans and their effects appear to be quite weak at the population level[36]. Due to the greater potential of SVs to impact gene function and regulation relative to SNVs and indels, this class of genetic variation has historically proven effective for illuminating new components of the genetic architecture of disease. Our findings provide a further demonstration of the utility of SV analysis for characterizing the genetic

regulatory elements that influence risk for autism.

Chapter 3 has been previously published in *Science* (William M. Brandler, Danny Antaki, Madhusudan Gujral, Morgan L. Kleiber, Joe Whitney, Michelle S. Maile, Oanh Hong, Timothy R. Chapman, Shirley Tan, Prateek Tandon, Timothy Pang, Shih C. Tang, Keith K. Vaux, Yan Yang, Eoghan Harrington, Sissel Juul, Daniel J. Turner, Bhooma Thiruvahindrapuram, Gaganjot Kaur, Zhuozhi Wang, Stephen F. Kingsmore, Joseph G. Gleeson, Denis Bisson, Boyko Kakaradov, Amalio Telenti, J Craig Venter, Roser Corominas, Claudio Toma, Bru Cormand, Isabel Rueda, Silvina Guijarro, Karen S. Messer, Caroline M. Nievergelt, Maria J. Arranz, Eric Courchesne, Karen Pierce, Alysson R. Muotri, Lilia M. Iakoucheva, Amaia Hervas, Stephen W. Scherer, Christina Corsello, and Jonathan Sebat. 2018. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* 360(6386):327-331). The dissertation author shares primary authorship with William M. Brandler and Madhusudan Gujral. William M. Brandler assisted with the conception, variant filtering and merging, and genetic analysis. Madhusudan Gujral assisted with sequence alignment and variant calling. The dissertation author was responsible for providing input on methodologies, software development, variant genotyping and filtering, and formal analysis. Full author contributions are detailed in the publication[10]. Jonathan Sebat supervised the project and provided invaluable advice.

Chapter 4

Different Maternal and Paternal Contributions to Autism

4.1 Abstract

The genetic basis of autism is known to consist of de novo and inherited loss of function mutations in haploinsufficient genes. Previous studies have reported that inherited risk is predominantly carried by mothers, believed to be due to incomplete penetrance of risk alleles in females. However, the distinct contributions of each parent to inherited risk for ASD has not been explored in depth. We investigated paternal and maternal contributions to autism by analyzing the transmission of private deletions in coding and cis-regulatory (CRE-SVs) regions of haploinsufficient genes in whole genomes of 10,015 individuals (2650 families). We also extended our transmission distortion analysis of private coding deletions to encompass of loss of function single nucleotide variants (SNVs) and insertion/deletion (INDELs), defined as a gain of a stop codon, frameshift mutation, or a splicing mutation, as well as private missense mutations stratified according to missense intolerance. Our goal is to untangle distinct modes of inheritance for autism risk, hypothesizing that fathers and mothers carry distinct risk contributions for the

disease and that, in some cases, autism risk is bilineal in nature. We report that mothers and fathers over-transmit loss of function variants within functionally constrained coding regions. However, fathers but not mothers tended to over-transmit CRE-SVs and missense variants within risk regions to affected offspring. When we test the segregation of loss of function variants stratified by sex of the offspring. We find that most of the genetic risk to sons is derived from the father, which is not consistent with the previous hypothesis that mothers are the sole contributors of inherited risk to sons. Our work demonstrates that the paternal origin effect of risk variants contributes genetic risk for autism in more complex ways than previously thought.

4.2 Introduction

Inherited risk for autism is thought to primarily derive from mothers[86]. This theory is known as the female protective effect model that Sebat and Wigler proposed. The basis behind this model is the fact that females require a higher genetic risk load than males, evident by the fact affected females have more de novo burden than males[69]. If a de novo mutation occurs in a female, she may not develop autism since she has increased tolerance for risk alleles. However, if she transmits this variant to a son, he may develop autism since he is more susceptible to autism risk. This theory is bolstered by the observation that more males are diagnosed than females with autism[64, 12], suggesting that males are more susceptible to the disease. Previous studies using exome sequencing have associated maternal but not paternal inherited LoF variants with autism[42]. So this brings us to ask if fathers contribute inherited risk to autism, outside of CRE-SVs[10].

The effect of inherited LoF mutations and missense mutations in autism haven't been explored in large datasets. To this end, we collected whole genome sequence data from 2650 families (10,015 individuals) and SNV, INDEL, and SV variant calls. We hypothesize that a significant component of missing heritability in autism can be explained by inherited variants.

We also would like to investigate the possibility for a bilineal model for autism, where each parent contributes inherited risk to the disease. We plan to stratify missense variants according to pathogenicity using PolyPhen-2[3] scores of deleteriousness and ExAC's[45] measure of missense intolerance for genes. We also plan to condition transmission disequilibrium tests to test for evidence of sex-specific inheritance, since it's been previously reported that mothers tend to transmit LoF mutations to sons[42]. In summary, we plan to measure the effect of rare inherited damaging mutations in autism, hypothesizing that many cases can be explained by inherited LoF mutations that derive from the father.

4.3 Methods

4.3.1 Study Design

Our study implemented the cohorts listed in the previous chapter, with the exception of the MSSNG data set, which lacks SNV and INDEL calls. Sample counts for the cohorts with SNV, INDEL, and SV calls are shown in Table 4.1. The REACH cohort represents a clinical sampling of autism; the only requirement for selection was an autism diagnosis. In contrast, the SSC1 cohort (518 families), was preselected on the condition that children did not carry a damaging loss of function (LoF) mutation[33, 69]. The remaining SSC cohorts do not meet this criteria, since many samples in this cohort have a previously reported LoF variant[33, 69]. Worthy of note, the SSC4 cohort is primary trios, which differs from the rest of the SSC quad families. In total, there are 10,015 individuals in the combined cohort (2650 families, 2703 cases, 2009 controls; Table 4.1). Of the 2703 affected offspring, 2314 were males and 389 were females (male:female ratio = 5.95); the male-female ratio for controls was 0.9 (949 males, 1060 females).

Table 4.1: Sample Counts by Cohort

Cohort	Families	Individuals	Cases	Controls
REACH	309	1095	362	112
SSC1	518	2072	518	518
SSC2	598	2392	598	598
SSC3	783	3125	783	776
SSC4	442	1331	442	5
Total	**2650**	**10015**	**2703**	**2009**

4.3.2 Whole Genome Sequencing

Samples from the REACH cohort were sequenced at Human Longevity Inc. (HLI) on Illumina HiSeq X10 machines with 150bp long paired-ends at a mean coverage of 50X. 204 individuals were sequenced on the older Illumina HiSeq 2500 platform that have been previously described[11]. The SSC1-4 cohorts were sequenced at the New York Genome Center on an Illumina HiSeq X10 (150bp paired-end reads, mean coverage 40X).

REACH genomes were aligned to the human hg19 (GRCh37) reference with bwa-mem[46]. The SSC1-4 were aligned with bwa-mem[46] to the human hg38 (GRCh38) reference build. Duplicate reads were flagged and removed from each sample. Additionally, base quality scores were recalibrated with GATK[51] BaseQualityScoreRecalibration. For each family, SNPs and INDELs for the REACH cohort were called using GATK[51] HaplotypeCaller and recalibrated using the default parameters for VariantQualityScoreRecalibration. For the SSC1-4 cohort, SNPs and INDELs were called using GATK[51] HaplotypeCaller and recalibrated with VariantQualityScoreRecalibration (VQSR), but were genotyped jointly in sub-cohorts of roughly 50 families.

4.3.3 Structural Variation Detection, Filtering, and Genotyping

Structural Variants were called in each sample using ForestSV[54], LUMPY[44], and Manta[16] and then genotyped with SV^2. We first removed variants that overlapped more than

66% with centromeres, segmental duplications, regions with low mappability[37], regions subjects to somatic V(D)J recombination such as antibody regions and T-cell receptors (data obtained from UCSC Table Browser[39]. LUMPY and Manta SVs were omitted if either breakpoint overlapped with one of these regions. Additionally, LUMPY/Manta SVs were omitted if the breakpoint overlapped a short tandem repeat.

We then filtered variants according to SV^2 filters. Variants that passed the standard filters of SV^2 were retained. Variants that overlapped within samples were collapsed to the variant with the highest median ALT genotype likelihood. For variants that were not genotyped by SV^2 because of masking or erratic coverage[5], we retained the variant if it was genotyped by either LUMPY or Manta and if the median ALT genotype likelihood passed filters that were previously determined[10].

We then extracted private variants (those that are unique to founder/parent) by testing the reciprocal overlap of variants in one family to variants in all families. We defined private variants as those with reciprocal overlap less than 50% to all other variants in unrelated pedigrees.

4.3.4 De Novo Mutation Detection

De novo SVs were called if they occurred in a child and were genotyped reference in both parents and the parent allele frequency for the variant was less than 1%. We also applied more stringent SV^2 genotype likelihood filters for de novo SVs and transmission disequilibrium analyses.

De novo SNVs were called with ForestDNM[55] for the REACH cohort. Variants were removed if they were found with greater than 1% allele frequency in either of the the 1000 Genomes[19] phase 3 samples or samples in the Genome Aggregation Database[45] (gnomAD). Additionally, putative de novo calls were removed if it resided in a segmental duplication. De novo SNVs were previously generated for the SSC1-4 cohorts as outlined previously[33, 70, 34, 58, 41]. Since we plan to use the de novo variants to classify affected offspring according to the presence

of a damaging LoF mutation, we did not need de novo calls for the entire genome. Hence, de novo SNV/INDEL and SV calls from exome sequencing and microarrays, respectively, were adequate enough for following analyses.

4.3.5 SNV and INDEL Merging and Filtering

We first transformed the REACH SNV and INDEL calls with hg19 positions to hg38 using GATK (V4.0.5.1) LiftoverVcf command. Variants that were unable to be mapped over to hg38 were omitted from subsequent analyses. We then merged SNVs and INDELs separated for every cohort, using GATK (V3.8-1) CombineVariant command with the ”uniquify” option for merging. This ensures merging does not occur for duplicate ids. Merging was done on 10Mb intervals, allow for faster execution time. Merged INDELs were left-aligned with GATK (V3.8-1) LeftAlignAndTrimVariants command with a 1000bp window. Unlike bcftools[56], GATK's method for left-aligning INDELs omits multiallelic variants, which can complicate downstream analysis. For all subsequent steps, multiallelic variants for SNVs or INDELs were removed.

After merging, we then calculated the allele frequency in founders for every variant with plink[63], which allowed us to extract private variants. We first removed variants with coverage values less than 3 standard deviations from the mean and greater than 6 standard deviations from the mean (mean coverage = 34.9, standard deviation = 9.66, accepted range: 5.92-92.9). Likewise, private variants with allele frequencies greater than 1% in either the 1000 Genomes[19] phase 3 database or the gnomAD[45] database were removed. We then sought filters for private variants by leveraging features produced by VQSR. For each private variant, we calculated the number of families that carry the variant, with the rationale that private variants present in unrelated offspring are likely to be false positives. We then visualized VQSR features for private variants recurrent in unrelated offspring (likely false positive) to variants truly unique to one family (likely true positive). It is important to mention that there is a chance that a private variant in an unrelated child could be a recurrent de novo, but true recurrent de novos are rare and wouldn't affect the

results of this exercise to design filters. We applied a cutoff of $\leq$2.5 to the Symmetric Odds Ratio (SOR) for strand bias and a threshold of $\leq$20 for the Phred-scaled Fisher's exact p-value for strand bias (FS), since a large porprotion of the recurrent private variants fell outside these cutoffs (Figure 4.1A-B). Likewise, we applied a cutoff of $\geq$5 for the variant confidence or quality by depth (QD) feature, also chosen because the majority of the recurrent variants had QD $<$5 evident by the kernel density plots shown in Figure 4.1C.

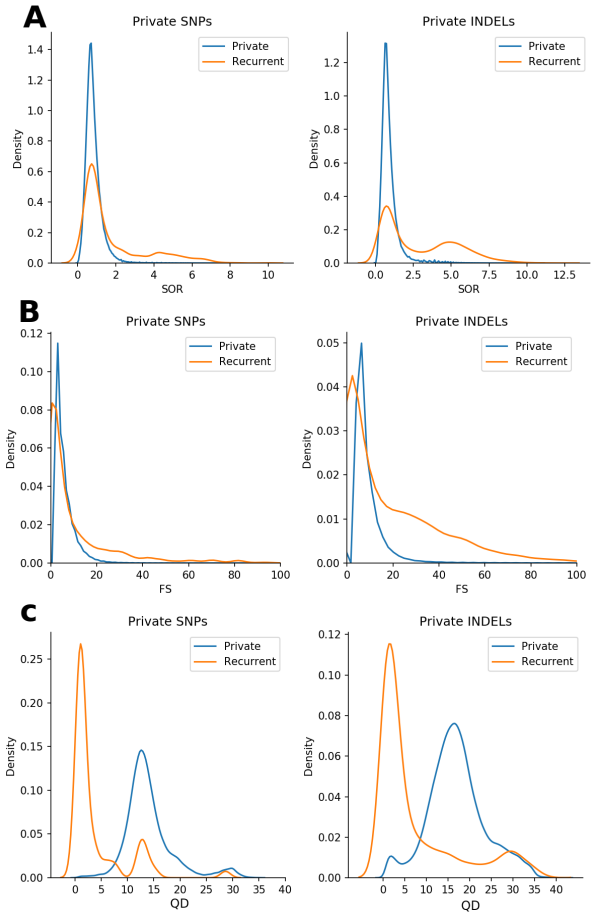

Figure 4.1: **Kernel Density Estimates of VQSR features for Private SNVs and INDELs.** We determined filters according to features from VQSR by comparing private variants that were unique to one family ("Private" in blue) and private variants that were found in unrelated children ("Recurrent" in orange). We chose a cutoff of $\leq$2 for SOR (A) and $\leq$ for FS (B). Variants with QD values less than 5 were omitted (C).

4.3.6 Structural Variant Functional Annotation

Each variant was prioritized according to overlap to functional elements. We created a hierarchy of annotations that classifies each variant into one of four groups: exonic, cis-regulatory, intronic, or noncoding. For this analysis only the exonic, cis-regulatory, and intronic variants were analyzed. Variants that intersected at least one exon were labeled as ”exonic”. We defined cis-regulatory as variants that do not intersect coding regions, but intersect either a 3’UTR, transcription start site, and/or promoter. Intronic variants do not overlap coding regions or cis-regulatory regions, but are in introns, which are not as constrained as coding regions.

