Modeling Rare Protein-Coding Variation to Identify Mutation-Intolerant Genes With Application to Disease

Modeling rare protein-coding variation to identify mutation-intolerant genes with application to disease

A dissertation presented

by

Kaitlin Elisabeth Samocha

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Genetics and Genomics

Harvard University

Cambridge, Massachusetts

April 2016

Dissertation Advisor: Professor Mark J. Daly Kaitlin Elisabeth Samocha

Modeling rare protein-coding variation to identify mutation-intolerant genes with application to disease

Abstract

Sequencing exomes—the 1% of the genome that codes for proteins—has increased the rate at which the genetic basis of a patient's disease is determined. Unfortunately, when a patient does not carry a well-established pathogenic variant, it is extremely challenging to establish which of the tens of thousands of variants identified in that individual is contributing to their disease. In these situations, variants must be prioritized to make further investigation more manageable. In this thesis, we have focused on creating statistical frameworks and models to aid in the interpretation of rare variants and towards establishing gene-level metrics for variant prioritization.

We developed a sensitive and specific workflow to detect newly arising (*de novo*) variants from exome sequencing data of parent-child trios, and created a sequence-context based mutational. This mutational model was the basis of a rigorous statistical framework to evaluate the significance of *de novo* variant burden not only globally, but also per gene. When we applied this framework to *de novo* variants identified in patients with an autism spectrum disorder, we found a global excess of *de novo* loss-of-function variants as well as two genes that harbored significantly more *de novo* loss-of-function variants than expected.

We also used the mutational model to predict the expected number of rare (minor allele frequency < 0.1%) variants in exome sequencing datasets of reference individuals. We found a significant depletion of missense and loss-of-function variants in a subset of genes, indicating that these genes are under strong evolutionary constraint. Specifically, we identified 3,230 genes that are intolerant of loss-of-function variation and that set of genes is enriched for established dominant and haploinsufficient disease genes. Similarly, we searched for regions within genes that were intolerant of missense variation. The most missense depleted 15% of the exome contains 83% of reported pathogenic variants found in haploinsufficient disease genes that cause severe disease. Additionally, both gene-level and region-level constraint metrics highlight a set of *de novo* variants from patients with a neurodevelopmental disorder that are more likely to be pathogenic, supporting the utility of these metrics when interpreting rare variants within the context of disease.

Table of Contents

Acknowledgements

I have a hard time expressing how grateful I am that Mark Daly agreed to be my mentor. We met during my interview weekend and got lost while exploring the medical school campus to find my next meeting. I am so happy that he allowed me explore science with him as well. He may not agree, but I felt as though taking me on as a student was a gamble; I was mostly untrained in both programming and statistics—the bread and butter of our group—and unaware of the field of human genetics. Mark was always very patient with me and knew exactly when to give subtle nudges to get me back on track. I would not be the scientist I am today without Mark's mentorship and I am deeply indebted to him for the years of guidance.

I am also lucky to have many other mentors with the ATGU. Ben Neale helped supervise my work since the very beginning and his knowledge of statistical genetics was crucial to any successes I've had. Additionally, it was wonderful to have Elise Robinson as a mentor, teacher, and role model. I am also happy that I was absorbed into Daniel MacArthur's lab. The opportunity to work with his group was amazing, but I most appreciate his advice on life as a scientist. I would also like to thank Steve Schaffner for answering my questions about population genetics and being a great crossword buddy, and Chris Cotsapas for the many dinners and life advice.

It was wonderful to be surrounded by an amazing set of lab members in the ATGU. Andrew Kirby remains a wealth of knowledge and I greatly appreciate all of his help. I was also fortunate to share my office with wonderful young minds: Jackie Goldstein, Nikita Artomov, and Jack Kosmicki. In particular, Jack has been a good friend and I know he'll do great things in the future. It has been a fun experience

watching the lab grow and change over the last 5 years. I especially appreciate the wave of postdocs who helped turn the lab into a more social place. I've had great times (scientific and not) with Konrad Karczewski, Taru Tukiainen, and Verneri Antilla. I also need to specifically thank our awesome admins Jill Harris, Beth Raynard, and Carla Hammond. Nothing would have gotten done around the lab without them.

Graduate school is a challenge and can be full of dark and stressful times. I certainly would not have made it through these many years without my friends in the BBS program. In particular, I have to thank Niroshi Senaratne for her amazing support, kindness, and dances. Kostadin Petrov was my friend since day one and has continued to be a wonderful source of scientific and life advice. I appreciate Ben Vincent's grounded attitude and sass and am happy to have him as a friend. Daniel Grubaugh has been one of the greatest sources of fun times in graduate school and an awesome partner in broventures. I respect Stephen Hinshaw's dedication to doing science the right way; he's also a great guy once you get to know him. More generally, the Program in Genetics and Genomics has been a wonderful community and I've truly enjoyed my time with the PGG members of my class: Alex Meeske, Danielle Heller, and some of the aforementioned individuals. You guys made it worth it.

Finally, there is no way I could ever thank my family enough for their love and support. My parents have always been my most steadfast advocates and influential mentors. From early ages, they encouraged pursing scientific endeavors and they have always been there for me when I was in tough situations. I'm also thankful to have such wonderful siblings, who can be relied upon for a laugh or a hug at any time. A huge thank you to my family for being the best.

Chapter 1

Introduction

Overview

A primary goal of medical genetics is to associate genetic variants with risk of disease. This goal is impeded by a variety of complicating factors, such as the vast amount of genetic variation found in each individual[1] and the fact that such variation can impact as much sequence as whole chromosomes to as little as single bases. Additionally, the genetic basis of human diseases varies in the complexity of its architecture: some diseases are monogenic and caused by high impact variants. These disorders are for the most part rare, and the one-to-one relationship between disease and disruption of a single gene often allows for identification of the risk locus in a small number of families. When the relevant gene has been identified, then specific variants can be established as pathogenic and screened for in new patients.

Common diseases, however, have a far more complex genetic architecture and typically involve variants spread across the genome, each of which has a small effect on the phenotype (polygenicity). The polygenicity of common diseases makes it much more challenging to identify specific genetic risk factors; association analyses often require tens of thousands of affected and unaffected individuals (e.g. >36,000 cases with schizophrenia and >113,000 controls)[2]. The small average effect size of any identified risk-contributing variant does not typically permit the nomination of a primary causal event.

Even when studying a disease that may be influenced by stronger acting variants, determining the specific variant or set of variants that are contributing to a patient's disease is challenging, particularly when the patient does not carry a well-established genetic risk factor. Unfortunately for these types of activities, each individual

harbors tens of thousands of variants (single base to larger structural changes). Focusing in on those variants that alter the coding sequence leaves thousands to examine, even if only considering alleles that are rare in the general population. Therefore, it is critical to be able to prioritize variants that are more likely to be contributing to disease. A primary focus on this thesis has been to establish methods to aid such prioritization.

Tying genetic variation to disease

It has long been observed that the frequency of some diseases in families is associated with the family members' degree of relatedness, which suggests that the disease has a genetic component. A measure of the degree to which inherited genetic variation is contributing to disease is referred to as heritability[3]. More specifically, heritability is the amount of phenotypic variability that can be explained by inherited genetic variation. Estimates of heritability can come from many sources, but one of the classical approaches is to compare the concordance of the disease in monozygotic twins versus the concordance seen in dizygotic twins[4]. Since monozygotic twins share nearly 100% of their genetic material and dizygotic twins only ~50%, a highly heritable trait would be expected to have a much higher concordance rate in monozygotic twins. Of note, most traits are influenced by both genetic and environmental factors; comparing concordance in twins is designed to control for most of the environmental influence.

Heritability, however, does not provide information about specific loci that are contributing to disease etiology. In order to find risk loci, researchers have taken a

variety of approaches, limited partially by available technology. One of the earliest approaches that could be used to identify risk loci was linkage mapping[5-8]. Linkage mapping relied on collecting families with many affected and unaffected members. To map, sites across the genome were used as marks of all variation nearby and the segregation of these markers in the family were compared to the segregation of the disease. Markers near areas of the genome that contribute to disease should therefore follow the inheritance of the disease. Linkage mapping is best suited to diseases that are caused by large effect variants that occur in a small number of genes, which makes the technology poorly suited for complex trait association.

For diseases caused by many loci of small effect, genome-wide association studies (GWAS)[6,9,10] are performed to find the contributing variants. In GWAS, sites of a common variation in the population are treated as markers of nearby variation, much like linkage mapping. However, instead of using families, GWAS use large numbers of unrelated affected and unaffected individuals and search for variants that are seen more often in the affected than unaffected individuals. Given that these loci individually have a small impact on risk, they are still seen commonly in unaffected individuals. GWAS are also affected by the polygenicity of a disease; for those diseases with many contributing loci, very large cohorts of affected and unaffected individuals are needed to identify specific risk variants.

There are diseases that have a strong genetic component, but whose contributing variants would be difficult to find in either linkage or association studies. These are disorders that are not often passed on because they are extremely severe and affected individuals either do not survive to maturity or do not have children of their

own. These diseases are therefore often influenced by newly arising (*de novo*) alleles. An example of such a disease is Hutchinson-Gilford progeria syndrome, a rare disorder in which affected individuals show signs of early aging, such as hair loss and scleroderma[11]. It is caused by *de novo* missense variants in *LMNA* with the most common risk allele leading to the activation of a cryptic splice site and creation of a truncated protein product of the impacted gene copy[12]. These alleles are never passed on from an affected individual to their child because individuals with progeria die at an average age of 13[11].

De novo variation and disease

Beyond examples like Hutchinson-Gilford progeria, *de novo* variation can also contribute to diseases that are not always lethal in childhood. Achondroplasia, a form of dwarfism, is caused by heterozygous (only one copy of the gene being affected) disruptions of *FGFR3*[13,14]. While the disease and risk allele can be inherited from an affected parent, most cases are caused by a *de novo* event[15].

It was noted in the early 1900s that sporadic cases of achondroplasia occurred more often in the last-born child[16] and it was later shown that a higher rate of achondroplasia is specifically associated with advanced paternal age[15,17]; a similar trend has been seen for other disorders as well[18]. Overwhelmingly, the causal allele was paternal in origin; in the case of achondroplasia, all 40 cases tested by Wilkin and colleagues were on the paternally inherited chromosome[19]. These results indicated not only that there is a higher mutation rate in males, but suggested that the number of mutations increases as the father ages. Germline mutations are introduced during DNA

replication in mitosis and the first half of meiosis. The female germline has 22 rounds of mitosis and 1 round of meiosis during development to produce an egg[18,20]. The male germline, however, undergoes far more mitotic divisions owing to lifelong sperm production, thereby having more opportunities to mutate. Additionally, the number of replication cycles affecting a particular sperm is higher for older males. It has been estimated that a 20-year old male has had approximately 150 rounds of replication where a 40-year old has had 610[18,20]. The increased number of mitotic divisions in the male germline, however, does not fully explain the increased rate of sporadic achondroplasia among the children of older fathers[21].

De novo variation also contributes to more complex disorders, such as schizophrenia and autism spectrum disorder, where no single *de novo* allele is likely to lead to a patient's disease. As these disorders involve a large number of contributing loci, it is more challenging to define the role *de novo* variation plays. In particular, determining which *de novo* variants, if any, are contributing is complicated by non-disease associated *de novo* variants: every individual is expected to carry 70-100 *de novo* variants across their genome[22-26]. Since the *de novo* variant signal is likely to be spread across many genes, studying *de novo* variation in these disorders requires careful consideration of the mutation rate.

The role of sequencing technology

Linkage and association studies rely on the ability to determine the allele at a specific locus, but historically a relatively limited number of loci were measured because sequencing and genotyping were slow and expensive. The advent of massively parallel

sequencing technologies opened the door to quickly and affordably interrogate variation at many locations and as small as single base changes.

Whole genome sequencing within families successfully identified risk loci for Charcot-Marie-Tooth[27] and severe hypercholesterolemia[28], but in both cases the risk loci resided in the exome – the 1% of the genome that codes for proteins. Since understanding the effects of non-coding variation remains a major challenge to the field, much of the sequence data produced in these studies is considered uninterpretable. The creation of exome-capture kits allowed researchers to sequence only coding segments, which is faster and cheaper than sequencing the whole genome, thereby accelerating the discovery of protein-coding disease-associated variation[29,30].

Early successes of exome sequencing studies came from rare, severe, and likely monogenic disorders, such as Kabuki syndrome[31], Schinzel-Giedion syndrome[32], and Miller syndrome[33]. In the case of Kabuki syndrome, the nonsynonymous variants in *KMT2B* (previously known as *MLL2*) that were considered causal were often *de novo* in the affected individual[31]. These early studies proved that sequencing technology is especially critical for identifying *de novo* variation.

Using evolutionary conservation to prioritize variants

When analyzing the thousands of protein-coding variants within a patient, it is critical to prioritize variants for further investigation. One way to do so is to focus on variants that occur in genes that have been buffered against mutation across evolutionary time. The Human[34] and Mouse[35] Genome Projects – whose aims were to create reference genomes for the species – allowed large-scale comparisons of genetic

sequence in between species. The similarity (conservation) of sequence between humans and mice was first used to aid in the annotation of the genomes: highly conserved sequences were considered likely to be functional elements. Sequence homology, therefore, helped define coding and regulatory sequences within both species[34,35]. The level of conservation of sequence between species has become a common metric to indicate the importance of the sequence. Particularly, once gene annotations were defined, a plethora of tools were built to leverage sequence homology to predict the likely deleteriousness of specific variants (e.g. SIFT[36], GERP[37], Polyphen2[38]).

Additionally, reference sequences of various species, and the identification of polymorphisms within the species, allowed estimation of evolutionary selection pressures on genes (both positive and negative). The classical approaches rely on comparing the rates of nonsynonymous and synonymous substitutions (e.g. d_N/d_S, $K_a/K_{s)}$[39-43]. These methods were also used to measure the strength of the selection, often given as a selective coefficient (s) where $s = 0$ indicates neutrality and $s = 1$ lethality[44-46].

While successful at identifying genes under the influence of weak negative selection (selective coefficient [s] < 10^{-3}), these methods rely on the observation of variation within the population. Severely deleterious alleles ($s > 10^{-2}$), however, will never become polymorphisms within a population. As modeled by Zuk et al[47], when $s \geq 10^{-2}$) the combined allele frequency of variants with that selective coefficient is approximately 0.0001, independent of the demographic history of the population studied (**Figure 1.1**, reproduced from the paper). These simulations reinforce that alleles that

contribute to traits which greatly reduce fecundity (reproductive rate) will never become common enough in the population to be included in conservation-based metrics.

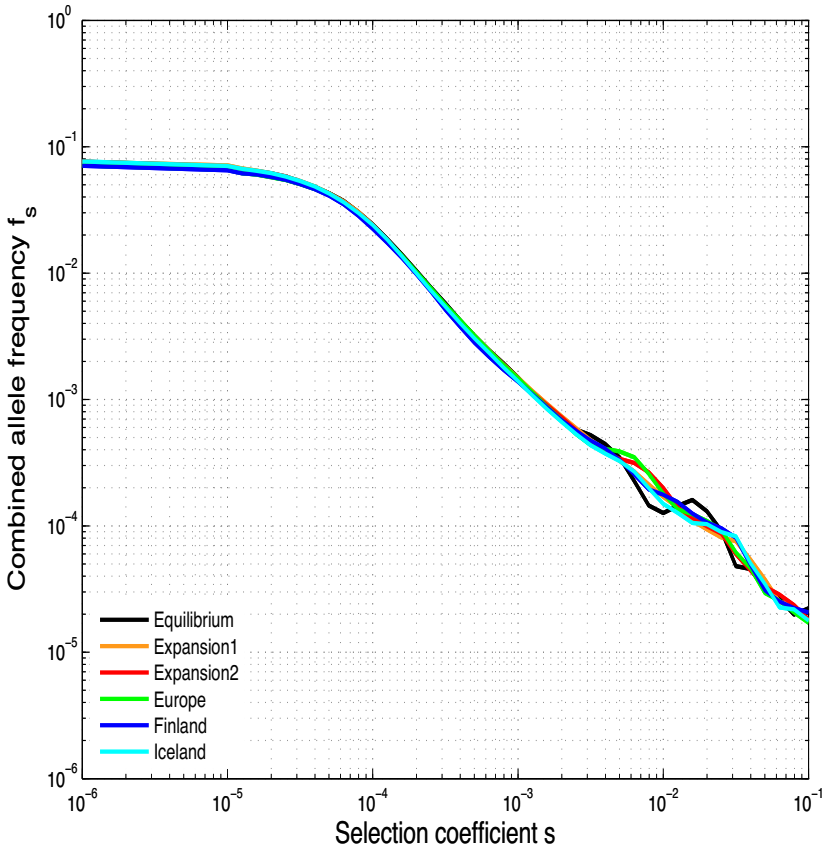


Figure 1.1. The relationship between combined allele frequency and selective coefficient (*s*) for various demographic models. As *s* increases, indicating stronger negative selection against those allele, the combined allele frequency of all variants with that *s* becomes smaller. This is supplementary Figure 1 from Zuk et al[47].

In order to determine how likely a variant, especially a *de novo* variant, seen in a patient is likely to be, we needed a measure of deleteriousness that captured larger values of *s*. In this thesis, we propose using the depletion of nonsynonymous variation within the human population as a reflection of the deleteriousness of variants that arise

within the gene. We are aided by recent large-scale sequencing efforts of reference populations, which provide power to determine significant depletion of variation.

Genetic basis of autism spectrum disorder

Autism spectrum disorders (ASDs) are a set of severe neurodevelopmental disorders that arise early in childhood and are characterized by impaired social interactions and communication, as well as restricted interests and repetitive behaviors. It has been recently estimated that the prevalence of ASDs in the United States is over 1%, with a notable excess of male cases[48]. The biological bases of ASDs are currently unknown.

ASDs are a class of disorders unlikely to show a strong evolutionary signature due to the strength of selection against the disorder. One way to measure how strongly selection is acting on a particular disease is to investigate the reproductive rate (fecundity) of affected individuals. In a study of a birth cohort in Sweden, Power and colleagues found that patients with ASD had dramatically reduced fecundity: male patients had 75% fewer children than their unaffected relatives, indicating very strong selective constraint (high *s*). Females showed a similar, but less severe, pattern (fecundity ratio 0.48)[49].

Various studies have established that there is a substantial genetic component to ASD risk. Estimates of the heritability of ASD are typically between 60 and 80%, indicating a large genetic component[50]. Unfortunately, research to find the genetic basis of ASDs has not been particularly successful.

Early linkage mapping efforts in ASD identified a very limited set of risk loci due to the highly polygenic and heterogeneous nature of ASD risk. Since linkage mapping works best for disorders caused by large-effect variants that fall into a small number of genes, linkage mapping successfully identified the causal loci for syndromic forms of ASDs, such as Fragile X syndrome[51] and Rett syndrome[52]. While linkage also identified a few universally accepted risk loci (*NLGN3*[53], *NLGN4X*[53], *NRXN1*[54], and *SHANK3*[55]), it mostly lead to long lists of candidate genes, whose association to ASD did not replicate in subsequent studies. Similarly, multiple GWAS of ASD did not report significant results, or found loci that never replicated[56-58]. The association studies were limited small sample size in conjunction with the fact that each risk variant has a very small effect on phenotype. This limitation will be overcome when large enough samples are aggregated and jointly analyzed.

The most successful early studies of the genetic basis of ASD were those that found associated copy number variants (CNVs)[59-64]. Researchers identified several CNVs that were strongly associated with risk for ASD (listed in **Table 1.1**), such as duplications and deletions in the 16p11.2 region[59,60,64,65]. These CNVs had larger effect sizes than are typically found for variants identified via GWAS and were often *de novo* in the affected individual[59,60,62-64]. Given the reduced fecundity of individuals with ASDs, it is not surprising that large effect variants are often *de novo*: these variants cannot be maintained in the population for multiple generations.