We then determined the gene and the corresponding pLI score each exonic, cis-regulatory, and intronic variant. If a variant intersected two or more genes, the gene with the highest (most functionally constrained) score was reported. Our target list were generated using the hg19 reference, for the SSC2-4 variants that were called on a hg38 reference, we transformed the coordinates to hg19 with liftOver[38], with a minimum of 75% matching bases to the destination locus.

4.3.7 SNV and INDEL Functional Annotation

Variants were annotated with Annovar[82] (hg38) using RefSeq[57] for annotating genic function. Annovar also provided exonic variant function annotations such, which allowed us to define loss of function mutations as those with splicing, frameshift insertion, frameshift deletion, frameshift block substitution, stopgain, or stoploss annotations.

We used PolyPhen-2[3] to assign risk scores to missense variants. PolyPhen-2 implements a naive Bayes classifier using features from 8 sequence-based and 3 protein structure-based datasets to predict the probability for a missense mutation to be damaging[3]. The model was trained on two datasets: HumDiv and HumVar. The first dataset is recommended for evaluating rare alleles that contribute risk to complex diseases, while the latter is recommended for Mendelian

disorders[2]. Hence we opted for the HumDiv training set for filtering missense variants. We used a threshold for the PolyPhen score at $\geq$0.9 to define deleterious alleles (PolyPhen recommends a 0.957 threshold[2]). Likewise, we leveraged the missense Z scores calculated by ExAC[45], which are Z scores of missense intolerance for genes. We chose a threshold of $\geq$5 sigma (1.33% of genes in ExAC, Total = 18,241) to select for missense intolerant genes. Note that there is no correlation between PolyPhen HumDiv scores and ExAC missense Z scores, since the former is assigned on a per-variant basis while the latter is applied on a per-gene basis (Figure 4.2). For LoF and missense mutations that overlapped more than one gene, we took the maximum pLI and missense Z score.

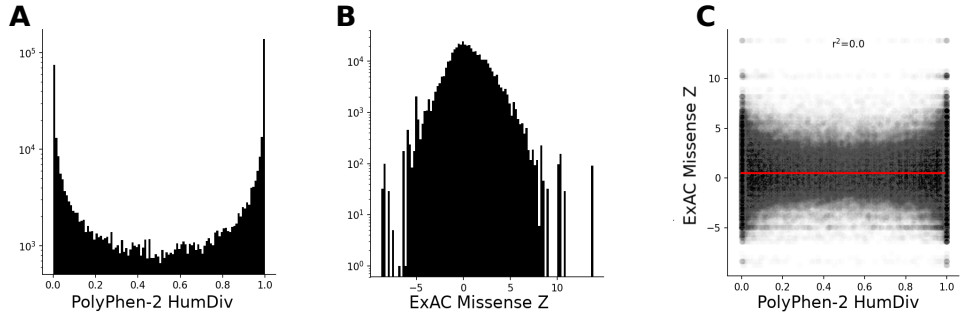

Figure 4.2: **Functional Constraint for Missense Variants.** Missense variants were annotated with PolyPhen-2 HumDiv scores (A) and ExAC missense Z scores (B). The distribution of PolyPhen-2 HumDiv scores for all private missense variants is bimodal (A). Higher HumDiv scores are associated with disease. Likewise, higher missense Z scores are associated with higher intolerance for missense mutations for a given gene (B). Note that there is no correlation between HumDiv scores and missense Z scores (panel C). This is likely due to the fact that PolyPhen assigns scores to variants, while the ExAC missense Z score is assigned to genes. We chose a threshold of $\geq$0.9 for HumDiv scores and $\geq$5 for missense Z scores.

4.3.8 Group-Wise Transmission Disequilibrium Test

Similar to the analysis performed in chapter 3, we determined the transmission of private variants using an adapted form of the transmission disequilibrium test. Our approach groups variants according to functional impact and total the number of transmitted and not transmitted variants in each group to determine the overall transmission rate. This approach is robust, since

it lowers the number of statistical tests that need to be performed[15]. However, the previous method that was implemented (GTDT[15]), contains bugs that makes analyzing large datasets difficult. For example, GTDT requires all variant ids in the VCF to be unique and produces errors if larger pedigrees are not deconstructed into individual trios. However, these bugs were reported and accounted for in the previous work[10] (personal communications with W. Brandler). Given this problem, we sought to determine parent of origin for inherited private variants with plink[63] using the *-tdt poo* command. The advantage of plink over GTDT, is (1) speed of execution (since plink leverages binary files) and (2) flexibility in the sense that we can test subsets of the cohort with simple input files (*-pheno* option allows to test in cases or controls). The group-wise test the then performed using an in-house python library, pytdt (github.com/dantaki/pytdt). This library was tested with the previous data set[10] and produced the same results as GTDT.

4.4 Results

4.4.1 De Novo Loss of Function Burden

Since many diagnoses of autism can be explained by a damaging de novo variant, we would like to test the effect of inherited risk in child that lack such a damaging variant. Hence, we collected de novo mutation predictions from our previous studies[11] and from previously published studies of the Simons Simplex Collection performed using exome sequencing and genotyping microarrays[33, 70, 34, 58, 41]. We then filtered the list to variants that were likely to cause a loss in function: stopgains or losses, frameshifts, splicing mutations, and exonic deletions. Then we further filtered the list of LoF de novo mutations to those that are in genes with pLI scores $\geq$0.9. We found an average of 0.1 damaging de novo mutations in cases compared to 0.03 in controls (Ttest p-value = 4.93×10^{-16}). The female protective effect is striking; affected females have more LoF burden than affected males (male mean = 0.09, female mean = 0.15, Ttest p-value = 1.89×10^{-4}; Figure 4.3). The difference between male and female controls was not significant

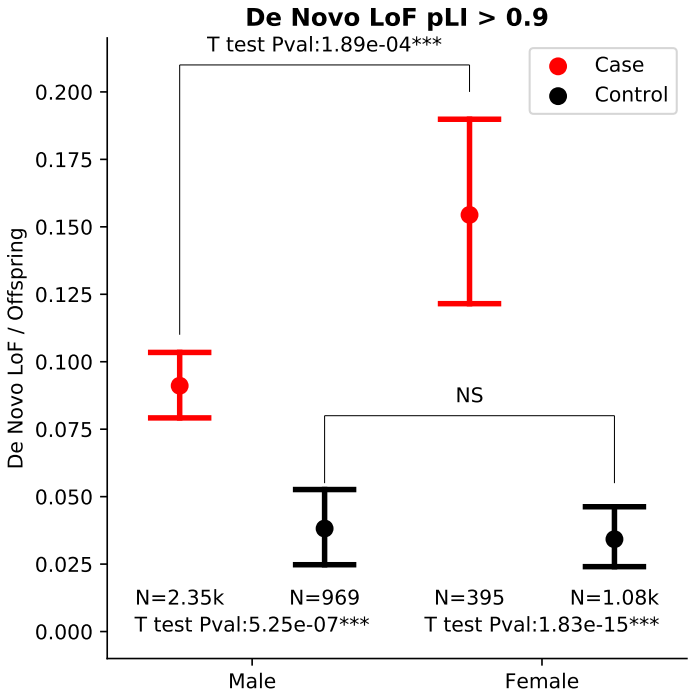

Figure 4.3: **Burden of Damaging De Novos in Autism.** We measured the burden of damaging de novo mutations defined as SNVs or INDELs or SVs that create a stopgain or loss, frameshift, splicing mutation, or coding deletion of haploinsuffcient genes (pLI $\geq$0.9). We used previously published data sets to annotate mutations in 2744 affected children and 2050 controls. Of these offspring 265 cases had one or more LoF de novo, while there were 69 control offspring with one or more damaging de novo. There is a striking female protective effect, since females carry more LoF burden (Ttest p-value = 1.89×10^{-4}). However, there was no difference of damaging burden between male and female controls (Ttest p-value = 0.653).

(male mean = 0.036, female mean = 0.038, Ttest p-value = 0.653; Figure 4.3). Of 2744 cases, 265 had one or more damaging de novo mutation, while 69 of the 2050 control offspring had one or more damaging de novo.

4.4.2 Both Fathers and Mothers Contribute Risk through Private Mutations

We measured the transmission of loss of function SNVs, INDELs, and SVs in haploinsufficient genes (pLI $\geq$0.9) in the combined REACH, SSC1-4 cohort (2650 families, 10,015

individuals). We found that mothers significantly over-transmitted these LoF variants to cases but not to controls (transmission rate = 0.538, binomial p-value = 1.99×10^{-3}, ChiSquare p-value = 0.0198), which has been previously reported[42, 35, 86]. However, in contrast to previous studies we also see a significant transmission of private LoF variants deriving from the paternal lineage (transmission rate = 0.548, binomial p-value = 1.1×10^{-4}, ChiSquare p-value = 9.6×10^{-5}; Figure 4.4A). This finding suggests revision is needed for the previous model where the mother solely contributes inherited risk.

Next, we measured the transmission of CRE-SVs associated with functionally constrained genes (Figure 4.4B). We found that fathers significantly over-transmitted CRE-SVs to cases (transmission rate = 0.567, binomial p-value = 0.048, ChiSquare p-value = 0.002), but under-transmitted CRE-SVs to controls (transmission rate = 0.401, binomial p-value = 0.013; Figure 4.4B). Likewise, we see a significant under-transmission of CRE-SV from mothers to controls (transmission rate = 0.386, binomial p-value = 0.011; Figure 4.4B). There was a slight under-transmission from mothers to affected offspring too (transmission rate = 0.442), similar to previous observations[10]. The observed under-transmission is intriguing, and maybe suggests that the ancestral allele is under strong selection operating as a selfish gene[30].

We selected potentially pathogenic missense variants with the aid of two datasets: PolyPhen-2[3] and ExAC[45] and tested their transmission rates (Figure 4.4C). PolyPhen-2 assigns a probability score of of pathogenicity to each missense variant. ExAC has provided a list of genes with missense intolerance Z scores, derived from the observed and expected number of missense mutations in that gene[45]. We used a threshold of $\geq$0.9 for PolyPhen-2 HumDiv scores and $\geq$5 sigma for missense Z scores from ExAC. For these potentially damaging variants we measured the transmission rate from parents and found no significant association (Figure 4.4C). However we observed a significant over-transmission of missense variants from the fathers (transmission rate = 0.528, binomial p-value = 0.018, ChiSquare p-value = 0.023).

We measured the effect of inherited rare damaging mutations (LoF, CRE-SVs, and

missense variants) as on autism. To maximize the effect of inherited risk, we excluded families with a damaging LoF de novo mutation in an affected child. We calculated odds ratios determined by transmission rates in affected and control offspring. Significance was determined by the Fischer's Exact test. We show that rare inherited LoF mutations in functionally constrained genes are strongly associated with autism (odds ratio = 1.29, 95% confidence interval = 1.14–1.44, p-value = 6.82×10^{-6}; Figure 4.4D), in concordance with previous studies[42]. Both fathers and mothers contribute inherited risk to autism via LoF mutations (paternal odds ratio = 1.36 95% confidence interval = 1.13–1.59, p-value = 0.0001; maternal odds ratio = 1.23, 95% confidence interval = 1.02–1.44, p-value = 0.011). The effect of LoF mutations is stronger when the variant is derived from the paternal lineage. Likewise, CRE-SVs are strongly associated with autism, but only from fathers (paternal odds ratio = 1.84, 95% confidence interval = 0.85–2.83, p-value = 0.006; Figure 4.4D). Missense variants from the paternal lineage are associated with autism but not maternal missense variants (paternal odds ratio = 1.19 95% confidence interval = 1.0–1.38, p-value = 0.026). When combined (LoF + CRE-SV + Missense) the inherited risk to autism is primarily derived from fathers (paternal odds ratio = 1.29 95% confidence interval = 1.15–1.43, p-value = 1.43×10^{-6}).

4.4.3 Fathers Primarily Contribute Risk to Sons

We then compared the transmission of private LoF variants in functionally constrained genes while stratifying on the offspring's gender. According to the previous model of the female protective effect, we should observe a significant transmission of LoF variants from mothers to affected sons[42, 35, 86]. In fact, we do observe a significant association from mothers to affected sons (maternal transmission rate = 0.527, binomial p-value = 0.04), but this effect is not as strong when compared to fathers and affected sons (paternal transmission rate = 0.552, binomial p-value = 8.5×10^{-5}, ChiSquare p-value = 9.5×10^{-4}; Figure 4.5A). Interestingly, mothers significantly transmit LoF variants to affected daughters (maternal transmission rate = 0.599,

binomial p-value = 0.002, ChiSquare p-value = 0.003; Figure 4.5A) but not fathers (paternal transmission rate = 0.519, binomial p-value = 0.629). We also tested if the maternal association to sons was due to the presence of LoF mutations on chromosome X, however when excluding variants on the sex chromosome, we see no difference in effect.

We then compared the transmission of private CRE-SVs in functionally constrained genes while stratifying on the gender of the offspring. Since we observed a significant sex-specific effect when leveraging LoF mutations where most of the autism risk to sons came from fathers and risk to daughters came from mothers, we then asked if a similar effect was present for CRE-SVs. We do observe a significant over-transmission of CRE-SVs from fathers to sons (transmission rate = 0.577, binomial p-value = 0.037, ChiSquare p-value = 0.0058; Figure 4.5B), but no over-transmission from mothers to daughters (Figure 4.5B). We do not observe a significant difference between cases and controls for the transmission of CRE-SVs from fathers to daughters. Our results would suggest that risk from CRE-SVs are more likely to confer risk to sons, but more data would be needed for that conclusion since the number of transmission events for fathers to daughters is only 37.

We compared the transmission of private potentially pathogenic missense SNVs while stratifying on the gender of the offspring. Since our previous results show signs of sex-specific associations, we explored the effect of gender has for the transmission of these missense variants. In contrast to our previous results, we do not observe a father to son bias nor a mother to daughter bias (Figure 4.5C). However, we do observe a significant father to daughter effect (transmission rate = 0.571, binomial p-value = 0.03, ChiSquare p-value = 0.049). Mothers seem to under-transmit damaging missense variants to affected daughters (transmission rate = 0.461) and over-transmit to control daughters (transmission rate = 0.53). However the difference between the two is marginal (ChiSquare p-value = 0.07). Our data suggests that fathers primarily contribute missense risk to daughters and potentially damaging mutations in mothers act in a protective manner, although those results are not significant.