Table 1.1. Recurrent *de novo* copy number variants associated with autism spectrum disorder (ASD). The "Del vs Dup" column lists whether duplications, deletions, or both in the locus are associated with ASD. Size is given in megabase pairs (Mbp). Both the size of the region and the number of genes are approximate.

Region	Size (Mbp)	Genes	Del vs Dup	References
1q21.1	1.3	14	Both	Sanders 2011[62]
7q11.23	1.4	22	Duplication	Levy 2011[59]; Sanders 2011[62]
15q11.2-13.1	4.9	12	Duplication	Levy 2011[59]; Marshall 2008[60]; Pinto 2010[61]; Sanders 2011[62]; Sebat 2007[63]
15q13.2-13.3	1.5	6	Both	Marshall 2008[60]; Sanders 2011[62]; Sebat 2007[63]
16p11.2	0.6	25	Both	Levy 2011[59]; Marshall 2008[60]; Pinto 2010[61]; Sanders 2011[62]; Sebat 2007[63]; Weiss 2008[64]
22q11.2	2.5	56	Both	Marshall 2008[60]; Pinto 2010[61]; Sanders 2011[62]

In light of these successes and the availability of exome sequencing data, our group began to study *de novo* single nucleotide (SNV) and insertions and deletions (indels) in ASD cases. A previous publication had sequenced 20 parent-child trios, but could not implicate any specific gene or pathway in ASD etiology[66], an unsurprising result given the high polygenicity and locus heterogeneity of ASDs. As described in Chapter 3 and many subsequent publications[67-73], it took analyzing many more trios to identify a significant, but minor[50], role of *de novo* variation in ASD.

Summary

The ability to sequence patient genomes has allowed researchers to study variation with base-pair resolution. Sequencing, however, identifies thousands upon thousands of variants that need to be filtered in order to find those that may be

contributing to a patient's disease. For this thesis, we wanted to create methods and tools that could be used to aid in such prioritization of variants.

We first determined a way to sensitively and specifically identify *de novo* variants from family sequencing studies (Chapter 2). In order to properly analyze these results, we created a mutational model and statistical framework to rigorously evaluate excesses of such variation that may be observed in a patient population (Chapter 3). In particular, we established an important, but modest, role for *de novo* loss-of-function—and to a lesser extent missense—variation in ASD.

Given the modest enrichment of *de novo* variation in ASD cases, we needed a way to identify those variants that were more likely to be contributing. We used the mutational model we created to identify genes that are intolerant of nonsynonymous variation. In particular, using a large exome sequencing data set, we found 3,230 genes that appear to be extremely loss-of-function intolerant to the point of near haploinsufficiency – meaning that heterozygous loss-of-function variants in these genes should cause disease (Chapter 4). These highly loss-of-function intolerant genes contain the majority of the signal for *de novo* loss-of-function variants found in ASD cases as well as cases with other neurodevelopmental disorders.

We also hoped to explain the modest excess of missense variation in ASD cases by searching for specifically missense constrained regions within genes. Using the intolerance to missense variation, as well as variant level predictors of deleteriousness, we created a score to predict how likely a missense variant is to be deleterious and show that is separates signal from noise in the *de novo* missense variant results from ASD cases (Chapter 5).

In summary, we have developed tools and metrics to better interrogate exome sequencing data and applied them to substantially clarify the role of rare variation in ASD risk. These approaches have been adopted by the broader community to both inform rare variant discovery and patient exome interpretation.

Bibliography

1. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

2. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-7 (2014).

3. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* **9**, 255-66 (2008).

4. Falconer, D.S.M.T.F.C. *Introduction to Quantitaive Genetics*, (Pearson, Longmans Green, Harlow, Essex, UK, 1996).

5. Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**, 314-31 (1980).

6. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881-8 (2008).

7. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* **11**, 241-7 (1995).

8. Ott, J., Wang, J. & Leal, S.M. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* **16**, 275-84 (2015).

9. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-7 (1996).

10. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95-108 (2005).

11. Hennekam, R.C. Hutchinson-Gilford progeria syndrome: review of the phenotype. *Am J Med Genet A* **140**, 2603-24 (2006).

12. Eriksson, M. *et al.* Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature* **423**, 293-8 (2003).

13. Shiang, R. *et al.* Mutations in the transmembrane domain of FGFR3 cause the most common genetic form of dwarfism, achondroplasia. *Cell* **78**, 335-42 (1994).

14. Rousseau, F. *et al.* Mutations in the gene encoding fibroblast growth factor receptor-3 in achondroplasia. *Nature* **371**, 252-4 (1994).

15. Oberklaid, F., Danks, D.M., Jensen, F., Stace, L. & Rosshandler, S. Achondroplasia and hypochondroplasia. Comments on frequency, mutation rate, and radiological features in skull and spine. *J Med Genet* **16**, 140-6 (1979).

16. Weinberg, W. Zur Vererbung des Zwergwuchses. *Arch. Rassen- u. Gesel. Biolog.* **9**, 710-718 (1912).

17. Penrose, L.S. Parental age and mutation. *Lancet* **269**, 312-3 (1955).

18. Risch, N., Reich, E.W., Wishnick, M.M. & McCarthy, J.G. Spontaneous mutation and parental age in humans. *Am J Hum Genet* **41**, 218-48 (1987).

19. Wilkin, D.J. *et al.* Mutations in fibroblast growth-factor receptor 3 in sporadic cases of achondroplasia occur exclusively on the paternally derived chromosome. *Am J Hum Genet* **63**, 711-6 (1998).

20. Vogel, F. & Rathenberg, R. Spontaneous mutation in man. *Adv Hum Genet* **5**, 223-318 (1975).

21. Tiemann-Boege, I. *et al.* The observed human sperm mutation frequency cannot explain the achondroplasia paternal age effect. *Proc Natl Acad Sci U S A* **99**, 14952-7 (2002).

22. Campbell, C.D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet* **44**, 1277-81 (2012).

23. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature genetics* **43**, 712-4 (2011).

24. Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* **21**, 12-27 (2003).

25. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-5 (2012).

26. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* **107**, 961-8 (2010).

27. Lupski, J.R. *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *The New England journal of medicine* **362**, 1181-91 (2010).

28. Rios, J., Stein, E., Shendure, J., Hobbs, H.H. & Cohen, J.C. Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. *Hum Mol Genet* **19**, 4313-8 (2010).

29. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182-9 (2009).

30. Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39**, 1522-7 (2007).

31. Ng, S.B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature genetics* **42**, 790-3 (2010).

32. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nature genetics* **42**, 483-5 (2010).

33. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics* **42**, 30-5 (2010).

34. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

35. Mouse Genome Sequencing, C. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).

36. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-4 (2003).

37. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-13 (2005).

38. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-9 (2010).

39. Bustamante, C.D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153-1157 (2005).

40. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652-4 (1991).

41. Li, W.H., Wu, C.I. & Luo, C.C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol* **2**, 150-74 (1985).

42. Suzuki, Y. & Gojobori, T. A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* **16**, 1315-28 (1999).

43. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17**, 32-43 (2000).

44. Nielsen, R. & Yang, Z. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol* **20**, 1231-9 (2003).

45. Yang, Z. & Nielsen, R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol* **25**, 568-79 (2008).

46. Keightley, P.D. & Eyre-Walker, A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**, 2251-61 (2007).

47. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**, E455-64 (2014).

48. Christensen, D.L. *et al.* Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012. *MMWR Surveill Summ* **65**, 1-23 (2016).

49. Power, R.A. *et al.* Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry* **70**, 22-30 (2013).

50. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat Genet* **46**, 881-5 (2014).

51. Verkerk, A.J. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905-14 (1991).

52. Amir, R.E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* **23**, 185-8 (1999).

53. Jamain, S. *et al.* Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. *Nature genetics* **34**, 27-29 (2003).

54. Kim, H.-G. *et al.* Disruption of neurexin 1 associated with autism spectrum disorder. *American journal of human genetics* **82**, 199-207 (2008).

55. Durand, C.M. *et al.* Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nature genetics* **39**, 25-27 (2007).

56. Ma, D. *et al.* A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Annals of human genetics* **73**, 263-73 (2009).

57. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528-33 (2009).

58. Weiss, L.A. *et al.* A genome-wide linkage and association scan reveals novel loci for autism. *Nature* **461**, 802-8 (2009).

59. Levy, D. *et al.* Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886-97 (2011).

60. Marshall, C.R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *American journal of human genetics* **82**, 477-88 (2008).

61. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-72 (2010).

62. Sanders, S.J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-85 (2011).

63. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445-9 (2007).

64. Weiss, L.A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *The New England journal of medicine* **358**, 667-75 (2008).

65. Kumar, R.A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Human molecular genetics* **17**, 628-38 (2008).

66. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature genetics* **43**, 585-9 (2011).

67. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).

68. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-21 (2014).

69. Iossifov, I. *et al.* De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* **74**, 285-299 (2012).

70. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-245 (2012).

71. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250 (2012).

72. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).

73. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241 (2012).

Chapter 2

Identifying and characterizing *de novo* variation

This chapter is based on methods reported in:

Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-245 (2012).

De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).

Motivation

A natural property of DNA is that it spontaneously mutates, which leads to the creation of new alleles. The mutation rate of single nucleotides is quite low, but still large enough to expect that any individual will carry 70-100 newly arising (*de novo*) single nucleotide alleles not present in the somatic genome of either parent, with roughly one of these *de novo* alleles falling into exomic sequence (the 1% of the genome that encodes protein)[1-5].

While it has been established that there is a large genetic component to autism spectrum disorder (ASD), both linkage and association studies had limited success identifying risk loci. Some of the most fruitful studies came from examining large copy number variants (CNVs)[6-11]. Researchers identified several CNVs that were strongly associated with risk for ASD, many of which were *de novo* in the affected child. Unfortunately, these CNVs are large and contain many genes, complicating studies to decipher the underlying biology. As an example, the most-reported CNVs tied to ASD are deletions in the 16p11.2 region, which span roughly 500-600 kilobases and contain 25 genes[7,8,10,12]. Understanding how these ASD-associated CNVs contribute to disease is further complicated by both incomplete penetrance and associations to other neurodevelopmental disorders[8,10,11,13]. All together, these CNVs account for less than 3% of the heritability of ASD, indicating that there is much more to find[14].

The development of exome-capture kits, in combination with the falling costs of sequencing DNA, allowed the study of *de novo* single nucleotide variants (SNVs) and small insertions and deletions (indels) found in coding sequence[15]. *De novo* SNVs in single genes have been tied to a number of rare, severe, and likely monogenic

disorders[16,17]. There were also a few studies of *de novo* variation in more complex traits, such as schizophrenia, with less success implicating specific risk genes[18,19].

Subsequently, our group and others began to sequence parent-child families (known as trios) to define the role of *de novo* variation in ASD and discover genes or pathways associated with disease risk. While in theory identifying *de novo* variation should be straight forward—finding alleles in the child that neither parent has—it is complicated by such occurrences being rare and looking like genotyping or sample tracking errors. We therefore needed to establish specific and sensitive quality thresholds to determine trustworthy candidate *de novo* events.

Data generation

Identifying *de novo* variation requires genetic information, specifically sequencing data, from both parents and their child (a trio). Our earliest work was with the Autism Consortium, a Boston-based group of collaborators, which collected whole blood or cell lines from 96 trios. DNA was extracted and sheared into 200-300 base long fragments, which were then end-repaired, adenylated, and had adaptor oligonucleotides ligated in preparation for sequencing. PCR amplification with primers specific to the adaptor oligonucleotides was performed to enrich for fragments with attached adaptors. Exons were captured using Agilent 38Mb SureSelect v2. After capture, a round of ligation-mediated PCR was performed to increase the quantity of DNA available for sequencing. All libraries were sequenced using an Illumina HiSeq 2000 instrument. The data were processed with the Picard software, which uses base quality score recalibration and local realignment at known indels[20] and Burrows-Wheeler Aligner (BWA)[21] to map reads

to hg19. Variants were called using the Genome Analysis Toolkit (GATK) software for all trios jointly[20,22]. The resulting output was a standard Variant Call Format (VCF) file containing genotypes for sequenced members of the trios at positions where at least one individual in the data set had a non-reference allele. All sequencing was performed at the Broad Institute.

Key parameters to identify *de novo* variants

We created a Python script to identify candidate *de novo* variants from sequencing data with two required inputs: a GATK-generated VCF file that contains the variant information and a family relation file—often referred to as a pedigree file—that describes sample relatedness. Our first requirement was that variants passed all of the standard quality filters of the genotyper (here, the GATK Unified Genotyper), which was indicated by a PASS in the FILTER field of the VCF. Of the high quality sites, we focused on those where a child had a heterozygous genotype and both parents were homozygous reference. Given the small size of the original data set (n = 96 trios), we assumed that any site where the alternative allele was seen in another individual in the data set was likely to not be a true *de novo* and therefore removed such sites from further consideration[23]. This assumption was later dropped (discussed below).

We then sought to remove miscalled genotypes by imposing a threshold on the observed allele balance (the percentage of non-reference, or alternative, reads). Since the child should be heterozygous for the alternative allele, roughly 50% of all sequencing reads at the site in the child should have the alternative allele. There is a slight reference bias—it is easier to capture sequences with the reference allele than

the alternative—as well as normal sampling error. Given these two properties, we allowed the child's allele balance to be as low as 30%. Additionally, we wanted to avoid the possibility of a missed heterozygous genotype in the parents and required that their allele balance be no greater than 5%. Instances where genotypes that appear to indicate a *de novo* event, but fail these expected allele balances, can arise from poor read mapping or biases in data generation but may still lead to confident genotyping calls if the site has high sequencing depth. We also removed sites where the child's read depth was < 10% of the total depth of reads in both parents, which was meant to remove instances where the child may have been poorly sequenced or, less likely, had a deletion at the site.

We next explored filtering variants based on the Phred-scaled likelihood (PL) of the data conditional on the genotype calls. The PL represents $-10 * log_{10}(p)$, where *p* is the likelihood ratio of each genotype. In the case of a site with a single alternative allele, a PL is provided for each genotype: PL(AA) for the homozygous reference, PL(AB) for the heterozygote, and PL(BB) for homozygous alternative. The most likely genotype is assigned a PL score of 0 and all others are scaled relative to the most likely genotype. Therefore, a PL of 30 corresponds to the genotype in question being a thousand times less likely to be the true genotype than the reported most likely genotype. To determine an appropriate PL filter, we set a threshold of *T* and required sites to have a PL ≥ *T* for the child's homozygous reference genotype—PL(AA)—and for the parents' heterozygous genotypes—PL(AB). The relationship between *T* and the number of retained *de novo* events is depicted in **Figure 2.1**. As the genotypes become increasingly confident (greater PLs), the number of *de novo* events drops until

plateauing at a PL of ~20-30[23]. We therefore set 30 as our required threshold, *T*, when evaluating *de novo* events.

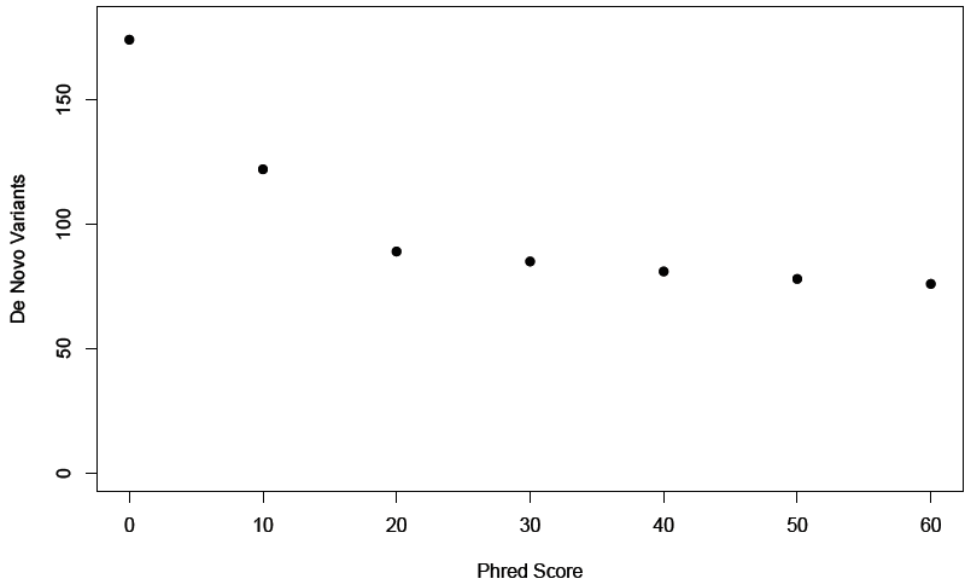


Figure 2.1. The relationship between the genotype likelihood threshold (Phred Score) used and the number of *de novo* variants found in 96 trios.

We sought to validate some of the identified *de novo* variants to support our choice of thresholds. Overall, nearly 95% of variants were confirmed to be *de novo* using an alternative sequencing technology, indicating high specificity[23]. To insure sensitivity of the PL threshold, we also attempted to validate variants that had a PL in between 20 and 30: all four of these variants failed to validate. Further investigation indicated that the most likely culprit for a falsely identified *de novo* event was missing a heterozygous genotype in one of the parents due to under-sampling the alternative allele, often because of low depth of sequencing.

After settling on the filtering parameters, we found 161 coding *de novo* variants in 175 ASD trios (the additional 78 were sequenced at other centers)[23]. The number of *de novo* variants per trio matched a Poisson distribution (**Table 2.1**). We also observed the expected relationship between variant deleteriousness and the number of alternative alleles observed in the data set (**Table 2.2**). More common alternative alleles, as a class, have a lower percentage of nonsynonymous variants. In addition, these more common alternative alleles have lower percentages of missense variants that are predicted to be damaging by Polyphen2[24], a program that estimates variant deleterious using conservation of the amino acid across species and whether the change is predicted to destroy important structural features of the protein, among other features.

Table 2.1. The number of *de novo* single nucleotide events per trio compared to the expected number of such events. We are including only single nucleotide variants (SNVs) and not insertions and deletions. The expected number of trios with a given number of *de novo* variants was determined by the Poisson with a lambda of 0.92, the median number of *de novo* events per trio.

Events per trio	Observed *de novo* SNVs	Expected *de novo* SNVs
0	71	69.7
1	62	64.2
2	28	29.5
3	10	9.1
4	2	2.1
5	1	0.4

Table 2.2. The percentage of variants by mutation type for ASD cases and their parents. Only single nucleotide variants are included. Singletons (alternative allele seen once in the data set), doubletons (alternative allele seen twice in the data set), and variants where the alternative allele was seen ≥ 3 times in the data set were only those variants found in the 192 parents of the original 96 trios.

Type of mutation	*De novo*	Singletons	Doubletons	≥ 3
Synonymous	31.1%	39.3%	43.8%	50.8%
Missense	62.7	59.5	55.4	48.8
Nonsense	6.2	1.2	0.8	0.4
PolyPhen-2 missense classification				
Benign	35.0%	46.6%	51.3%	63.4%
Possibly damaging	21.0	18.8	17.7	15.1
Probably damaging	44.0	34.7	31.0	21.4

Population frequency aware *de novo* identification

Our early work only considered *de novo* variants that were singletons, where the alternative allele is seen only once in the data set. Of course, it is possible for a true *de novo* event to arise at a site that has been mutated in another individual, an occurrence that becomes increasingly likely as the sample size of the data set increases. The logic behind our original choice to only consider singletons was that a *de novo* variant should be a private event and unlikely to be seen in another individual, especially given the limited sample size at the beginning of the study.

When we dropped the requirement that any *de novo* variant had to be a singleton, we found that many of the additional events identified had low read depth in all three members of the trio or borderline allele balances, indicating that there was likely under-calling of a heterozygous genotype in one of the parents.

We therefore had to refine our filters and thresholds to have strong confidence in variants that were seen in another individual in the dataset, or as a standing variant in the population, but appeared *de novo* in a trio. As before, we first identified candidate *de*

novo variants at the highest quality (PASSing) sites as defined by a standard GATK pipeline where the child is called heterozygous and the parents both reference homozygotes. We maintained our requirement that the proportion of alternate allele reads was no more than 5% in each parent, but allowed the child to have as few as 20% alternate allele reads. We also removed variants where the depth of sequence coverage in the child was less than 10% of the total depth of the two parents. However, we dropped the PL requirement from 30 to 20 since we were adding other filters to produce confident *de novo* variant calls (discussed below).