4.4.4 Evidence for an Inherited Bilineal Model

We then tested whether a bilineal model could explain inherited autism risk by observing the transmission of private LoF and CRE-SV variants (Figure 4.6). We conditioned the test with respect to the other parent, requiring the other parent to have transmitted a LoF to an affected child. For example, we test the association of paternally inherited private LoF variants in families where the mother already transmitted a LoF mutation. We then combine the transmissions of fathers in families where the mother transmitted a damaging variant, and mothers in families where fathers transmitted a damaging variant. This condition allows us to test the association of inherited LoF and CRE-SV mutations from both parental lineages. Additionally, we removed families with a damaging de novo LoF mutation present in an affected offspring. In total, there is a significant over-transmission of private LoF and CRE-SVs to affected children (transmission rate = 0.539, binomial p-value = 1.99×10^{-6}, ChiSquare p-value = 2.46×10^{-7}; Figure 4.6). However, we found no association of private LoF variants when conditioned on the presence of a damaging de novo mutation in an affected child (Figure 4.6), which suggests that LoF de novo mutations are generally casual for autism in those children. When testing for the bilineal association, we observed a significant over-transmission of private LoF variants and CRE-SVs (transmission rate = 0.533, binomial p-value = 0.0465, ChiSquare p-value = 0.0453; Figure 4.6). These results suggest that inherited risk in the form of LoF and CRE-SVs contribute significantly to autism. There is a significant, not strong, but statistically significant contribution of bilineally inherited variants, where both parents contribute a damaging mutation. This observation might imply that the mutations in trans might be involved in similar functional pathways, and that the risk from both parents is additive[9].

4.4.5 Contribution of LoF, CRE-SV, and Missense Variants to Autism Risk

We then measured the effect of inherited rare damaging mutation in affected males and females separately (Figure 4.7). We found that the effect of LoF inherited mutations is strongest from fathers to sons (paternal odds ratio = 1.41, 95% confidence interval = 1.08–1.74, p-value = 0.0017) and from mothers to daughters (maternal odds ratio = 1.57, 95% confidence interval = 0.99–2.15, p-value = 0.005). Likewise, the effect of paternal CRE-SVs was strongest in male offspring (Figure 4.7). Interestingly, the effect of missense variants is strongest from fathers to daughters (paternal odds ratio = 1.43, 95% confidence intervals = 0.9-1.96, p-value = 0.026). When combined, inherited autism risk in males and females seems to be strongest from the fathers (father-to-son odds ratio = 1.3, 95% confidence interval = 1.11–1.49, p-value = 0.0002; father-to-daughter odds ratio = 1.3, 95% confidence interval = 0.98–1.62, p-value = 0.021). Although, mothers contribute significant inherited risk in the form of LoF mutations to sons and daughters.

4.5 Discussion

We sought to explore the effects of rare inherited damaging mutations in autism and show that there is a difference in how fathers and mother contribute to the disease. Both fathers and mothers contribute rare LoF mutations to affected children. Previously, it was only assumed that mothers contributed this risk[86], due to the observation that risk variants are less penetrant in females (”female protective effect”). The foundation for the female protective effect is rooted in the idea that females can tolerate more loss of function burden; hence, a de novo in a female is less likely to cause disease and could be transmitted to a son, who has lower tolerance for the mutation. In fact, our data suggests a female protective effect evident by the observation that affected females carry more de novo LoF mutations on average (Figure 4.3). These data imply that for a female to develop autism, there needs to be an accumulation of risk variants and that males require less burden to develop autistic traits. Such a models of additive and omnigenic

effects have been postulated for complex disorders like autism[9]. We did observe evidence for a bilineal model in families where both parents contribute a damaging private variant ("Bilineal 2nd Hit" Figure 4.6), and these results would imply an additive or omnigenic model. However, paternally inherited LoF mutations have greater effect than maternal LoFs (paternal odds ratio = 1.36 [1.13–1.59]; maternal odds ratio = 1.23 [1.02–1.44]; Figure 4.4D). These results could imply that males have a higher tolerate for risk variants than previously thought, since the fathers in our cohort are not autistic. However, it could be the case that fathers that carry these risk variants show some signs of autism, but we lack the appropriate phenotype data to directly test this hypothesis.

Furthermore we show that inherited LoF risk in sons is primarily derived from the paternal lineage, while for daughters most inherited LoF risk is inherited from the mothers (Figure 4.7). Mothers also contributed LoF risk to sons, but this effect was not as strong as father to son transmission. For daughters with autism, paternally inherited LoFs are not significantly associated with inherited risk (father-to-daughter odds ratio = 1.18 [0.71–1.65], p-value = 0.35). Given the observation that paternal LoF risk is primarily transmitted to sons and that these carrier fathers lack an autism diagnosis, we can propose an amendment to the female protective effect model which can account for these observations. A strong parent of origin effect for inherited variants suggests imprinting as a biological mechanism[43]. Strong parent of origin effects have been observed in human diseases, such as Bipolar Disorder where the authors suggest imprinting as a biological mechanism[74]. However, we do not know many imprinted regions of the human genome, save but a few examples such as the 15q11-13 locus responsible for Prader-Willi and Angelman disorder[48]. Previous studies have shown that the gene largely responsible for Angelman syndrome, *UBE3A*, exhibits maternal but not paternal expression in the hippocampus and cerebellum but not other tissues[80, 68]. Additional evidence for tissue-specific expression for imprinted genes have been reported for developing embryos[8] and in brains[20, 27]. Thus it's within the realm of possibility that many genes responsible for the development of brains

in humans are imprinted. If these genes are imprinted, then it would indicate that disruption of monoallelic expression could cause a disease phenotype. With this in mind, we can posit an explanation for the LoF burden in unaffected fathers: it could be the case that such fathers inherited these imprinted LoF variants from their mothers. Hence, maternal imprinting of these variants is protective against autism, explaining the apparent incomplete penetrance shown in carrier fathers. When this variant is transmitted from a carrier father, the maternal epigenetic protection is erased. Since males have decreased tolerance for LoF mutations, these paternally risk-imprinted variants confer greater risk to male offspring than female. Hence, daughters that inherited these variants can tolerant more burden and may not exhibit autistic phenotypes. However, they pose the risk of transmitting them to a son, who could later transmit a risk variant to future offspring. In order to test this theory, we would require multi-generational pedigrees to track the inheritance of these rare mutations. If paternal LoF variants are derived from his maternal haplotype, then this would give credence to an imprinting amendment to the female protective effect model.

CRE-SVs, on the other hand, exhibit a strong paternal bias to affected offspring while being under-transmitted from the mother. Such a pattern of inheritance also suggests imprinting[48] or possibly a selfish element[30]. Likewise, CRE-SVs were predominantly transmitted to affected children that did not inherit a damaging LoF mutation. This result suggests that CRE-SVs might confer more risk than previously thought[10]. However, damaging CRE-SVs are extremely rare, when compared to LoF or missense mutations (Figure 4.4D), so more data needs to be collected before reaching a solid conclusion on the effect of these variants. Oddly enough, missense variants exhibited a strong father to affected daughter bias and an over-transmission from mothers to controls. The effect of missense variants is not as strong as LoF or CRE-SVs (Figure 4.4D), but given the large number of variants to test ($>$5690 damaging missense variants) we can feel confident about our results. This distinct pattern of missense variants also suggests imprinting; maternally derived damaging missense mutations may be silenced, explaining the opposite association in

cases (conferring protection to controls), while paternally derived missense mutations are in genes that are paternally expressed. In all, imprinting may offer a harmonious explanation for the striking parent of origin effect we observe in our data with fathers contributing more epigenetic risk. However, before conclusions are drawn, a study into the effect of inherited variants over multiple generations is needed.

Chapter 4, in part, is currently being perpared for submissions for publication of this material. Danny Antaki, Madhusudan Gujral, Jonathan Sebat. Madhusudan Gujral assisted with variant calling and data processing. The disseration author is the primary investigator on this material, while Jonathan Sebat supervised the project and provided advice.

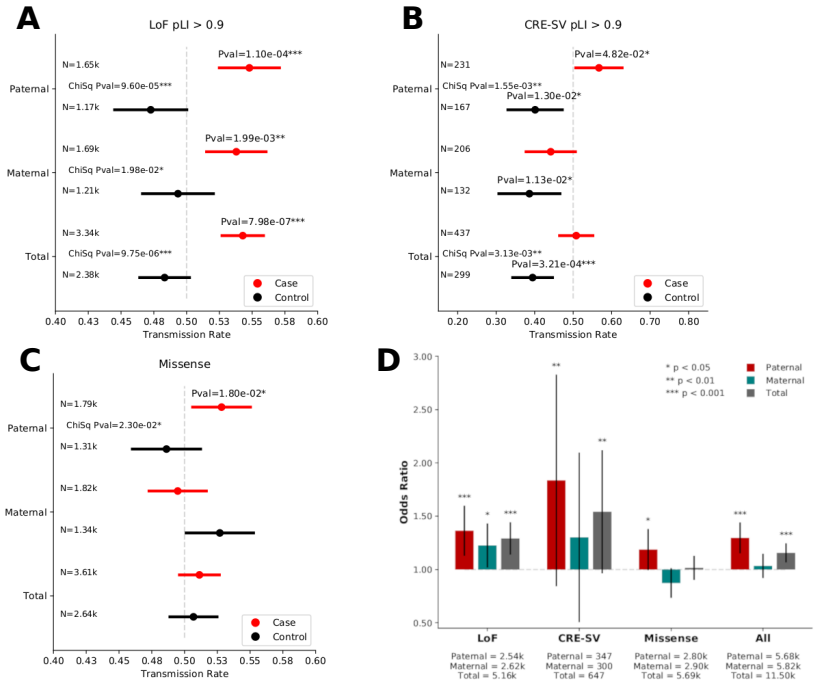

Figure 4.4: **Fathers and Mothers Contribute Inherited Risk to Autism.** Transmission disequilibrium tests for private SNVs, INDELs, and SVs that disrupt functionally constrained genes (A), deletions in functionally constrained cis-regulatory elements (B), and potentially damaging missense variants (C) stratified by parent of origin and combined ("Total"). We find significant over-transmission of LoF mutations (A) to cases but not to controls (paternal transmission rate = 0.548, maternal transmission rate = 0.538). Additionally, we find significant over-transmission of CRE-SVs (B) to cases but not to controls from the father but not the mother (paternal transmission rate = 0.567). Interestingly fathers and mothers under-transmit variants to controls (paternal transmission rate = 0.401, maternal transmission rate = 0.386). For potentially damaging missense variants (C), we found no association from parents to affected children, but observed a significant association from fathers to affected offspring (transmission rate = 0.528). We then measured the effect of rare inherited damaging LoF, CRE-SV, and missense mutations on autism (D). To maximize the effect of inherited risk, we excluded families with a damaging LoF de novo mutation in an affected child. Significance was defined by the Fischer's Exact test. Whiskers represent 95% confidence intervals. We show that rare inherited LoF mutations in functionally constrained genes are strongly associated with autism (odds ratio = 1.29, 95% confidence interval = 1.14–1.44, p-value = 6.82×10^{-6}). Both fathers and mothers contribute inherited risk to autism via LoF mutations (paternal odds ratio = 1.36 95% confidence interval = 1.13–1.59, p-value = 0.0001; maternal odds ratio = 1.23, 95% confidence interval = 1.02–1.44, p-value = 0.011). Likewise, paternal CRE-SVs are strongly associated with autism (paternal odds ratio = 1.84, 95% confidence interval = 0.85–2.83, p-value = 0.006). Missense variants from the paternal lineage are associated with autism but not maternal missense variants (paternal odds ratio = 1.19 95% confidence interval = 1.0–1.38, p-value = 0.026). When combined (LoF + CRE-SV + Missense) the inherited risk to autism is primarily derived from fathers (paternal odds ratio = 1.29 95% confidence interval = 1.15–1.43, p-value = 1.43×10^{-6}).

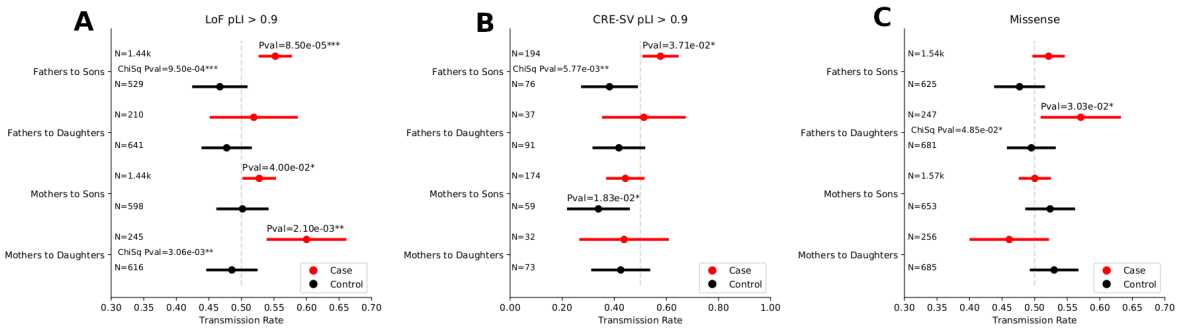

Figure 4.5: **Fathers Contribute More Risk to Autistic Sons than Mothers.** Transmission disequilibrium tests for private SNVs, INDELs, and SVs that disrupt functionally constrained genes (A), CRE-SVs (B), and potentially damaging missense variants (C), stratified by gender of offspring. Given the previous model of maternal inherited risk, we should expect an over-transmission of LoF variants from mothers to affected sons. Our results are in concordance with this model given the significant over-transmission we observed (maternal transmission rate = 0.527, binomial p-value = 0.04). However, this effect is not as strong as the father to affected son group (paternal transmission rate = 0.552, binomial p-value = 8.5×10^{-5}, ChiSquare p-value = 9.5×10^{-4}). Interestingly, we find a significant sex-specific association, also highlighted by the observation that mothers contribute risk to daughters (maternal transmission rate = 0.599, binomial p-value = 0.002, ChiSquare p-value = 0.003), but not fathers. We do observe a significant over-transmission of CRE-SVs (B) from fathers to sons (transmission rate = 0.577), but no over-transmission from mothers to daughters (transmission rate = 0.438, N= 32). We do not observe a significant difference between cases and controls for the transmission of CRE-SVs from fathers to daughters. Our results would suggest that risk from CRE-SVs are more likely to confer risk to sons. In contrast to LoF mutations and CRE-SVs, we do not observe a significant over-transmission of missense variants (C) from fathers to affected sons (transmission rate = 0.521) nor from mothers to affected daughters (transmission rate = 0.461). However, we do observe a significant father to daughter effect (transmission rate = 0.571), suggesting that missense risk is primarily derived from fathers to daughters.