The major error mode of falsely called *de novo* events is when one parent is truly heterozygous, but has been incorrectly called homozygous reference due to under-sampling of the alternative allele. We therefore implemented a novel algorithm that uses population and sample allele frequency information to provide a Bayesian probability estimate that an apparent *de novo* variant constitutes a true *de novo*, as opposed to a missed heterozygous call in the parent.

While the PL information from the parents provides an accurate picture of the probability of the data given the genotype, the prior probability of a heterozygous genotype must be derived from population data. To calculate this, we conservatively take the maximum allele frequency from two sources: the extensively curated National Heart, Lung, and Blood Institute's Exome Sequence Project reference database and the sequenced population sample from which the trio is drawn. Including both data sets permits use of both the accuracy that comes from the size of well-curated reference but insures against false low frequency estimates should there be an occasional variant missed in the reference resource but present in many copies in the current data. The

probability of a site being present in a parent but absent from the reference data and all other samples in our data is simply the average number of singleton sites unique to an individual (~100) divided by the exome target size in base pairs, whereas the prior probability of a *de novo* mutation is the mean number of *de novo* variants (~1) divided by the same exome size in base pairs.

The probabilities of the two hypotheses are then calculated using Bayes' theorem and the relative probability,

$$P(de\ novo) = \frac{P(true\ de\ novo \mid data)}{P(true\ de\ novo \mid data)+P(missed\ het\ in\ parent \mid data)},$$

is reported as the probability of *de novo* variant. Sites for which $P(de\ novo)$ was estimated to be greater than 0.99 were considered high quality sites and constitute the set of variants included in all analyses. We also combined $P(de\ novo)$ with the allele balance of the variant and its allele count in the data set to assign it to one of three categories: high, medium, and low likelihood of validating as a true *de novo* event.

We applied the improved version of the *de novo* identification script to the exome sequencing data from 1,474 trios where the child was diagnosed with ASD as part of the Autism Sequencing Consortium[25]. Extensive validation of sites via Sanger sequencing was performed and found that only three out of 200 high quality sites (both SNVs and indels) were inherited, confirming the validity of the $P(de\ novo)$ > 0.99 estimate (**Table 2.3**). Additionally, we tested 56 sites that were considered to have a medium likelihood of validating: 30 (53.6%) of these were confirmed to be *de novo*. These results further supported the validity of the probability estimate. As these variants constituted a small but significantly real category (estimated to add ~2% true events), they were included in all analysis of the *de novo* variants.

Table 2.3. Validation of *de novo* variants by their likelihood of validating. No validation was attempted for variants that fell into the low likelihood of validating category.

Likelihood of validating	Number of variants tested	Variants validated (%)	Confirmed *de novo* variants (%)
High	200	196 (98.0%)	193 (97.5%)
Medium	56	36 (64.3%)	30 (53.6%)

Author contributions

Kaitlin Samocha: method design, data analyses (exceptions below), writing

Mark Daly: method design, writing, overall guidance

Benjamin Neale: Poisson analysis in Table 2.1 ("Expected *de novo* SNVs" column), extracted and analyzed singleton, doubleton, and variants with ≥ 3 alternative alleles in Table 2.2, guidance, writing

Silvia De Rubeis: molecular validation listed in Table 2.3

Samples were provided by the Autism Consortium and Autism Sequencing Consortium Principal investigators: Eric Boerwinkle, Joseph Buxbaum, Edwin Cook Jr, Mark Daly, Bernie Devlin, Richard Gibbs, Michael Gill, Kathryn Roeder, Gerard Schellenberg, Matthew State, James Sutcliffe, Michael Zwick

Bibliography

1. Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* **21**, 12-27 (2003).

2. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* **107**, 961-8 (2010).

3. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature genetics* **43**, 712-4 (2011).

4. Campbell, C.D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet* **44**, 1277-81 (2012).

5. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-5 (2012).

6. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445-9 (2007).

7. Marshall, C.R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *American journal of human genetics* **82**, 477-88 (2008).

8. Weiss, L.A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *The New England journal of medicine* **358**, 667-75 (2008).

9. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-72 (2010).

10. Levy, D. *et al.* Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886-97 (2011).

11. Sanders, S.J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-85 (2011).

12. Kumar, R.A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Human molecular genetics* **17**, 628-38 (2008).

13. Sebat, J., Levy, D.L. & McCarthy, S.E. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends in genetics : TIG* **25**, 528-35 (2009).

14. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat Genet* **46**, 881-5 (2014).

15. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics* **42**, 30-5 (2010).

16. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nature genetics* **42**, 483-5 (2010).

17. Ng, S.B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature genetics* **42**, 790-3 (2010).

18. Girard, S.L. *et al.* Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature genetics* **43**, 860-3 (2011).

19. Xu, B. *et al.* Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nature genetics* **43**, 864-8 (2011).

20. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-8 (2011).

21. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).

22. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-303 (2010).

23. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-245 (2012).

24. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-9 (2010).

25. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).

Chapter 3

Using a mutational model to evaluate *de novo* findings and identify genes intolerant of missense variation

This chapter was previously published as:

Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).

Abstract

Spontaneously arising *(de novo)* variants play an important role in medical genetics. For diseases with extensive locus heterogeneity – such as autism spectrum disorders (ASDs) – the signal from *de novo* variants is distributed across many genes, making it difficult to distinguish disease-relevant variants from background variation. We provide a statistical framework for the analysis of *de novo* variant excesses per gene and gene set by calibrating a model of *de novo* mutation. We applied this framework to *de novo* variants collected from 1,078 ASD trios and – while affirming a significant role for loss-of-function variants – found no excess of *de novo* loss-of-function variants in cases with IQ above 100, suggesting that the role of *de novo* variants in ASD may reside in fundamental neurodevelopmental processes. We also used our model to identify ~1,000 genes that are significantly lacking functional coding variation in non-ASD samples and are enriched for *de novo* loss-of-function variants identified in ASD cases.

Introduction

Exome sequencing has allowed for the identification of *de novo* (newly arising) events and has already been effectively put to use in identifying causal variants in rare, mendelian diseases. In the case of Kabuki syndrome, the observation of a *de novo* variant in *KMT2D* (previously *MLL2*) in 9 out of the10 patients strongly implicated the loss of *KMT2D* function as causal[1]. The conclusion that *KMT2D* is important in Kabuki syndrome etiology based on the *de novo* variant findings relies upon the unlikely accumulation of independent and infrequently occurring events in the vast majority of

these unrelated cases. By contrast, *de novo* variants (DNVs) play a smaller role in the pathogenesis of heritable complex traits, such as autism spectrum disorders (ASDs), and associated DNVs are spread across multiple genes. These differences in the etiologic architecture of complex traits make the task of identifying “causal” genes considerably more challenging. For example, recent exome sequencing studies demonstrated a significant excess of *de novo* loss-of-function (LoF) variants in ASD cases, but lacked the ability to directly implicate more than a very few genes[2-6].

The main complicating factor for interpreting the number of observed DNVs for a particular gene is the background rate of *de novo* mutation, which can vary greatly between genes. As more individuals are sequenced, multiple DNVs will inevitably be observed in the same gene by chance. However, if *de novo* variation plays a role in a given disease, then we would expect to find that genes associated to disease should contain more DNVs than expected by chance.

Here, we develop a statistical model of *de novo* mutation in order to evaluate the findings from exome sequencing data. With this model, we establish a statistical framework to evaluate the rate of DNVs not only on a per-gene basis (in a frequentist manner analogous to common genome-wide association analysis), but also globally and by gene set. We further use this model to predict the expected amount of rare standing variation per gene and to detect those genes that are significantly and specifically deficient in functional variation, likely reflecting processes of selective constraint. Consequently, since selection has reduced standing functional variation in these genes, it is reasonable to hypothesize that mutations in these genes are more likely to be deleterious.

We used the mutational model along with our list of highly constrained genes to evaluate the relationship between *de novo* variation and ASDs. Most of the families employed in these analyses were included in a set of previous studies of *de novo* variation, which reported an overall excess of *de novo* LoF variants in ASD cases, as well as multiple DNVs in specific genes[2-5]. We build on those studies to examine the aggregate rates of DNVs, the excess of multiply mutated genes, and the overlap of DNVs with gene sets, which highlights the complex relationship between intellectual functioning and the genetic architecture of ASD.

Results

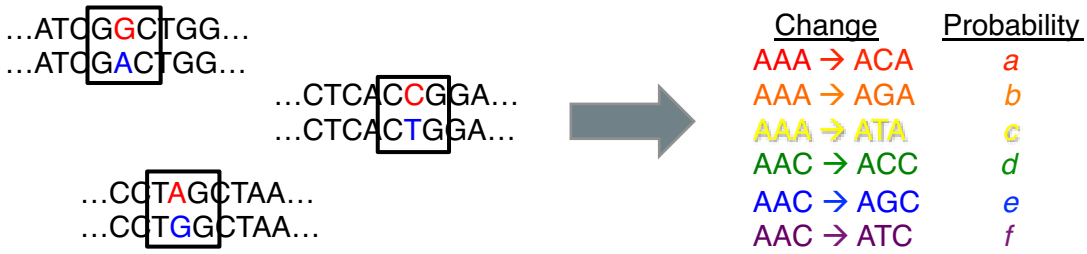
Basis of the mutational model

Accurate estimation of the expected rate of *de novo* mutation in a gene requires a precise estimate of each gene's mutability. While gene length is an obvious factor in a gene's mutability, local sequence context is also a well-known source of mutation rate differences[7]. Accordingly, we extended a previous model of *de novo* mutation based on sequence context and developed gene-specific probabilities for different types of mutation: synonymous, missense, nonsense, essential splice site, and frameshift (see Materials and Methods; **Figure 3.1**)[3]. Underscoring the importance of the sequence context factors in the model, this genome-wide rate yields an expected mutation rate of 1.67×10^{-8} per base per generation for the exome alone. Using counts of rare (minor allele frequency < 0.001) synonymous variants identified in the National Heart, Lung, and Blood Institute's (NHLBI's) Exome Sequencing Project (ESP), we found that our

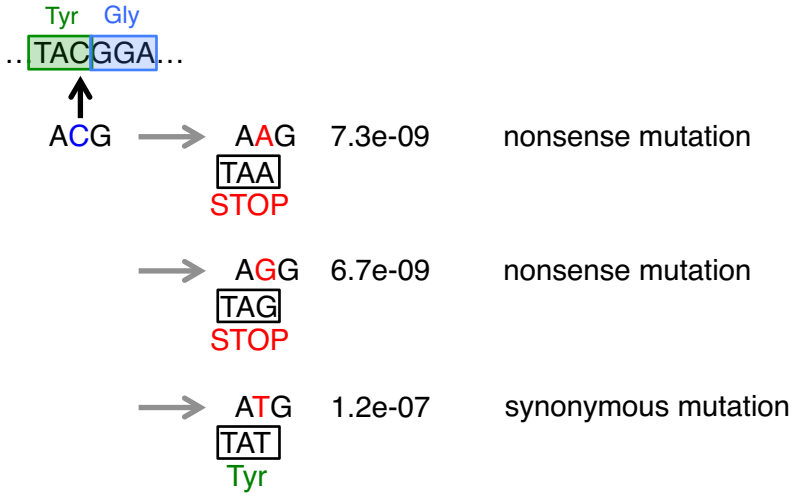
per-gene probabilities of mutation were significantly more correlated (r = 0.940) with these counts than gene length alone $p < 10^{-16}$; Materials and Methods).

Having established accurate per-gene probabilities of mutation, we could then investigate the rates and distribution of DNVs found in sequencing studies. Specifically, we wished to systematically assess a) whether cases had genome-wide excesses of certain functional categories of *de novo* variation; b) whether individual genes could be associated via *de novo* variation with genome-wide statistical significance; c) whether specific sets of genes collectively showed significant enrichment of *de novo* variants and d) whether there were genome-wide excesses of genes with multiple *de novo* variants. Below we demonstrate the utility of the statistical framework in addressing all of these questions with respect to recently generated family exome sequencing for autism and intellectual disability.

1. Create a mutation rate table from intergenic SNPs for all possible trinucleotide to trinucleotide changes



2. Use the sequence context to determine the probability of each base changing to each other base for all bases in the coding region and those in the conserved splice site
3. Determine the outcome of each type of change on the amino acid coded for by the base



4. Add up the probabilities for each outcome across a gene to create a probability per gene for different types of mutations

Figure 3.1. An outline of the steps used in the model of *de novo* mutation probability. A graphic illustration of the steps taken to determine the per-gene probabilities of mutation. A mutation rate table was created from intergenic single nucleotide polymorphisms (SNPs) from the 1000 Genomes Project. This mutation rate table was then applied to every coding base and the bases of conserved splice site to create a gene specific probability of mutation, split by mutation type.

Identifying genes under selective constraint

There has been a long standing interest in identifying genes in the human genome that are sensitive to mutational changes, as these genes would be the most likely to contribute to disease. Recent work made use of the ESP data to create a metric evaluating the proportion of common functional variation in each gene, thereby identifying genes that appeared to be intolerant of mutation[8]. Along these lines, we correlated our calculated per-gene probabilities of mutation with the observed counts of rare missense variants in the ESP data set. In contrast to the high consistency between predicted synonymous mutation rates and observed synonymous counts (expected if the category is under no specific selection), we observed a significant number of genes with severe deficit of missense variants compared to the expectation generated from predicted mutation rates ($p < 10^{-16}$). Such a deficit is consistent with strong evolutionary constraint: when damaging mutations arise, they are quickly removed from the population by purifying selection. To avoid erroneously identified constrained genes, we removed 134 genes with either significantly elevated or depressed synonymous and nonsynonymous rates (both $p < 0.001$; Materials and Methods).

Comparing both the synonymous and missense predictions of our model to the ESP data set, we identified a list of excessively constrained genes (missense Z score > 3.09; corresponding to $p < 0.001$) that represented roughly 5% of all genes. A high proportion of the most significantly constrained genes (missense constraint $p < 10^{-6}$) were associated with autosomal or X-linked dominant, largely sporadic, mendelian disease entries listed in the Online Mendelian Inheritance in Man database (OMIM; n = 27/86). By contrast, a set of genes for which the missense constraint was very close to

expectation (n = 111, -0.01 < Z < 0.01) had only two *de novo* or dominant disease inheritance entries in OMIM, a number significantly different from that for the highly constrained set ($p < 10^{-8}$). For the 86 most highly constrained genes, no autosomal recessive mendelian disorders have been documented. However, 11 of the 111 genes with average levels of constraint have been identified as causal in autosomal recessive mendelian disorders. The significant excess of recessive disease-causing genes in the middle part of the distribution in comparison to the constrained set ($p < 0.003$) underscores the idea that recessive inheritance models do not induce strong constraint.

Mutation rates for ASD and intellectual disability

We applied the model to two primary data sets: published results from ASD sequencing studies[2-6] with a collection of additional unpublished ASD trios, and published results from patients with severe intellectual disability[9,10]. **Table 3.1a** shows the comparison between the predicted number of variants per exome and the observed data from the 1,078 ASD cases as well as 343 sequenced unaffected siblings[2-6]. The model's predictions match the observed data for the unaffected siblings well, but the cases show a significant excess of *de novo* LoF variants consistent with the findings of the individual sequencing studies ($p = 2.05 \times 10^{-7}$). Using our model to simulate null DNV sets, we found that there are significantly more genes with two or more *de novo* LoF variants than would be expected by chance ($p < 0.001$, 6 observed when less than one was expected; **Table 3.1b**. Importantly, while we do not observe a global excess of *de novo* missense variants, we do observe an excess of genes with two or more functional (LoF or missense) *de novo* variants (observed 48 such genes when the average

expected is 27; $p < 0.001$) and genes with two or more *de novo* missense variants alone (observed 33 such genes when average expectation was 21, $p = 0.007$ for missense, **Table 3.1b**). No such excess of genes containing multiple DNVs was seen in the unaffected siblings (**Table 3.1b**). Of note, our framework also supports the assessment of many other weightings and combinations of alleles – such as missense variants only (optimal for pure gain-of-function disease models), predicted damaging missense variants only, and exact probability estimates for specific combinations of LoF and missense variants - than those shown above.

Table 3.1. Evaluation of the rates of *de novo* variants in ASD cases and unaffected siblings. The observed and expected rate of variants by type per exome for unaffected siblings[2] and ASD cases[2-6] (a). (b) The number of genes with multiple *de novo* variants in unaffected siblings and ASD cases across studies. The average number of expected genes with multiple *de novo* variants was determined by simulation. LoF = Loss-of-function. DNVs = *de novo* variants. For (a), a two-tailed test was performed for synonymous and missense; a one-tailed test for loss-of-function.

a) Genome-wide excesses of mutational events

Unaffected Siblings

Mutation Type	Observed events per exome	Expected events per exome	p-value
Synonymous	0.21	0.27	0.0218
Missense	0.61	0.62	0.8189
Loss-of-Function	0.09	0.09	0.4508

n = 343

ASD Cases

Mutation Type	Observed events per exome	Expected events per exome	p-value
Synonymous	0.25	0.27	0.1065
Missense	0.64	0.62	0.5721
Loss-of-Function	0.13	0.09	2.05×10^{-7}

n = 1,078

Table 3.1 (Continued)

b) Genome-wide excesses of multiply hit genes

Unaffected Siblings

Mutation Type	Observed genes with 2+ DNVs	Average expected genes with 2+ DNVs	p-value
Synonymous	0	0.5	1.0
Missense	5	2.5	0.1049
Loss-of-Function	0	0.04	1.0
LoF+missense	6	3	0.0779

n = 343

ASD Cases

Mutation Type	Observed genes with 2+ DNVs	Average expected genes with 2+ DNVs	p-value
Synonymous	4	3.8	0.5186
Missense	33	21.4	0.0070
Loss-of-Function	6	0.5	< 0.001
LoF+missense	48	27.2	< 0.001

n = 1,078

Table 3.2 lists all of the genes that have two or more *de novo* missense or LoF variants across the 1,078 ASD subjects. A conservative significance threshold of $1x10^{-6}$ was used, correcting for 18,271 genes and two tests. Considering this set of 1,078 trios as a single experiment, two genes (*DYRK1A* and *SCN2A*) exceeded this conservative genome-wide significance threshold for more *de novo* LoF variants than predicted. *SCN2A* also had significantly more functional *de novo* variants than expected. *CHD8*, with three *de* novo LoF variants and one missense, was very close to the significance threshold in these studies ($p = 1.76x10^{-6}$ for LoF; $p = 3.20x10^{-5}$ for functional). However, a recent targeted sequencing study found 7 additional *CHD8 de novo* LoF variants in ASD cases[11]. This brought the total number of *de novo* LoF variants in *CHD8* to 10, which was highly significant ($p = 8.38x10^{-20}$ when accounting for the total number of trios – 2,750 – examined in the combination of the targeted and exome-wide study).

These results offer the encouraging point that, as with genome-wide association studies (GWAS), larger collaborative exome efforts for trios will define unambiguous risk factors. It is important to note, however, that not all genes with a large number of *de novo* variants in them had significant p-values. For example, *TTN* had four missense DNVs in ASD cases, but a p-value that is not even nominally significant due to the enormous size of the gene ($p = 0.18$). Even having two *de novo* LoF variants was on occasion not enough to provide compelling significance (*POGZ*, two frameshifts, $p = 8.93 \times 10^{-5}$). In comparison, none of the genes found to contain multiple DNVs in the unaffected siblings crossed the significance threshold (**Table 3.3**).

Table 3.2. Significance of genes with multiple *de novo* variants (DNVs) in autism spectrum disorder (ASD) cases. Loss-of-function (LoF) mutations include nonsense, frameshift, and splice site-disrupting mutations. "# LoF Expected" refers to the expected number of *de novo* LoF variants based on the probability of mutation for the gene as determined by our model. The genome-wide significance threshold is $1x10^{-6}$. "." = no data available.