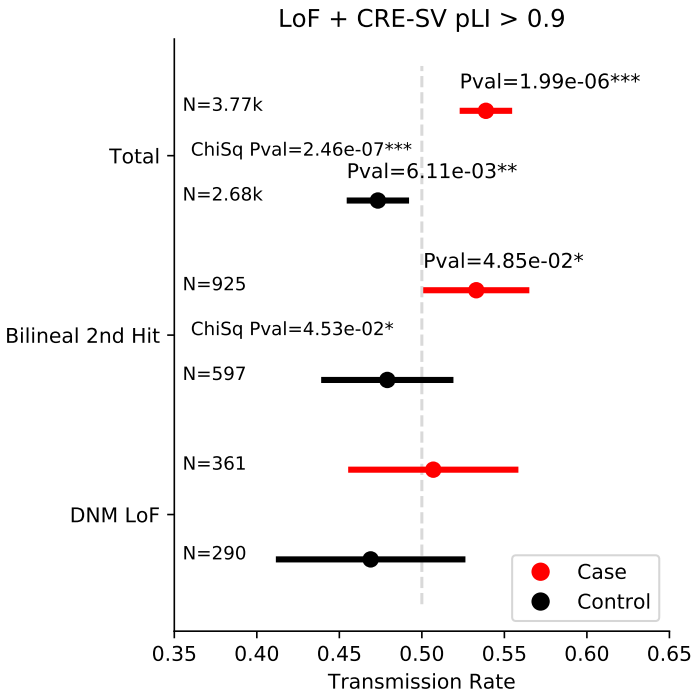

Figure 4.6: **Evidence for Inherited Bilineal Risk in Autism.** Transmission disequilibrium tests for private SNVs, INDELs, and SVs that disrupt functionally constrained genes and CRE-SVs. Private LoF and CRE-SVs in haploinsufficient genes are associated with autism ("Total"); both mothers and fathers contribute inherited risk (transmission rate = 0.539). We then test the hypothesis that inherited mutations in trans can contribute risk to autism. We conditioned these tests by measuring the transmission of private LoF and CRE-SVs in parents in families where the other parent transmitted a damaging variant. That is to say, what is the transmission of damaging variants from fathers in families where the mother transmitted a damaging variant to a case? We then total the number of transmissions from respective tests (fathers when mothers transmit damaging variants and vice versa) and show that there is a significant contribution of damaging inherited mutations in trans ("Bilineal 2nd Hit", transmission rate = 0.0533). However this effect is slightly significant, suggesting that most inherited risk might act in cis or additivity. Additionally, we show that children with a damaging de novo mutation do not inherit damaging variants more often than controls ("DNM LoF", transmission rate = 0.507, binomial p-value = 0.833). This implies that damaging de novo mutations are likely to be the main contributors of risk in those children.

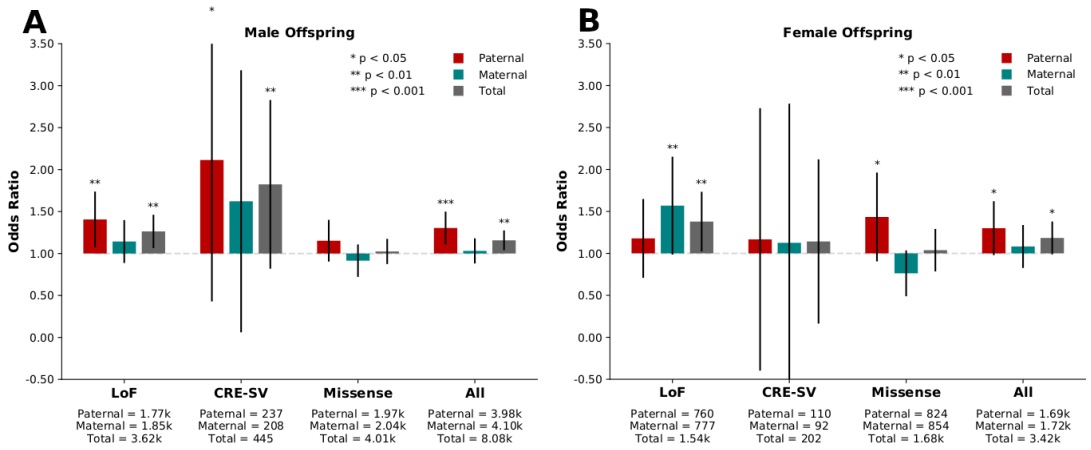

Figure 4.7: **Fathers Contribute More Inherited Risk to Autism than Mothers.** To maximize the effect of inherited risk, we excluded families with a damaging LoF de novo mutation in an affected child. Odds ratios were determined by transmission rates in affected and control offspring, and significance was defined by the Fischer's Exact test. Whiskers represent 95% confidence intervals. The effect of LoF inherited mutations is strongest from fathers to sons (paternal odds ratio = 1.41, 95% confidence interval = 1.08–1.74, p-value = 0.0017) and from mothers to daughters (maternal odds ratio = 1.57, 95% confidence interval = 0.99–2.15, p-value = 0.005). Interestingly, the effect of missense variants is greatest from fathers to daughters (paternal odds ratio = 1.43, 95% confidence intervals = 0.9-1.96, p-value = 0.026). When combined, inherited autism risk in males and females seems to be strongest from the fathers (father-to-son odds ratio = 1.3, 95% confidence interval = 1.11–1.49, p-value = 0.0002; father-to-daughter odds ratio = 1.3, 95% confidence interval = 0.98–1.62, p-value = 0.021). Although, mothers contribute significant inherited risk in the form of LoF mutations to sons and daughters.

Chapter 5

Discussion

Autism is a neurodevelopmental disorder marked with repetitive behaviors and impaired social interaction. Autism is has a strong genetic basis, evident by decades of reports on concordance of the disease between monozygotic twins[7] and recent investigations into mutations in coding regions[33] and large structural mutations found using genotyping microarrays[71, 69]. Genetically, autism is a complex disorder; there are many genes that contribute to the disorder with a wide range of risk, potentially 400–1000 genes may contribute risk to autism[24]. Large de novo mutations typically act in a dominant fashion with high penetrance that are unlikely to be transmitted given the low fecundity of autism[62] (0.25 children for males, 0.48 children for females, relative to general population). However only 10% of cases can be explained by de novo SVs[67, 12]. The remaining risk of de novo LoF mutations is approximately 9–15%[33, 67]. De novo missense variants can explain between 12-15% of autism[33, 67]. At best 30–37% of diagnoses can be attributed to a damaging de novo mutation[67, 12, 33]. Rare inherited risk is thought to explain 3% of cases[23]. With a large portion of heritability that is unexplained, and given the limited scope of previous studies (exome sequencing and genotyping arrays), we leveraged whole genome sequencing in order to associate noncoding mutations to autism. To this end, we amassed variant calls from over 10,000 whole genomes to test our hypothesis.

We first opted to search for structural variants in noncoding regions, since SVs are much larger than SNVs and INDELs they are more likely to elicit a functional change. However, detection of SV in whole genome sequence data carries a higher rate of false positives than SNVs and INDELs[11]. We then constructed a machine learning genotyper for SVs in whole genome sequence data to rectify SV calls. Our algorithm, SV^2 genotypes deletions and duplications with low false positive rates[5]. However, there is always room for improvement. Larger data sets are readily available which can supplement training and evaluation of newer models. Additionally, SV^2 is rather simple compared to more sophisticated machine learning models since SV^2 uses at most 3 different features of structural change in paired-end sequencing data. Hence, the addition of more informative features can potentially create a better model. Such features include confidence intervals of the SV position, GC context, deviations of coverage, strand information, mapping quality, and clipped reads. Likewise, SV detection by sequencing is migrating to a newer platform of longer single molecule reads. Longer reads allow for better alignment to repeat elements, in addition to entirely sequencing many transposable elements like LINEs. Thus, making a model that can either be platform blind or combine features from the two platforms would be ideal. One such solution might be convolutional neural networks that are trained on images. In fact, a deep learning model dubbed DeepVariant[61] outperforms GATK[51] at calling SNVs and INDELs. This model was trained on images that consist of a two dimensional matrix of pixels. Each row in the matrix is an alignment; each column is a nucleotide in the reference. The RGB channels of the alignments correspond to features such as strand orientation and mapping quality. With this in mind, we can create images of structural variants using alignments from either paired-end or single molecule read libraries. I have created a prototype that generates images (github.com/dantaki/SVanGogh) for SVs on either platform. The design of the two dimension matrix is still in use, except the image of the SV is a composite of two of these matrices, comprising the left and right breakpoint. One of the RGB channels encodes for clipped and split-reads, with positions of split-reads being pixelated at the highest RGB value (255). The

other channels correspond to mapping quality and strand orientation. I have prototype methods for all simple SVs: deletions, duplications, insertions, and inversions. Training on simple SVs would allow the model to extrapolate structural predictions on complex SVs, a powerful feature of convolutional neural networks. However, training a model to the required precision requires extremely large training sets on the order of hundreds of thousands to millions of examples. Simulation of SVs can aid in supplementing the needed data, but better simulation tools would be needed for single molecule libraries[83]. In all, the future of precise SV interrogation in paired-end and single molecule sequencing libraries is contingent on the development of newer methods, and deep learning may provide a tenable solution to this problem.

After creating a precise genotyping model for SVs[5], we then investigated the burden of SVs in cis-regulatory regions. At this time, annotating noncoding variants that do not impact genic elements, such as UTRs or promoters, is difficult. Previous attempts at defining evolutionary accelerated regions of the genome has produced a small list of loci[60]. Hence, we opted to test the association of noncoding variants that overlapped either 3'UTRs, promoters, or transcription start sites. If we assume autism risk not dependent on a recessive model, which is valid since most of our families have no history of the disorder, then genes that are dosage intolerant are more likely to harbor risk alleles. At our disposal was the ExAC[45] dataset which contains probability scores of likely to be intolerant to loss of function mutations (pLI) assigned to genes. The pLI score is a function that considers the expected and observed number of LoF mutations in the exomes of over 60,000 people. A threshold of $\geq$0.9 is recommended for disease association, since these genes are under extreme functional constraints. We tested the association of deletions that intersect cis-regulatory elements of functionally constrained genes and found that fathers but not mothers over-transmitted this class of mutation to affected children (Figure 3.6). Such a strong parent of origin effect implies imprinting[48] or a selfish gene system[30] as a possible biological mechanism. For example, deletion of CREs can lead to de-repression of imprinted genes[78]. However, an epigenetic mechanism could only explain our results if non-canonical imprinting

of regulatory elements is widespread. More data needs to be collected, but many studies have demonstrated that imprinting is differential throughout tissues[80, 68] and development[27]. Therefore, it's not outside the realm of possibility to imagine that the development of the human brain relies on the expression of monoallelic genes via parental imprinting, as seen in mice[73]. A carrier father might have inherited the mutation from a mother. The maternal imprinting acts in a protective manner, silencing the expression of the maternal copy in the son. However, since the paternal copy is expressed, when a carrier father transmits CRE-SVs, the protective epigenetic imprinting is erased and the offspring virtually is a knock-out for the tissue with monoallelic expression. To answer this question of a possible epigenetic mechanism for CRE-SVs, we would need to phase the variants with respect to the parent of origin for carrier fathers. This does not require whole genome sequencing, rather simple genotyping from PCR or targeted sequencing for problematic regions can determine the phase of variants, given access to the grandparents' genomes.

With over 10,000 whole genomes and variant calls at our disposal, we then measured the effect of rare inherited LoF mutations has on risk for autism. We measured the transmission of SNVs, INDELs, and SVs that disrupted coding regions of haploinsufficient genes and found that both mothers and fathers over-transmit damaging LoF mutations to affected offspring (Figure 4.4A). The association of maternally derived LoF mutations in autism has been previously reported[42], and a unifying model has been proposed by Sebat and Wigler[86] where can be explained autism by either highly penetrant de novo mutations or inherited LoF mutations from the maternal lineage. The rationale for the maternal origin model is that autistic females typically require a greater genetic load of burden to be affected[86], which is also seen in our own datasets (Figure 4.3). Thus is a damaging variant is de novo in a female, due to the female protective effect, she may not develop autism. But if she transmits that variant to a male offspring, his diminished tolerance for risk variants significantly increases the odds that he may develop autism. However, our data suggests that fathers also contribute inherited LoF risk to affected children. The female

protective effect model rests on the assumption that males should not tolerate these mutations, so what biological explanation can there be for a paternal LoF origin. The simplest explanation is that the fathers might be slightly affected. Autism is a spectrum disorder with a wide degree of intellectual capabilities[12, 64]; hence, carrier fathers might exhibit some autistic traits. We tested this hypothesis for parents with Social Responsiveness Scale (SRS) records and found no association with carrier status and SRS. These SRS scores were recorded by the spouse and age, cognitive level, and language capabilities can influence SRS scores[31]. We also considered the age of the parent at birth of the first child with the rationale that carrier fathers might delay childhood due to social impairments. However, we did not see any difference between carrier and non-carrier fathers or mothers, which is not surprising since many factors such as education, social economic status, and health influence when people decide to produce offspring. Without clinical grade phenotype data, we cannot test this hypothesis, but it should seriously be considered for future studies.

Another explanation as to why fathers carry these risk alleles and are not affected could be epigenetics. As outlined above, there is a lack of understanding of how monoallelic expression operates in developing humans. This epigenetic effect is observed in mice[27, 73] and there is some evidence that imprinted genes have tissue-specific expression in humans[80]. Therefore it's not absurd to suggest that some of these risk variants are imprinted in developing tissues. In fact, an epigenetic explanation can harmonize the female protective effect and our observation that (presumably) unaffected fathers carry LoF risk. Consider a LoF is in a gene that is silenced on the maternal haplotype, if a mother has a LoF in this gene and passes it along to a son, the copy with the LoF mutation is not expressed (the son expresses the wild-type paternal copy). However, the imprinting for maternal silencing is removed in the germ cells of the son. Therefore if he transmits this LoF mutation to a son, the offspring will express the gene on the paternal haplotype with the LoF mutation, since the offspring's maternal wild-type copy is silenced. Therefore a maternally transmitted LoF mutation in a maternally silenced gene confers risk for autism

when it is transmitted from male offspring in the next generation. As for female offspring of carrier fathers, they already have an increased tolerance for LoF mutations and can take on more burden. This is why we only see a significant effect of LoF mutations from fathers to sons and not from fathers to daughters (Figure 4.5A). Individually, LoF variants might contribute additive risk, since we observe support for a bilineal model (Figure 4.6); such an observation supports an additive or omnigenic model for autism[9]. Interestingly, for missense variants that are likely to be damaging, we find an opposite association where fathers transmit risk variants to daughters but not sons. Such a striking sex-dependent observation might indicate complex epigenetic effects such as sex-dependent imprinting, where the epigenetic effect is dependent on the gender of the individual. Such effects have been observed in mice[28], but have not truly explored in humans outside of clinical observation[6]. Additionally, mothers over-transmitted these missense variants to controls, suggesting a protective effect.

It is quite evident from all the above observations that autism is a very diverse disorder, genetically speaking. Previous studies have attributed de novo LoF variants to consist of 15% of diagnoses[12, 67] and de novo missense variants to consist of 12% of cases[33]. Studies into the effect of rare inherited LoF mutations in autism can explain about 12% of cases[42]. With larger cohorts (2703 affected children) and better tools such as whole genome sequencing, we show that rare inherited risk for autism can explain a significant component of missing heritability (25% of diagnoses; Figure 5.1). We find that 3% of cases inherited a LoF mutation from both parents (”Bilineal”). Likewise, CRE-SVs and damaging missense variants from fathers can explain 1% and 14% of cases respectively. Currently about 35% of cases remain unexplained, an improvement from 5 years ago. However, other approaches for explaining missing heritability need to be applied. For example, assigning polygenic risk scores of autism for individuals might explain many cases. Indeed, the combination of de novo, rare, and common mutations seem to all contribute risk to autism. Hence, it’s prudent to consider genomic effects rather than individual genic effects[9]. I feel confident that in the future researchers are going to be successful at

finding novel association for autism, further decreasing the proportion of cases with unexplained diagnoses. The integration of large datasets, rare and common variants, better phenotyping and cohort matching, and better algorithms for genetic analyses will ensure that the genetic mystery of autism will soon be a thing of the past.

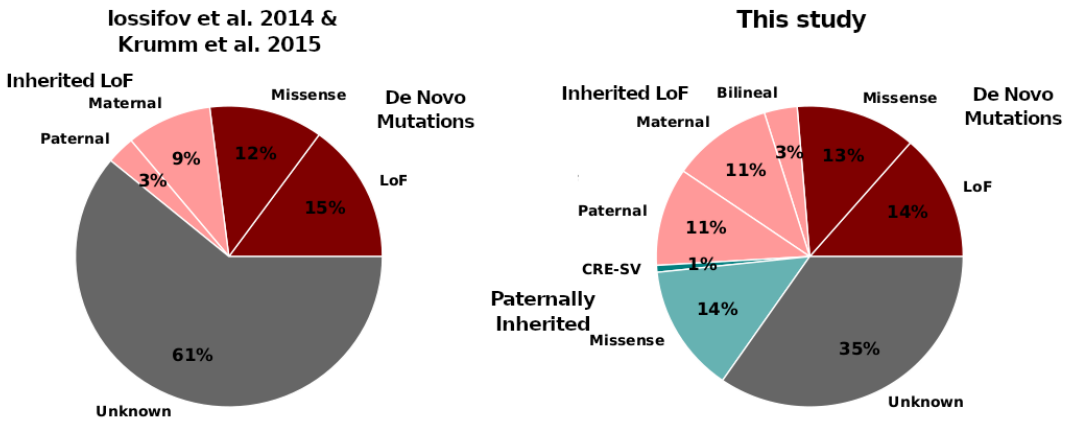

Figure 5.1: **Rare Inherited Variants can Explain a Significant Component of Missing Heritability.** Pie charts showing the percentage of autism diagnoses that can be explained by a casual variant for previous studies (left) and this study (right). Previous forays into the genetic causes of autism using exome sequencing and microarrays have attributed de novo mutations to about 30% of cases[33, 69]. This includes LoF mutations and missense mutations, which are thought to explain 15% and 12% of diagnoses respectively. Studies into the effect of rare inherited LoF mutations in autism can explain about 12% of cases[42]. With larger cohorts (2703 affected children) and better tools such as whole genome sequencing, we show that rare inherited risk for autism can explain a significant component of missing heritability (25% of diagnoses). We find that 3% of cases inherited a LoF mutation from both parents (”Bilineal”). Likewise, CRE-SVs and damaging missense variants from fathers can explain 1% and 14% of cases respectively. Currently about 35% of cases remain unexplained, an improvement from 5 years ago.

Bibliography

[1] Alexej Abyzov, Alexander E. Urban, Michael Snyder, and Mark Gerstein. Cnvnator: An approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. In *Genome Research*, 2011.

[2] Ivan A. Adzhubei, Daniel M. Jordan, and Shamil R. Sunyaev. Predicting functional effect of human missense mutations using polyphen-2. In *Current Protocols in Human Genetics*, 2014.

[3] Ivan A. Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. A method and server for predicting damaging missense mutations. In *Nature Methods*. Nature Publishing Group SN -, 2010.

[4] Kimberly A. Aldinger, Christianne J. Lane, Jeremy Veenstra-VanderWeele, and Pat Levitt. Patterns of risk for multiple co-occurring medical conditions replicate across distinct cohorts of children with autism spectrum disorder. In *Autism Research*, 2015.

[5] Danny Antaki, William M. Brandler, and Jonathan Sebat. Sv2: accurate structural variation genotyping and de novo mutation detection from whole genomes. In *Bioinformatics*, 2018.

[6] S Hasan Arshad, Wilfried Karmaus, Abid Raza, Ramesh J. Kurukulaaratchy, Sharon M. Matthews, John W. Holloway, Alireza Sadeghnejad, Hongmei Zhang, Graham Roberts, and Susan L. Ewart. The effect of parental allergy on childhood allergic diseases depends on the sex of the child. In *Journal of Allergy and Clinical Immunology*. Elsevier, 2012.

[7] A. Bailey, A. Le Couteur, I. Gottesman, P. Bolton, E. Simonoff, E. Yuzda, and M. Rutter. Autism as a strongly genetic disorder: evidence from a british twin study. In *Psychological Medicine*, 1995.

[8] S.C. Barton, A.C. Ferguson-Smith, R. Fundele, and M.A. Surani. Influence of paternally imprinted genes on development. In *Development*. The Company of Biologists Ltd, 1991.

[9] Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. An expanded view of complex traits: From polygenic to omnigenic. In *Cell*. Elsevier, 2017.

[10] William M. Brandler, Danny Antaki, Madhusudan Gujral, Morgan L. Kleiber, Joe Whitney, Michelle S. Maile, Oanh Hong, Timothy R. Chapman, Shirley Tan, Prateek Tandon, Timothy Pang, Shih C. Tang, Keith K. Vaux, Yan Yang, Eoghan Harrington, Sissel Juul, Daniel J. Turner, Bhooma Thiruvahindrapuram, Gaganjot Kaur, Zhuozhi Wang, Stephen F. Kingsmore, Joseph G. Gleeson, Denis Bisson, Boyko Kakaradov, Amalio Telenti, J. Craig Venter, Roser Corominas, Claudio Toma, Bru Cormand, Isabel Rueda, Silvina Guijarro, Karen S. Messer, Caroline M. Nievergelt, Maria J. Arranz, Eric Courchesne, Karen Pierce, Alysson R. Muotri, Lilia M. Iakoucheva, Amaia Hervas, Stephen W. Scherer, Christina Corsello, and Jonathan Sebat. Paternally inherited cis-regulatory structural variants are associated with autism. In *Science*. American Association for the Advancement of Science, 2018.

[11] William M. Brandler, Danny Antaki, Madhusudan Gujral, Amina Noor, Gabriel Rosanio, Timothy R. Chapman, Daniel J. Barrera, Guan Ning Lin, Dheeraj Malhotra, Amanda C. Watts, Lawrence C. Wong, Jasper A. Estabillo, Therese E. Gadomski, Oanh Hong, Karin V. Fuentes Fajardo, Abhishek Bhandari, Renius Owen, Michael Baughn, Jeffrey Yuan, Terry Solomon, Alexandra G. Moyzis, Michelle S. Maile, Stephan J. Sanders, Gail E. Reiner, Keith K. Vaux, Charles M. Strom, Kang Zhang, Alysson R. Muotri, Natacha Akshoomoff, Suzanne M. Leal, Karen Pierce, Eric Courchesne, Lilia M. Iakoucheva, Christina Corsello, and Jonathan Sebat. Frequency and complexity of de novo structural mutation in autism. In *The American Journal of Human Genetics*, 2016.

[12] William M. Brandler and Jonathan Sebat. From de novo mutations to personalized therapeutic interventions in autism. In *Annual Review of Medicine*, 2015.

[13] Jose Castresana. Estimation of genetic distances from human and mouse introns. In *Genome Biology*, 2002.

[14] Mark J.P. Chaisson, Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, Oscar Rodriguez, Li Guo, Ryan L. Collins, Xian Fan, Jia Wen, Robert E. Handsaker, Susan Fairley, Zev N. Kronenberg, Xiangmeng Kong, Fereydoun Hormozdiari, Dillon Lee, Aaron M. Wenger, Alex Hastie, Danny Antaki, Peter Audano, Harrison Brand, Stuart Cantsilieris, Han Cao, Eliza Cerveira, Chong Chen, Xintong Chen, Chen-Shan Chin, Zechen Chong, Nelson T. Chuang, Christine C. Lambert, Deanna M. Church, Laura Clarke, Andrew Farrell, Joey Flores, Timur Galeev, David Gorkin, Madhusudan Gujral, Victor Guryev, William Haynes Heaton, Jonas Korlach, Sushant Kumar, Jee Young Kwon, Jong Eun Lee, Joyce Lee, Wan-Ping Lee, Sau Peng Lee, Shantao Li, Patrick Marks, Karine Viaud-Martinez, Sascha Meiers, Katherine M. Munson, Fabio Navarro, Bradley J. Nelson, Conor Nodzak, Amina Noor, Sofia Kyriazopoulou-Panagiotopoulou, Andy Pang, Yunjiang Qiu, Gabriel Rosanio, Mallory Ryan, Adrian Stutz, Diana C.J. Spierings, Alistair Ward, AnneMarie E. Welch, Ming Xiao, Wei Xu, Chengsheng Zhang, Qihui Zhu, Xiangqun Zheng-Bradley, Ernesto Lowy, Sergei Yakneen, Steven McCarroll, Goo Jun, Li Ding, Chong Lek Koh, Bing Ren, Paul Flicek, Ken Chen, Mark B. Gerstein, Pui-Yan Kwok, Peter M. Lansdorp, Gabor Marth, Jonathan Sebat, Xinghua Shi, Ali Bashir, Kai Ye, Scott E. Devine, Michael Talkowski, Ryan E. Mills, Tobias Marschall, Jan O. Korbel,

Evan E. Eichler, and Charles Lee. Multi-platform discovery of haplotype-resolved structural variation in human genomes. In *biorxiv*, 2018.

[15] Rui Chen, Qiang Wei, Xiaowei Zhan, Xue Zhong, James S. Sutcliffe, Nancy J. Cox, Edwin H. Cook, Chun Li, Wei Chen, and Bingshan Li. A haplotype-based framework for group-wise transmission/disequilibrium tests for rare variant association analysis. In *Bioinformatics*, 2015.

[16] Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, and Christopher T. Saunders. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. In *Bioinformatics*, 2016.

[17] Colby Chiang, Ryan M. Layer, Gregory G. Faust, Michael R. Lindberg, David B. Rose, Erik P. Garrison, Gabor T. Marth, Aaron R. Quinlan, and Ira M. Hall. Speedseq: ultra-fast personal genome analysis and interpretation. In *Nature Methods*, 2015.

[18] Donald F. Conrad, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, T. Daniel Andrews, Chris Barnes, Peter Campbell, Tomas Fitzgerald, Min Hu, Chun Hwa Ihm, Kati Kristiansson, Daniel G. MacArthur, Jeffrey R. MacDonald, Ifejinelo Onyiah, Andy Wing Chun Pang, Sam Robson, Kathy Stirrups, Armand Valsesia, Klaudia Walter, John Wei, The Wellcome Trust Case Control Consortium, Chris Tyler-Smith, Nigel P. Carter, Charles Lee, Stephen W. Scherer, and Matthew E. Hurles. Origins and functional impact of copy number variation in the human genome. In *Nature*, 2009.