Gene	# LoF	# Missense	# DNVs Expected	p-value	Test
DYRK1A	3	0	0.0072	$6.15x10^{-8}$	LoF
SCN2A	3	2	0.0177	$9.20x10^{-7}$	LoF
CHD8	3	1	0.0221	$1.76x10^{-6}$	LoF
KATNAL2	2	0	0.0049	$1.19x10^{-5}$	LoF
POGZ	2	0	0.0134	$8.93x10^{-5}$	LoF
ARID1B	2	0	0.0178	$1.57x10^{-4}$	LoF
SCN2A	3	2	0.1334	$3.15x10^{-7}$	LoF+mis
CHD8	3	1	0.1724	$3.20x10^{-5}$	LoF+mis
SUV420H1	1	2	0.0602	$3.48x10^{-5}$	LoF+mis
PLEKHA8	0	2	0.0302	$4.46x10^{-4}$	LoF+mis
TUBA1A	0	2	0.0338	$5.59x10^{-4}$	LoF+mis
SLCO1C1	0	2	0.0394	$7.55x10^{-4}$	LoF+mis
NTNG1	0	2	0.0413	$8.29x10^{-4}$	LoF+mis
TSNARE1	0	2	0.0498	$1.20x10^{-3}$	LoF+mis
TBR1	1	1	0.0541	$1.41x10^{-3}$	LoF+mis
MEGF11	0	2	0.0552	$1.47x10^{-3}$	LoF+mis
KRBA1	0	2	0.0642	$1.98x10^{-3}$	LoF+mis
SRBD1	0	2	0.0645	$1.99x10^{-3}$	LoF+mis
KIRREL3	0	2	0.0652	$2.03x10^{-3}$	LoF+mis
NR3C2	1	1	0.0655	$2.05x10^{-3}$	LoF+mis
UBE3C	0	2	0.0775	$2.85x10^{-3}$	LoF+mis
AGAP2	0	2	0.0825	$3.22x10^{-3}$	LoF+mis
ABCA13	0	3	0.2890	$3.24x10^{-3}$	LoF+mis
ADCY5	0	2	0.1098	$5.61x10^{-3}$	LoF+mis
KIAA0182	0	2	0.1114	$5.76x10^{-3}$	LoF+mis
ZNF423	0	2	0.1131	$5.94x10^{-3}$	LoF+mis

Table 3.2 (Continued)

Gene	# LoF	# Missense	# DNVs Expected	p-value	Test
ZNF638	1	1	0.1212	6.78×10^{-3}	LoF+mis
SCN1A	0	2	0.1352	8.36×10^{-3}	LoF+mis
LAMB2	0	2	0.1604	1.16×10^{-2}	LoF+mis
MYO7B	0	2	0.1616	1.17×10^{-2}	LoF+mis
KIAA0100	1	1	0.1619	1.18×10^{-2}	LoF+mis
PLXNB1	1	1	0.1718	1.32×10^{-2}	LoF+mis
CACNA1D	0	2	0.1732	1.34×10^{-2}	LoF+mis
ZFYVE26	1	1	0.1753	1.37×10^{-2}	LoF+mis
SBF1	0	2	0.1808	1.45×10^{-2}	LoF+mis
BRCA2	0	2	0.1928	1.64×10^{-2}	LoF+mis
TRIO	0	2	0.2374	2.41×10^{-2}	LoF+mis
ALMS1	0	2	0.2422	2.50×10^{-2}	LoF+mis
RELN	1	1	0.2429	2.51×10^{-2}	LoF+mis
ANK2	1	1	0.2591	2.83×10^{-2}	LoF+mis
MLL3	1	1	0.3159	4.05×10^{-2}	LoF+mis
DNAH5	1	1	0.3219	4.19×10^{-2}	LoF+mis
FAT1	0	2	0.3343	4.49×10^{-2}	LoF+mis
GPR98	0	2	0.3761	5.53×10^{-2}	LoF+mis
AHNAK2	0	2	0.4172	6.62×10^{-2}	LoF+mis
SYNE1	0	2	0.5931	1.20×10^{-1}	LoF+mis
TTN	0	4	2.1947	1.80×10^{-1}	LoF+mis
MUC5AC	0	2	.	.	LoF+mis
RFX8	0	2	.	.	LoF+mis
EFCAB8	0	2	.	.	LoF+mis

Table 3.3. Significance of specific genes with multiple *de novo* variants (DNVs) in unaffected siblings. Loss-of-function (LoF) mutations include nonsense, frameshift, and splice site-disrupting mutations. "# LoF Expected" refers to the expected number of *de novo* LoF variants based on the probability of mutation for the gene as determined by our model. The genome-wide significance threshold is $1x10^{-6}$. "." = no data available.

Gene	# LoF	# Missense	# DNVs Expected	p-value	Test
CSNK1G3	1	1	0.0098	$4.78x10^{-5}$	LoF+mis
UGT2B4	0	2	0.0102	$5.12x10^{-5}$	LoF+mis
USP34	0	2	0.0717	$2.45x10^{-3}$	LoF+mis
AHNAK2	0	2	0.1327	$8.07x10^{-3}$	LoF+mis
SYNE2	0	2	0.1369	$8.56x10^{-3}$	LoF+mis
TTN	0	2	0.6983	$1.55x10^{-1}$	LoF+mis

These analyses were also applied to the results from the sequencing studies of moderate to severe (IQ < 60) intellectual disability[9,10] (n = 151). Intellectual disability, like ASD, showed a significant excess of LoF DNVs ($p = 6.49x10^{-7}$; **Table 3.4a**). Even with a much smaller sample size there were genes with significantly more LoF and functional DNVs than predicted by the model (**Table 3.4c**). The intellectual disability data also have significantly more genes with multiple *de novo* missense, LoF, and functional variants than predicted ($p = 0.009$ for missense, $p < 0.001$ for LoF and functional; **Table 3.4b**).

Table 3.4. Evaluation of the rates of *de novo* variants in cases with intellectual disability. (a) The observed and expected rate of variants by type per exome for cases of intellectual disability (ID, n = 151 families)[9,10]. A two-tailed test was performed for synonymous and missense; a one-tailed test for loss-of-function. (b) The number of genes with multiple *de novo* variants in intellectual disability cases across studies. The average number of expected genes with multiple *de novo* variants was determined by simulation. (c) Genes with multiple functional *de novo* variants in the ID cases[9,10]. LoF variants include nonsense, frameshift, and splice site-disrupting events. The genome-wide significance threshold is $1x10^{-6}$. The number of variants is either compared to the expected number for LoF only or for both LoF and missense, as indicated by the "# DNVs Expected" and "Test" columns. LoF = Loss-of-function. DNVs = *de novo* variants.

a) Genome-wide excesses of mutational events

Mutation Type	Observed events per exome	Expected events per exome	p-value
Synonymous	0.19	0.27	0.0267
Missense	0.70	0.62	0.2380
Loss-of-Function	0.24	0.09	$6.49x10^{-7}$

b) Genome-wide excesses of multiply hit genes

Mutation Type	Observed genes with 2+ DNVs	Average expected genes with 2+ DNVs	p-value
Synonymous	1	0.09	0.0879
Missense	3	0.5	0.0090
LoF	2	0.01	< 0.001
LoF+missense	6	0.6	< 0.001

c) Genes with multiple *de novo* missense and loss-of-function variants

Gene	# LoF	#Missense	# DNVs Expected	p-value	Test
SYNGAP1	3	0	0.0017	$8.15x10^{-10}$	LoF
SCN2A	3	1	0.0025	$2.56x10^{-9}$	LoF
SCN2A	3	1	0.0187	$5.01x10^{-9}$	LoF+mis
STXBP1	1	2	0.0071	$5.87x10^{-8}$	LoF+mis
TCF4	0	2	0.0069	$2.39x10^{-5}$	LoF+mis
GRIN2A	0	2	0.0162	$1.34x10^{-4}$	LoF+mis
TRIO	0	2	0.0333	$5.60x10^{-4}$	LoF+mis

In our ASD sample, we then investigated the rate of *de novo* events as a function of IQ; roughly 80% of this sample had an IQ assessment attempted. We found that the rate of *de novo* LoF mutation in ASD cases with a measured IQ above average was no different than expectation (IQ ≥ 100; n = 229; 0.08 *de novo* LoF variants per exome compared to expected 0.09, p = 0.59). By contrast, the rate in the rest of the sample was substantially higher than expectation (n = 572; rate of 0.17 *de novo* LoF variants per exome, $p = 1.17 \times 10^{-10}$). Furthermore, when directly compared (rather than to our expectation), these two groups were significantly different from each other, confirming a difference in genetic architecture among ASDs as a function of IQ (**Table 3.5a-b**, $p < 0.001$). These conclusions are unchanged in separate analyses of nonverbal and verbal IQ as well as full scale IQ (**Table 3.5c**).