[19] The 1000 Genomes Project Consortium, Adam Auton, Gonçalo R. Abecasis, David M. Altshuler (Co-Chair), Richard M. Durbin (Co-Chair), David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs (Principal Investigator), Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Jun Wang (Principal Investigator), Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Eric S. Lander (Principal Investigator), David M. Altshuler, Stacey B. Gabriel (Co-Chair), Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Paul Flicek (Principal Investigator), Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, David R. Bentley (Principal Investigator), Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Hans Lehrach (Principal Investigator), Ralf Sudbrak (Project Leader),

Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie-Laure Yaspo, Elaine R. Mardis (Co-Principal Investigator) (Co-Chair), Richard K. Wilson (Co-Principal Investigator), Lucinda Fulton, Robert Fulton, Stephen T. Sherry (Principal Investigator), Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Gil A. McVean (Principal Investigator), Richard M. Durbin (Principal Investigator), Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Jeanette P. Schmidt (Principal Investigator), Christopher J. Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yiping Zhan, Adam Auton (Principal Investigator), Christopher L. Campbell, Yu Kong, Anthony Marcketta, Fuli Yu (Project Leader), Lilian Antunes, Matthew Bainbridge, Aniko Sabo, Zhuoyi Huang, Lachlan J. M. Coin, Lin Fang, Qibin Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Gabor T. Marth (Principal Investigator), Erik P. Garrison (Project Lead), Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Mark J. Daly (Principal Investigator), Mark A. DePristo (Project Leader), Robert E. Handsaker (Project Leader), Eric Banks, Gaurav Bhatia, Guillermo del Angel, Giulio Genovese, Heng Li, Seva Kashin, Steven A. McCarroll, James C. Nemesh, Ryan E. Poplin, Seungtai C. Yoon (Principal Investigator), Jayon Lihm, Vladimir Makarov, Andrew G. Clark (Principal Investigator), Srikanth Gottipati, Alon Keinan, Juan L. Rodriguez-Flores, Jan O. Korbel (Principal Investigator), Tobias Rausch (Project Leader), Markus H. Fritz, Adrian M. Stütz, Kathryn Beal, Avik Datta, Javier Herrero, Graham R. S. Ritchie, Daniel Zerbino, Pardis C. Sabeti (Principal Investigator), Ilya Shlyakhter, Stephen F. Schaffner, Joseph Vitti, David N. Cooper (Principal Investigator), Edward V. Ball, Peter D. Stenson, Bret Barnes, Markus Bauer, R. Keira Cheetham, Anthony Cox, Michael Eberle, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Eimear E. Kenny (Principal Investigator), Mark A. Batzer (Principal Investigator), Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur (Principal Investigator), Monkol Lek, Ralf Herwig, Elaine R. Mardis (Co-Principal Investigator), Li Ding, Daniel C. Koboldt, David Larson, Kai Ye, Simon Gravel, Anand Swaroop, Emily Chew, Tuuli Lappalainen (Principal Investigator), Yaniv Erlich (Principal Investigator), Melissa Gymrek, Thomas Frederick Willems, Jared T. Simpson, Mark D. Shriver (Principal Investigator), Jeffrey A. Rosenfeld (Principal Investigator), Carlos D. Bustamante (Principal Investigator), Stephen B. Montgomery (Principal Investigator), Francisco M. De La Vega (Principal Investigator), Jake K. Byrnes, Andrew W. Carroll, Marianne K. DeGorter, Phil Lacroute, Brian K. Maples, Alicia R. Martin, Andres Moreno-Estrada, Suyash S. Shringarpure, Fouad Zakharia, Eran Halperin (Principal Investigator), Yael Baran, Charles Lee (Principal Investigator), Eliza Cerveira, Jaeho Hwang, Ankit Malhotra (Co-Project Lead), Dariusz Plewczynski, Kamen Radew, Mallory Romanovitch, Chengsheng Zhang (Co-Project Lead), Fiona C. L. Hyland, David W. Craig (Principal Investigator), Alexis Christoforides, Nils Homer, Tyler Izatt, Ahmet A.

Kurdoglu, Shripad A. Sinari, Kevin Squire, Chunlin Xiao, Jonathan Sebat (Principal Investigator), Danny Antaki, Madhusudan Gujral, Amina Noor, Kenny Ye, Esteban G. Burchard (Principal Investigator), Ryan D. Hernandez (Principal Investigator), Christopher R. Gignoux, David Haussler (Principal Investigator), Sol J. Katzman, W. James Kent, Bryan Howie, Andres Ruiz-Linares (Principal Investigator), Emmanouil T. Dermitzakis (Principal Investigator), Scott E. Devine (Principal Investigator), Gonçalo R. Abecasis (Principal Investigator) (Co-Chair), Hyun Min Kang (Project Leader), Jeffrey M. Kidd (Principal Investigator), Tom Blackwell, Sean Caron, Wei Chen, Sarah Emery, Lars Fritsche, Christian Fuchsberger, Goo Jun, Bingshan Li, Robert Lyons, Chris Scheller, Carlo Sidore, Shiya Song, Elzbieta Sliwerska, Daniel Taliun, Adrian Tan, Ryan Welch, Mary Kate Wing, Xiaowei Zhan, Philip Awadalla (Principal Investigator), Alan Hodgkinson, Yun Li, Xinghua Shi (Principal Investigator), Andrew Quitadamo, Gerton Lunter (Principal Investigator), Gil A. McVean (Principal Investigator) (Co-Chair), Jonathan L. Marchini (Principal Investigator), Simon Myers (Principal Investigator), Claire Churchhouse, Olivier Delaneau, Anjali Gupta-Hinch, Warren Kretzschmar, Zamin Iqbal, Iain Mathieson, Androniki Menelaou, Andy Rimmer, Dionysia K. Xifara, Taras K. Oleksyk (Principal Investigator), Yunxin Fu (Principal Investigator), Xiaoming Liu, Momiao Xiong, Lynn Jorde (Principal Investigator), David Witherspoon, Jinchuan Xing, Evan E. Eichler (Principal Investigator), Brian L. Browning (Principal Investigator), Sharon R. Browning (Principal Investigator), Fereydoun Hormozdiari, Peter H. Sudmant, Ekta Khurana (Principal Investigator), Matthew E. Hurles (Principal Investigator), Chris Tyler-Smith (Principal Investigator), Cornelis A. Albers, Qasim Ayub, Yuan Chen, Vincenza Colonna, Luke Jostins, Klaudia Walter, Yali Xue, Mark B. Gerstein (Principal Investigator), Alexej Abyzov, Suganthi Balasubramanian, Jieming Chen, Declan Clarke, Yao Fu, Arif O. Harmanci, Mike Jin, Donghoon Lee, Jeremy Liu, Xinmeng Jasmine Mu, Jing Zhang, Yan Zhang, Erik P. Garrison, Steven A. McCarroll (Principal Investigator), Chris Hartl, Khalid Shakir, Jeremiah Degenhardt, Jan O. Korbel (Principal Investigator) (Co-Chair), Sascha Meiers, Benjamin Raeder, Tobias Rausch, Francesco Paolo Casale, Oliver Stegle, Eric-Wubbo Lameijer, Li Ding (Principal Investigator), Ira Hall, Charles Lee (Principal Investigator) (Co-Chair), Ankit Malhotra, Chengsheng Zhang, Vineet Bafna, Jacob Michaelson, Eugene J. Gardner (Project Leader), Gonçalo R. Abecasis (Principal Investigator), Ryan E. Mills (Principal Investigator), Gargi Dayama, Ken Chen (Principle Investigator), Xian Fan, Zechen Chong, Tenghui Chen, Evan E. Eichler (Principal Investigator) (Co-Chair), Mark J. Chaisson, John Huddleston, Maika Malig, Bradley J. Nelson, Nicholas F. Parrish, Ben Blackburne, Sarah J. Lindsay, Zemin Ning, Yujun Zhang, Hugo Lam, Cristina Sisu, Richard A. Gibbs (Principal Investigator) (Co-Chair), Danny Challis, Uday S. Evani, James Lu, Uma Nagaswamy, Jin Yu, Wangshen Li, Gabor T. Marth (Principal Investigator) (Co-Chair), Mark A. DePristo, Elaine R. Mardis (Principal Investigator), Hyun Min Kang, Lukas Habegger, Haiyuan Yu (Principal Investigator), Fiona Cunningham, Ian Dunham, Kasper Lage (Principal Investigator), Jakob Berg Jespersen, Heiko Horn, Chris Tyler-Smith (Principal Investigator) (Co-Chair), Mark B. Gerstein (Principal Investigator) (Co-Chair), Donghoon Kim, Rob Desalle, Apurva Narechania, Melissa A. Wilson Sayres, Robert E. Handsaker, Yaniv Erlich, Carlos D. Bustamante (Principal Investigator) (Co-Chair), Fernando L. Mendez, G. David Poznik, Peter A. Under-

hill, Lachlan Coin (Principal Investigator), David Mittelman, Ruby Banerjee, Maria Cerezo, Thomas W. Fitzgerald, Sandra Louzada, Andrea Massaia, Graham R. Ritchie, Fengtang Yang, Divya Kalra, Walker Hale, Xu Dan, Paul Flicek (Principal Investigator) (Co-Chair), Laura Clarke (Project Lead), Ralf Sudbrak (Project Lead), Stephen T. Sherry (Principal Investigator) (Co-Chair), Aravinda Chakravarti (Co-Chair), Bartha M. Knoppers (Co-Chair), Kathleen C. Barnes, Christine Beiswanger, Esteban G. Burchard, Carlos D. Bustamante, Hongyu Cai, Hongzhi Cao, Richard M. Durbin, Brenna Henn, Danielle Jones, Lynn Jorde, Jane S. Kaye, Alastair Kent, Angeliki Kerasidou, Rasika Mathias, Pilar N. Ossorio, Michael Parker, Charles N. Rotimi, Charmaine D. Royal, Karla Sandoval, Yeyang Su, Ralf Sudbrak, Zhongming Tian, Sarah Tishkoff, Chris Tyler-Smith, Marc Via, Yuhong Wang, Huanming Yang, Ling Yang, Jiayong Zhu, Walter Bodmer, Gabriel Bedoya, Andres Ruiz-Linares, Zhiming Cai, Yang Gao, Jiayou Chu, Leena Peltonen, Andres Garcia-Montero, Alberto Orfao, Julie Dutil, Juan C. Martinez-Cruzado, Taras K. Oleksyk, Rasika A. Mathias, Anselm Hennis, Harold Watson, Colin McKenzie, Firdausi Qadri, Regina LaRocque, Pardis C. Sabeti, Xiaoyan Deng, Danny Asogun, Onikepe Folarin, Christian Happi, Omonwunmi Omoniwa, Matt Stremlau, Ridhi Tariyal, Muminatou Jallow, Fatoumatta Sisay Joof, Tumani Corrah, Kirk Rockett, Dominic Kwiatkowski, Jaspal Kooner, Trâ‘n Tinh Hiê‘n, Sarah J. Dunstan, Nguyen Thuy Hang, Richard Fonnie, Robert Garry, Lansana Kanneh, Lina Moses, John Schieffelin, Donald S. Grant, Carla Gallo, Giovanni Poletti, Danish Saleheen, Asif Rasheed, Lisa D. Brooks, Adam L. Felsenfeld, Jean E. McEwen, Yekaterina Vaydylevich, Audrey Duncanson, Michael Dunn, Jeffery A. Schloss, and Jonathan L. Marchini. A global reference for human genetic variation. In *Nature*, 2015.

[20] William Davies, Anthony R. Isles, and Lawrence S. Wilkinson. Imprinted gene expression in the brain. In *Neuroscience & Biobehavioral Reviews*, 2005.

[21] Ryan N. Doan, Byoung-Il Bae, Beatriz Cubelos, Cindy Chang, Amer A. Hossain, Samira Al-Saad, Nahit M. Mukaddes, Ozgur Oner, Muna Al-Saffar, Soher Balkhy, Generoso G. Gascon, Marta Nieto, and Christopher A. Walsh. Mutations in human accelerated regions disrupt cognition and social behavior. In *Cell*. Elsevier, 2016.

[22] Mayada Elsabbagh, Gauri Divan, Yun-Joo Koh, Young Shin Kim, Shuaib Kauchali, Carlos Marcín, Cecilia Montiel-Nava, Vikram Patel, Cristiane S. Paula, Chongying Wang, Mohammad Taghi Yasamy, and Eric Fombonne. Global prevalence of autism and other pervasive developmental disorders. In *Autism Research*, 2012.

[23] Trent Gaugler, Lambertus Klei, Stephan J. Sanders, Corneliu A. Bodea, Arthur P. Goldberg, Ann B. Lee, Milind Mahajan, Dina Manaa, Yudi Pawitan, Jennifer Reichert, Stephan Ripke, Sven Sandin, Pamela Sklar, Oscar Svantesson, Abraham Reichenberg, Christina M. Hultman, Bernie Devlin, Kathryn Roeder, and Joseph D. Buxbaum. Most genetic risk for autism resides with common variation. In *Nature Genetics*, 2014.

[24] Daniel H. Geschwind and Matthew W. State. Gene hunting in autism spectrum disorder: on the path to precision medicine. In *The Lancet Neurology*, 2015.

[25] Santhosh Girirajan, Jill A. Rosenfeld, Gregory M. Cooper, Francesca Antonacci, Priscillia Siswara, Andy Itsara, Laura Vives, Tom Walsh, Shane E. McCarthy, Carl Baker, Heather C. Mefford, Jeffrey M. Kidd, Sharon R. Browning, Brian L. Browning, Diane E. Dickel, Deborah L. Levy, Blake C. Ballif, Kathryn Platky, Darren M. Farber, Gordon C. Gowans, Jessica J. Wetherbee, Alexander Asamoah, David D. Weaver, Paul R. Mark, Jennifer Dickerson, Bhuwan P. Garg, Sara A. Ellingwood, Rosemarie Smith, Valerie C. Banks, Wendy Smith, Marie T. McDonald, Joe J. Hoo, Beatrice N. French, Cindy Hudson, John P. Johnson, Jillian R. Ozmore, John B. Moeschler, Urvashi Surti, Luis F. Escobar, Dima El-Khechen, Jerome L. Gorski, Jennifer Kussmann, Bonnie Salbert, Yves Lacassie, Alisha Biser, Donna M. McDonald-McGinn, Elaine H. Zackai, Matthew A. Deardorff, Tamim H. Shaikh, Eric Haan, Kathryn L. Friend, Marco Fichera, Corrado Romano, Jozef Gacz, Lynn E. DeLisi, Jonathan Sebat, Mary-Claire King, Lisa G. Shaffer, and Evan E. Eichler. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. In *Nature Genetics*. Nature Publishing Group SN -, 2010.

[26] Joseph T. Glessner, Kai Wang, Guiqing Cai, Olena Korvatska, Cecilia E. Kim, Shawn Wood, Haitao Zhang, Annette Estes, Camille W. Brune, Jonathan P. Bradfield, Marcin Imielinski, Edward C. Frackelton, Jennifer Reichert, Emily L. Crawford, Jeffrey Munson, Patrick M. A. Sleiman, Rosetta Chiavacci, Kiran Annaiah, Kelly Thomas, Cuiping Hou, Wendy Glaberson, James Flory, Frederick Otieno, Maria Garris, Latha Soorya, Lambertus Klei, Joseph Piven, Kacie J. Meyer, Evdokia Anagnostou, Takeshi Sakurai, Rachel M. Game, Danielle S. Rudd, Danielle Zurawiecki, Christopher J. McDougle, Lea K. Davis, Judith Miller, David J. Posey, Shana Michaels, Alexander Kolevzon, Jeremy M. Silverman, Raphael Bernier, Susan E. Levy, Robert T. Schultz, Geraldine Dawson, Thomas Owley, William M. McMahon, Thomas H. Wassink, John A. Sweeney, John I. Nurnberger, Hilary Coon, James S. Sutcliffe, Nancy J. Minshew, Struan F. A. Grant, Maja Bucan, Edwin H. Cook, Joseph D. Buxbaum, Bernie Devlin, Gerard D. Schellenberg, and Hakon Hakonarson. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. In *Nature*, 2009.

[27] Christopher Gregg, Jiangwen Zhang, James E. Butler, David Haig, and Catherine Dulac. Sex-specific parent-of-origin allelic expression in the mouse brain. In *Science*. American Association for the Advancement of Science, 2010.

[28] Reinmar Hager, James M. Cheverud, Larry J. Leamy, and Jason B. Wolf. Sex dependent imprinting effects on complex traits in mice. In *BMC Evolutionary Biology*, 2008.

[29] Michael M. Hoffman and Ewan Birney. Estimating the neutral rate of nucleotide substitution using introns. In *Molecular Biology and Evolution*, 2007.

[30] Gregory D. D. Hurst and John H. Werren. The role of selfish genetic elements in eukaryotic evolution. In *Nature Reviews Genetics*. Nature Publishing Group SN -, 2001.

[31] Vanessa Hus, Somer Bishop, Katherine Gotham, Marisela Huerta, and Catherine Lord.

Factors influencing scores on the social responsiveness scale. In *Journal Child Psychology Psychiatry*, 2013.

[32] J. W. IJdo, A. Baldini, D. C. Ward, S. T. Reeders, and R. A. Wells. Origin of human chromosome 2: an ancestral telomere-telomere fusion. In *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 1991.

[33] Ivan Iossifov, Brian J. O'Roak, Stephan J. Sanders, Michael Ronemus, Niklas Krumm, Dan Levy, Holly A. Stessman, Kali T. Witherspoon, Laura Vives, Karynne E. Patterson, Joshua D. Smith, Bryan Paeper, Deborah A. Nickerson, Jeanselle Dea, Shan Dong, Luis E. Gonzalez, Jeffrey D. Mandell, Shrikant M. Mane, Michael T. Murtha, Catherine A. Sullivan, Michael F. Walker, Zainulabedin Waqar, Liping Wei, A. Jeremy Willsey, Boris Yamrom, Yoon-ha Lee, Ewa Grabowska, Ertugrul Dalkic, Zihua Wang, Steven Marks, Peter Andrews, Anthony Leotta, Jude Kendall, Inessa Hakker, Julie Rosenbaum, Beicong Ma, Linda Rodgers, Jennifer Troge, Giuseppe Narzisi, Seungtai Yoon, Michael C. Schatz, Kenny Ye, W. Richard McCombie, Jay Shendure, Evan E. Eichler, Matthew W. State, and Michael Wigler. The contribution of de novo coding mutations to autism spectrum disorder. In *Nature*, 2014.

[34] Ivan Iossifov, Michael Ronemus, Dan Levy, Zihua Wang, Inessa Hakker, Julie Rosenbaum, Boris Yamrom, Yoon-ha Lee, Giuseppe Narzisi, Anthony Leotta, Jude Kendall, Ewa Grabowska, Beicong Ma, Steven Marks, Linda Rodgers, Asya Stepansky, Jennifer Troge, Peter Andrews, Mitchell Bekritsky, Kith Pradhan, Elena Ghiban, Melissa Kramer, Jennifer Parla, Ryan Demeter, Lucinda L Fulton, Robert S Fulton, Vincent J Magrini, Kenny Ye, Jennifer C Darnell, Robert B Darnell, Elaine R Mardis, Richard K Wilson, Michael C Schatz, W. Richard McCombie, and Michael Wigler. De novo gene disruptions in children on the autistic spectrum. In *Neuron*. Elsevier, 2012.

[35] Sébastien Jacquemont, Bradley P. Coe, Micha Hersch, Michael H. Duyzend, Niklas Krumm, Sven Bergmann, Jacques S. Beckmann, Jill A. Rosenfeld, and Evan E. Eichler. A higher mutational burden in females supports a female protective model in neurodevelopmental disorders. In *The American Journal of Human Genetics*. Elsevier, 2014.

[36] Alec J. Jeffreys and Rita Neumann. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. In *Nature Genetics*. Nature Publishing Group, 2002.

[37] Mehran Karimzadeh, Carl Ernst, Anshul Kundaje, and Michael M. Hoffman. Umap and bismap: quantifying genome and methylome mappability. In *Nucleic Acids Research*, 2018.

[38] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent. The ucsc genome browser database. In *Nucleic Acids Research*. Oxford University Press, 2003.

[39] Donna Karolchik, Angela S. Hinrichs, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet, David Haussler, and W. James Kent. The ucsc table browser data retrieval tool. In *Nucleic Acids Research*, 2004.

[40] Eric B. Keverne, Reinald Fundele, Maithreyi Narasimha, Sheila C. Barton, and M.Azim Surani. Genomic imprinting and the differential roles of parental genomes in brain development. In *Developmental Brain Research*, 1996.

[41] Niklas Krumm, Peter H. Sudmant, Arthur Ko, Brian J. O'Roak, Maika Malig, Bradley P. Coe, NHLBI Exome Sequencing Project, Aaron R. Quinlan, Deborah A. Nickerson, and Evan E. Eichler. Copy number variation detection and genotyping from exome sequence data. In *Genome Research*. Cold Spring Harbor Laboratory Press, 2012.

[42] Niklas Krumm, Tychele N. Turner, Carl Baker, Laura Vives, Kiana Mohajeri, Kali Witherspoon, Archana Raja, Bradley P. Coe, Holly A. Stessman, Zong-Xiao He, Suzanne M. Leal, Raphael Bernier, and Evan E. Eichler. Excess of rare, inherited truncating mutations in autism. In *Nature Genetics*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. SN -, 2015.

[43] Heather A. Lawson, James M. Cheverud, and Jason B. Wolf. Genomic imprinting and parent-of-origin effects on complex traits. In *Nature Reviews Genetics*, 2013.

[44] Ryan M. Layer, Colby Chiang, Aaron R. Quinlan, and Ira M. Hall. Lumpy: a probabilistic framework for structural variant discovery. In *Genome Biology*, 2014.

[45] Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, James S. Ware, Andrew J. Hill, Beryl B. Cummings, Taru Tukiainen, Daniel P. Birnbaum, Jack A. Kosmicki, Laramie E. Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N. Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I. Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M. Peloso, Ryan Poplin, Manuel A. Rivas, Valentin Ruano-Rubio, Samuel A. Rose, Douglas M. Ruderfer, Khalid Shakir, Peter D. Stenson, Christine Stevens, Brett P. Thomas, Grace Tiao, Maria T. Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M. Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C. Florez, Stacey B. Gabriel, Gad Getz, Stephen J. Glatt, Christina M. Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M. Neale, Aarno Palotie, Shaun M. Purcell, Danish Saleheen, Jeremiah M. Scharf, Pamela Sklar, Patrick F. Sullivan, Jaakko Tuomilehto, Ming T. Tsuang, Hugh C. Watkins, James G. Wilson, Mark J. Daly, Daniel G. MacArthur, and Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. In *Nature*, 2016.

[46] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. In *Bioinformatics*, 2009.

[47] H. A. Lubs. A marker x chromosome. In *The American Journal of Human Genetics*, 1969.

[48] Angela M. Mabb, Matthew C. Judson, Mark J. Zylka, and Benjamin D. Philpot. Angelman syndrome: insights into genomic imprinting and neurodevelopmental phenotypes. In *Trends in Neurosciences*. Elsevier, 2011.

[49] Dheeraj Malhotra and Jonathan Sebat. Cnvs: Harbingers of a rare variant revolution in psychiatric genetics. In *Cell*, 2012.

[50] Christian R. Marshall, Abdul Noor, John B. Vincent, Anath C. Lionel, Lars Feuk, Jennifer Skaug, Mary Shago, Rainald Moessner, Dalila Pinto, Yan Ren, Bhooma Thiruvahindrapduram, Andreas Fiebig, Stefan Schreiber, Jan Friedman, Cees E.J. Ketelaars, Yvonne J. Vos, Can Ficicioglu, Susan Kirkpatrick, Rob Nicolson, Leon Sloman, Anne Summers, Clare A. Gibbons, Ahmad Teebi, David Chitayat, Rosanna Weksberg, Ann Thompson, Cathy Vardy, Vicki Crosbie, Sandra Luscombe, Rebecca Baatjes, Lonnie Zwaigenbaum, Wendy Roberts, Bridget Fernandez, Peter Szatmari, and Stephen W. Scherer. Structural variation of chromosomes in autism spectrum disorder. In *The American Journal of Human Genetics*, 2008.

[51] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. In *Genome Research*, 2010.

[52] James C. McPartland, Brian Reichow, and Fred R. Volkmar. Sensitivity and specificity of proposed dsm-5 diagnostic criteria for autism spectrum disorder. In *Journal of the American Academy of Child & Adolescent Psychiatry*, 2012.

[53] Heather C. Mefford, Andrew J. Sharp, Carl Baker, Andy Itsara, Zhaoshi Jiang, Karen Buysse, Shuwen Huang, Viv K. Maloney, John A. Crolla, Diana Baralle, Amanda Collins, Catherine Mercer, Koen Norga, Thomy de Ravel, Koen Devriendt, Ernie M.H.F. Bongers, Nicole de Leeuw, William Reardon, Stefania Gimelli, Frederique Bena, Raoul C. Hennekam, Alison Male, Lorraine Gaunt, Jill Clayton-Smith, Ingrid Simonic, Soo Mi Park, Sarju G. Mehta, Serena Nik-Zainal, C. Geoffrey Woods, Helen V. Firth, Georgina Parkin, Marco Fichera, Santina Reitano, Mariangela Lo Giudice, Kelly E. Li, Iris Casuga, Adam Broomer, Bernard Conrad, Markus Schwerzmann, Lorenz Räber, Sabina Gallati, Pasquale Striano, Antonietta Coppola, John L. Tolmie, Edward S. Tobias, Chris Lilley, Lluis Armengol, Yves Spysschaert, Patrick Verloo, Anja De Coene, Linde Goossens, Geert Mortier, Frank Speleman, Ellen van Binsbergen, Marcel R. Nelen, Ron Hochstenbach, Martin Poot, Louise Gallagher, Michael Gill, Jon McClellan, Mary-Claire King, Regina Regan, Cindy Skinner, Roger E. Stevenson, Stylianos E. Antonarakis, Caifu Chen, Xavier Estivill, Björn Menten, Giorgio Gimelli, Susan Gribble, Stuart Schwartz, James S. Sutcliffe, Tom Walsh, Samantha J.L. Knight, Jonathan Sebat, Corrado Romano, Charles E. Schwartz, Joris A. Veltman, Bert B.A. de Vries, Joris R. Vermeesch, John C.K. Barber, Lionel Willatt, May Tassabehji, and Evan E. Eichler. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. In *New England Journal of Medicine*, 2008.

[54] Jacob J. Michaelson and Jonathan Sebat. forestsv: structural variant discovery through statistical learning. In *Nature Methods*, 2012.

[55] Jacob J. Michaelson, Yujian Shi, Madhusudan Gujral, Hancheng Zheng, Dheeraj Malhotra, Xin Jin, Jian Minghan, Guangming Liu, Douglas Greer, Abhishek Bhandari, Wenting Wu, Roser Corominas, Áine Peoples, Amnon Koren, Athurva Gore, Shuli Kang, Guan Ning Lin, Jasper Estabillo, Therese Gadomski, Balvindar Singh, Kun Zhang, Natacha Akshoomoff, Christina Corsello, Steven McCarroll, Lilia M. Iakoucheva, Yingrui Li, Jun Wang, and Jonathan Sebat. Whole genome sequencing in autism identifies hotspots for de novo germline mutation. In *Cell*, 2012.

[56] Vagheesh Narasimhan, Petr Danecek, Aylwyn Scally, Yali Xue, Chris Tyler-Smith, and Richard Durbin. Bcftools/roh: a hidden markov model approach for detecting autozygosity from next-generation sequencing data. In *Bioinformatics*, 2016.

[57] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. In *Nucleic Acids Research*. Oxford University Press, 2016.

[58] Brian J. O'Roak, Laura Vives, Santhosh Girirajan, Emre Karakoc, Niklas Krumm, Bradley P. Coe, Roie Levy, Arthur Ko, Choli Lee, Joshua D. Smith, Emily H. Turner, Ian B. Stanaway, Benjamin Vernot, Maika Malig, Carl Baker, Beau Reilly, Joshua M. Akey, Elhanan Borenstein, Mark J. Rieder, Deborah A. Nickerson, Raphael Bernier, Jay Shendure, and Evan E. Eichler. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. In *Nature*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. SN -, 2012.

[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. In *Journal of Machine Learning Research*, 2011.

[60] Katherine S. Pollard, Sofie R. Salama, Bryan King, Andrew D. Kern, Tim Dreszer, Sol Katzman, Adam Siepel, Jakob S. Pedersen, Gill Bejerano, Robert Baertsch, Kate R. Rosen-

bloom, Jim Kent, and David Haussler. Forces shaping the fastest evolving regions in the human genome. In *PLOS Genetics*. Public Library of Science, 2006.

[61] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T. Afshar, Sam S. Gross, Lizzie Dorfman, Cory Y. McLean, and Mark A. DePristo. Creating a universal snp and small indel variant caller with deep neural networks. In *biorxiv*, 2018.

[62] Robert A. Power, Simon Kyaga, Rudolf Uher, James H. MacCabe, Niklas Laangstrom, Mikael Landen, Peter McGuffin, Cathryn M. Lewis, Paul Lichtenstein, and Anna C. Svensson. Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. In *JAMA Psychiatry*, volume 70, pages 22–30. American Medical Association, 2013.

[63] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. Plink: A tool set for whole-genome association and population-based linkage analyses. In *The American Journal of Human Genetics*, 2007.

[64] Gokul Ramaswami and Daniel H. Geschwind. Genetics of autism spectrum disorder. In *Handbook of Clinical Neurology*, volume 147, 2018.

[65] Rosenberg RE, Law J, Yenokyan G, McGready J, Kaufmann WE, and Law PA. Characteristics and concordance of autism spectrum disorders among 277 twin pairs. In *Archives of Pediatrics & Adolescent Medicine*, 2009.

[66] Elise B. Robinson, Beate St Pourcain, Verneri Anttila, Jack A. Kosmicki, Brendan Bulik-Sullivan, Jakob Grove, Julian Maller, Kaitlin E. Samocha, Stephan J. Sanders, Stephan Ripke, Joanna Martin, Mads V. Hollegaard, Thomas Werge, David M. Hougaard, iPSYCH-SSI-Broad Autism Group, Thomas D. Als, Marie Baekvad-Hansen, Richard Belliveau, Ditte Demontis, Ashley Dumont, Jacqueline Goldstein, Jonas Grauholm, Christine S. Hansen, Thomas F. Hansen, Daniel Howrigan, Francesco Lescai, Manuel Mattheisen, Jennifer Moran, Ole Mors, Merete Nordentoft, Bent Norgaard-Pedersen, Timothy Poterba, Jesper Poulsen, Christine Stevens, Raymond Walters, Benjamin M. Neale, David M. Evans, David Skuse, Preben Bo Mortensen, Anders D. Børglum, Angelica Ronald, George Davey Smith, and Mark J. Daly. Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. In *Nature Genetics*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. SN -, 2016.

[67] Michael Ronemus, Ivan Iossifov, Dan Levy, and Michael Wigler. The role of de novo mutations in the genetics of autism spectrum disorders. In *Nature Reviews Genetics*. Nature Publishing Group, 2014.

[68] Claire Rougeulle, Heather Glatt, and Marc Lalande. The angelman syndrome candidate gene, ube3aie6-ap, is imprinted in brain. In *Nature Genetics*. Nature Publishing Group SN -, 1997.

[69] Stephan J. Sanders, Xin He, A. Jeremy Willsey, A. Gulhan Ercan-Sencicek, Kaitlin E. Samocha, A. Ercument Cicek, Michael T. Murtha, Vanessa H. Bal, Somer L. Bishop, Shan Dong, Arthur P. Goldberg, Cai Jinlu, John F. III Keaney, Lambertus Klei, Jeffrey D. Mandell, Daniel Moreno-De-Luca, Christopher S. Poultney, Elise B. Robinson, Louw Smith, Tor Solli-Nowlan, Mack Y. Su, Nicole A. Teran, Michael F. Walker, Donna M. Werling, Arthur L. Beaudet, Rita M. Cantor, Eric Fombonne, Daniel H. Geschwind, Dorothy E. Grice, Catherine Lord, Jennifer K. Lowe, Shrikant M. Mane, Donna M. Martin, Eric M. Morrow, Michael E. Talkowski, James S. Sutcliffe, Christopher A. Walsh, Timothy W. Yu, David H. Ledbetter, Christa Lese Martin, Edwin H. Cook, Joseph D. Buxbaum, Mark J. Daly, Bernie Devlin, Kathryn Roeder, and Matthew W. State. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. In *Neuron*. Elsevier, 2015.

[70] Stephan J. Sanders, Michael T. Murtha, Abha R. Gupta, John D. Murdoch, Melanie J. Raubeson, A. Jeremy Willsey, A. Gulhan Ercan-Sencicek, Nicholas M. DiLullo, Neelroop N. Parikshak, Jason L. Stein, Michael F. Walker, Gordon T. Ober, Nicole A. Teran, Youeun Song, Paul El-Fishawy, Ryan C. Murtha, Murim Choi, John D. Overton, Robert D. Bjornson, Nicholas J. Carriero, Kyle A. Meyer, Kaya Bilguvar, Shrikant M. Mane, Nenad Sestan, Richard P. Lifton, Murat Günel, Kathryn Roeder, Daniel H. Geschwind, Bernie Devlin, and Matthew W. State. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. In *Nature*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. SN -, 2012.

[71] Jonathan Sebat, B. Lakshmi, Dheeraj Malhotra, Jennifer Troge, Christa Lese-Martin, Tom Walsh, Boris Yamrom, Seungtai Yoon, Alex Krasnitz, Jude Kendall, Anthony Leotta, Deepa Pai, Ray Zhang, Yoon-Ha Lee, James Hicks, Sarah J. Spence, Annette T. Lee, Kaija Puura, Terho Lehtimäki, David Ledbetter, Peter K. Gregersen, Joel Bregman, James S. Sutcliffe, Vaidehi Jobanputra, Wendy Chung, Dorothy Warburton, Mary-Claire King, David Skuse, Daniel H. Geschwind, T. Conrad Gilliam, Kenny Ye, and Michael Wigler. Strong association of de novo copy number mutations with autism. In *Science*, 2007.

[72] Fritz J. Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C. Schatz. Accurate detection of complex structural variations using single-molecule sequencing. In *Nature Methods*, 2018.

[73] Yonatan Stelzer, Hao Wu, Yuelin Song, Chikdu S. Shivalila, Styliani Markoulaki, and Rudolf Jaenisch. Parent-of-origin dna methylation dynamics during mouse development. In *Cell Reports*, 2016.

[74] Colin O. Stine, Jianfeng Xu, Rebecca Koskela, Francis J. McMahon, Michele Gschwend, Carl Friddle, Chris D. Clark, Melvin G. McInnis, Sylvia G. Simpson, Theresa S. Breschel, Eva Vishio, Kelly Riskin, Harriet Feilotter, Eugene Chen, Susan Shen, Susan Folstein, Deborah A. Meyers, David Botstein, Thomas G. Marr, and J. Raymond DePaulo. Evidence for linkage of bipolar disorder to chromosome 18 with a parent-of-origin effect. In *The American Journal of Human Genetics*, 1995.

[75] Michael R. Stratton and Nazneen Rahman. The emerging landscape of breast cancer susceptibility. In *Nature Genetics*. Nature Publishing Group, 2008.

[76] Peter H. Sudmant, Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, Miriam K. Konkel, Ankit Malhotra, Adrian M. Stütz, Xinghua Shi, Francesco Paolo Casale, Jieming Chen, Fereydoun Hormozdiari, Gargi Dayama, Ken Chen, Maika Malig, Mark J. P. Chaisson, Klaudia Walter, Sascha Meiers, Seva Kashin, Erik Garrison, Adam Auton, Hugo Y. K. Lam, Xinmeng Jasmine Mu, Can Alkan, Danny Antaki, Taejeong Bae, Eliza Cerveira, Peter Chines, Zechen Chong, Laura Clarke, Elif Dal, Li Ding, Sarah Emery, Xian Fan, Madhusudan Gujral, Fatma Kahveci, Jeffrey M. Kidd, Yu Kong, Eric-Wubbo Lameijer, Shane McCarthy, Paul Flicek, Richard A. Gibbs, Gabor Marth, Christopher E. Mason, Androniki Menelaou, Donna M. Muzny, Bradley J. Nelson, Amina Noor, Nicholas F. Parrish, Matthew Pendleton, Andrew Quitadamo, Benjamin Raeder, Eric E. Schadt, Mallory Romanovitch, Andreas Schlattl, Robert Sebra, Andrey A. Shabalin, Andreas Untergasser, Jerilyn A. Walker, Min Wang, Fuli Yu, Chengsheng Zhang, Jing Zhang, Xiangqun Zheng-Bradley, Wanding Zhou, Thomas Zichner, Jonathan Sebat, Mark A. Batzer, Steven A. McCarroll, The 1000 Genomes Project Consortium, Ryan E. Mills, Mark B. Gerstein, Ali Bashir, Oliver Stegle, Scott E. Devine, Charles Lee, Evan E. Eichler, and Jan O. Korbel. An integrated map of structural variation in 2,504 human genomes. In *Nature*, volume 526, 2015.

[77] Peter Szatmari, Andrew D. Paterson, Lonnie Zwaigenbaum, Wendy Roberts, Jessica Brian, Xiao-Qing Liu, John B. Vincent, Jennifer L. Skaug, Ann P. Thompson, Lili Senman, Lars Feuk, Cheng Qian, Susan E. Bryson, Marshall B. Jones, Christian R. Marshall, Stephen W. Scherer, Veronica J. Vieland, Christopher Bartlett, La Vonne Mangin, Rhinda Goedken, Alberto Segre, Margaret A. Pericak-Vance, Michael L. Cuccaro, John R. Gilbert, Harry H. Wright, Ruth K. Abramson, Catalina Betancur, Thomas Bourgeron, Christopher Gillberg, Marion Leboyer, Joseph D. Buxbaum, Kenneth L. Davis, Eric Hollander, Jeremy M. Silverman, Joachim Hallmayer, Linda Lotspeich, James S. Sutcliffe, Jonathan L. Haines, Susan E. Folstein, Joseph Piven, Thomas H. Wassink, Kacie J. Meyer, Val Sheffield, Daniel H. Geschwind, Maja Bucan, W. Ted Brown, Rita M. Cantor, John N. Constantino, T. Conrad Gilliam, Martha Herbert, Clara LaJonchere, David H. Ledbetter, Christa Lese-Martin, Janet Miller, Stan Nelson, Carol A. Samango-Sprouse, Sarah Spence, Matthew State, Rudolph E. Tanzi, Hilary Coon, Geraldine Dawson, Bernie Devlin, Annette Estes, Pamela Flodman, Lambertus Klei, William M. McMahon, Nancy Minshew, Jeff Munson, Elena Korvatska, Patricia M. Rodier, Gerard D. Schellenberg, Moyra Smith, M. Anne Spence, Chris Stodgell, Ping Guo Tepper, Ellen M. Wijsman, Chang-En Yu, Bernadette Roge, Carine Mantoulan, Kerstin Wittemeyer, Annemarie Poustka, Barbel Felder, Sabine M. Klauck, Claudia Schuster, Fritz Poustka, Sven Bolte, Sabine Feineis-Matthews, Evelyn Herbrecht, Gabi Schmatzer, John Tsiantis, Katerina Papanikolaou, Elena Maestrini, Elena Bacchelli, Francesca Blasi, Simona Carone, Claudio Toma, Herman Van Engeland, Maretha de Jonge, Chantal Kemner, Frederieke Koop, Marjolein Langemeijer, Channa Hijmans, Wouter G. Staal, Gillian Baird, Patrick F. Bolton, Michael L. Rutter, Emma Weisblatt, Jonathan Green, Catherine Aldred,

Julie-Anne Wilkinson, Andrew Pickles, Ann Le Couteur, Tom Berney, Helen McConachie, Anthony J. Bailey, Kostas Francis, Gemma Honeyman, Aislinn Hutchinson, Jeremy R. Parr, Simon Wallace, Anthony P. Monaco, Gabrielle Barnby, Kazuhiro Kobayashi, Janine A. Lamb, Ines Sousa, Nuala Sykes, Edwin H. Cook, Stephen J. Guter, Bennett L. Leventhal, Jeff Salt, Catherine Lord, Christina Corsello, Vanessa Hus, Daniel E. Weeks, Fred Volkmar, Maite Tauber, Eric Fombonne, and Andy Shih. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. In *Nature Genetics*. Nature Publishing Group SN -, 2007.

[78] Joanne L. Thorvaldsen, Kristen L. Duran, and Marisa S. Bartolomei. Deletion of the h19 differentially methylated domain results in loss of imprinted expression of h19 and igf2. In *Genes Development*. Cold Spring Harbor Laboratory Press, 1998.

[79] Tychele N. Turner, Bradley P. Coe, Diane E. Dickel, Kendra Hoekzema, Bradley J. Nelson, Michael C. Zody, Zev N. Kronenberg, Fereydoun Hormozdiari, Archana Raja, Len A. Pennacchio, Robert B. Darnell, and Evan E. Eichler. Genomic patterns of de novo mutation in simplex autism. In *Cell*. Elsevier, 2017.

[80] Thanh H. Vu and Andrew R. Hoffman. Imprinting of the angelman syndrome gene, ube3a, is restricted to brain. In *Nature Genetics*. Nature Publishing Group SN -, 1997.

[81] Binbin Wang, Taoyun Ji, Xueya Zhou, Jing Wang, Xi Wang, Jingmin Wang, Dingliang Zhu, Xuejun Zhang, Pak Chung Sham, Xuegong Zhang, Xu Ma, and Yuwu Jiang. Cnv analysis in chinese children of mental retardation highlights a sex differentiation in parental contribution to de novo and inherited mutational burdens. In *Scientific Reports*. The Author(s) SN -, 2016.

[82] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. In *Nucleic Acids Research*. Oxford University Press, 2010.

[83] Yuchao Xia, Yun Liu, Minghua Deng, and Ruibin Xi. Pysim-sv: a package for simulating structural variation data with gc-biases. In *BMC Bioinformatics*, 2017.

[84] Kenny Ye, Ivan Iossifov, Dan Levy, Boris Yamrom, Andreas Buja, Abba M. Krieger, and Michael Wigler. Measuring shared variants in cohorts of discordant siblings with applications to autism. In *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 2017.

[85] Samir Zaidi, Murim Choi, Hiroko Wakimoto, Lijiang Ma, Jianming Jiang, John D. Overton, Angela Romano-Adesman, Robert D. Bjornson, Roger E. Breitbart, Kerry K. Brown, Nicholas J. Carriero, Yee Him Cheung, John Deanfield, Steve DePalma, Khalid A. Fakhro, Joseph Glessner, Hakon Hakonarson, Michael J. Italia, Jonathan R. Kaltman, Juan Kaski, Richard Kim, Jennie K. Kline, Teresa Lee, Jeremy Leipzig, Alexander Lopez, Shrikant M. Mane, Laura E. Mitchell, Jane W. Newburger, Michael Parfenov, Itsik Pe'er, George Porter, Amy E. Roberts, Ravi Sachidanandam, Stephan J. Sanders, Howard S. Seiden,

Mathew W. State, Sailakshmi Subramanian, Irina R. Tikhonova, Wei Wang, Dorothy Warburton, Peter S. White, Ismee A. Williams, Hongyu Zhao, Jonathan G. Seidman, Martina Brueckner, Wendy K. Chung, Bruce D. Gelb, Elizabeth Goldmuntz, Christine E. Seidman, and Richard P. Lifton. De novo mutations in histone-modifying genes in congenital heart disease. In *Nature*, 2013.

[86] Xiaoyue Zhao, Anthony Leotta, Vlad Kustanovich, Clara Lajonchere, Daniel H. Geschwind, Kiely Law, Paul Law, Shanping Qiu, Catherine Lord, Jonathan Sebat, Kenny Ye, and Michael Wigler. A unified genetic theory for sporadic and inherited autism. In *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 2007.