Table 3.5. Investigating the rate of *de novo* mutation as a function of IQ. (a) The observed and expected rate of *de novo* variants by mutation class for the autism spectrum disorder cases with full scale IQ ≥ 100. (b) The observed and expected rate of *de novo* variants by mutation class for the autism spectrum disorder cases that did not have a full scale IQ above 100. (c) The observed rate of *de novo* loss-of-function (LoF) mutations split by verbal IQ and nonverbal IQ. For (a) and (b), a two-tailed test was performed for synonymous and missense; a one-tailed test for loss-of-function.

a) Full Scale IQ scored above 100 (n = 229)

Mutation Type	Observed events per exome	Expected events per exome	p-value
Synonymous	0.24	0.27	0.2346
Missense	0.66	0.62	0.4736
Loss-of-Function	0.08	0.09	0.5867

b) Full Scale IQ not scored above 100 (n = 572)

Mutation Type	Observed events per exome	Expected events per exome	p-value
Synonymous	0.22	0.27	0.0123
Missense	0.62	0.62	0.9946
Loss-of-Function	0.17	0.09	1.17×10^{-10}

c) IQ comparisons split between verbal and nonverbal IQ

Phenotypic Group	Number of samples	Observed *de novo* LoF events per exome	p-value
Verbal IQ ≥ 100	242	0.10	0.1903
Verbal IQ not scored above 100	712	0.15	2.43×10^{-8}
Nonverbal IQ ≥ 100	276	0.09	0.4829
Nonverbal IQ not scored above 100	678	0.16	1.09×10^{-9}

Gene set enrichment

Given the significant global excess of *de novo* LoF variants in ASD cases, we wanted to evaluate whether the set of genes harboring *de novo* LoF variants had significant overlap with several sets of genes proposed as relevant to autism or describing biochemical pathways. We used the probabilities of mutation to determine the fraction of LoF variants expected to fall into the given gene set. We then used the binomial distribution to evaluate the number of observed LoF variants overlapping the set compared to the established expectation. When we applied this analysis to a set of 112 genes reported as disrupted in individuals with ASD or autistic features, we observed no enrichment of *de novo* LoF variants (**Figure 3.2**, "Betancur")[12]. By contrast, we applied this analysis to a recent study of 842 genes found to interact with the Fragile X mental retardation protein (FMRP) *in vivo* and found a highly significant overlap (2.3-fold enrichment, $p < 0.0001$, **Figure 3.2**)[2,13]. This enrichment with the targets of FMRP held even when we removed the *de novo* variants identified in the Iossifov *et al* study that initially reported an enrichment of *de novo* variants in ASD cases with FMRP-associated genes (2.5-fold enrichment, $p < 0.0001$)[2].

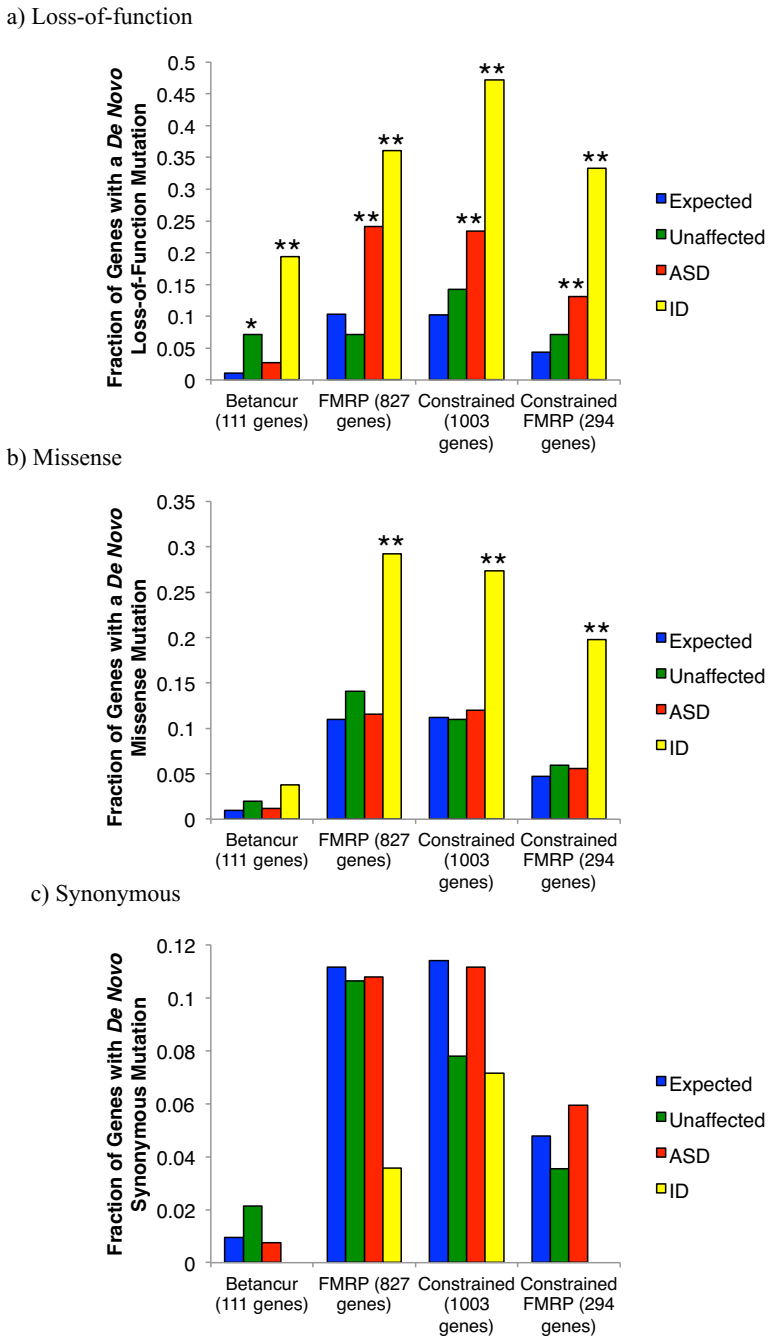


Figure 3.2. The expected and observed fraction of genes with a *de novo* variant in cases and controls for four gene sets of interest. ASD cases (n = 1,078), unaffected controls (n = 647), and intellectual disability (ID; n = 151) cases were sequenced across various studies[2-6,9,10]. "Betancur" refers to a set of genes reported as disrupted in individuals with ASD or autistic features; of the 112 on the list, we could evaluate 111[12]. "FMRP" refers to the genes whose mRNAs are bound and regulated by the Fragile X

Figure 3.2 (Continued) Mental Retardation Protein (FMRP), as identified by Darnell and colleagues[13]. The "constrained" category is a set of 1,003 genes that we defined as significantly lacking rare missense variation, indicating intolerance to mutation. The targets of FMRP that are also considered constrained by our metric make up the "Constrained FMRP" category. Loss-of-function variants are presented in (a); missense in (b) and synonymous in (c). * indicates $p < 0.01$; ** indicates $p < 10^{-4}$.

We then evaluated the group of individuals from the ASD studies who had a *de novo* LoF variant in one of the targets of FMRP. On average, these cases were enriched for having a measured IQ < 100 (Fisher's exact $p = 4.01 \times 10^{-4}$; **Table 3.6** as well as significantly reduced male:female ratio ($p = 0.02$; **Table 3.7**) as compared to the remaining sequenced cases (Materials and Methods). These individuals represent about 3% of the total sample, when at most a 1% overlap would be expected. The estimated odds ratio (OR) of *de novo* LoF variants in the set of FMRP target genes was around 6, very similar to the OR estimated for large CNVs that disrupt multiple genes[14]. In addition, the OR for the published cases of moderate to severe intellectual disability noted above (IQ < 60; not ascertained for ASDs) having a *de novo* LoF event in the set of FMRP targets was roughly 10.

Table 3.6. The number (and percentage) of individuals that have an IQ ≥ 100 or an IQ not scored above 100 split by whether they contain a *de novo* loss-of-function variant in a target of FMRP (FMRP-I) or not ("Rest of Cases"). In (a), individuals who started an IQ test but were not given an IQ score are included. Only individuals with IQ scores are included in (b).

a) IQ Attempted but unscored individuals included

	FMRP-I	**Rest of Cases**
IQ ≥ 100	1 (3%)	254 (31%)
IQ not above100	29 (97%)	575 (69%)

Fisher's exact p-value = 4.01×10^{-4}

b) Only scored individuals

	FMRP-I	**Rest of Cases**
IQ ≥ 100	1 (5%)	254 (35%)
IQ not above100	20 (95%)	469 (65%)

Fisher's exact p-value = 0.0021

Table 3.7. The number (and percentage) of individuals that are male and female split by containing a *de novo* loss-of-function mutation in a target of FMRP (FMRP-I) or not ("Rest of Cases").

	FMRP-I	**Rest of Cases**
Male	19 (63%)	658 (80%)
Female	11 (37%)	163 (20%)

Chi-square p-value = 0.02

The same analysis was applied to the list of *de novo* LoF variants from unaffected siblings of ASD cases and additional control individuals (n = 647)[2,4,5,15]. There was a significant enrichment when evaluating the overlap with the set of autism related genes (p = 0.0095, **Figure 3.2**). However, no significance was observed for the overlap with the *in vivo* targets of FMRP. The *de novo* LoF variants from the intellectual

disability individuals, on the other hand, were significant for both sets ($p < 10^{-4}$ for both sets; **Figure 3.2**). Even the *de novo* missense variants found in the intellectual disability cases showed significant overlap with both sets under study ($p = 0.02$ for autism-related genes, $p < 0.0001$ for the targets of FMRP, **Figure 3.2**).

Evaluating constrained genes

We further applied the enrichment analysis to our set of constrained genes and found that they contained more *de novo* LoF variants than expected by chance (2.3-fold enrichment, $p < 0.0001$, **Figure 3.2**). A greater fold enrichment was observed when focusing on the subset of constrained genes that were also identified in the FMRP study (3.0-fold enrichment, $p < 0.0001$, **Figure 3.2**)[13]. We note that the FMRP targets have a significant overlap with the constrained set of genes (odds ratio = 1.29, $p < 0.0001$), which is consistent with the report that the targets of FMRP are under greater purifying selection than expected[2]. All enrichments were demonstrated to be independent of gene size (Materials and Methods).

The genes that contained a *de novo* missense or LoF variant in the cases of intellectual disability also showed a significant enrichment for both the constrained gene set and the set of constrained targets of FMRP ($p < 0.0001$ for all lists). In comparison, no enrichment was found with either set and the list of genes that had a *de novo* LoF variant in unaffected siblings and control individuals.

In addition to treating constraint as a dichotomous trait, we also evaluated the missense Z score for each of the genes with a *de novo* LoF variant. We found that the

distribution of missense Z scores for genes with a *de novo* LoF variant in unaffected individuals was no different from the overall distribution of scores (Wilcoxon p = 0.8325; **Figure 3.3**). By contrast, both the genes with a *de novo* LoF variant in ASD and intellectual disability cases had values significantly shifted towards high constraint (Wilcoxon $p < 10^{-6}$ for both). Furthermore, we compared the distribution of Z scores among each of the three groups. Both the ASD and intellectual disability distributions were significantly different from the distribution of missense Z scores for unaffected individuals (p = 0.0148 and 0.0012, respectively). The intellectual disability missense Z scores were also significantly higher than the ASD values (p = 0.0319).

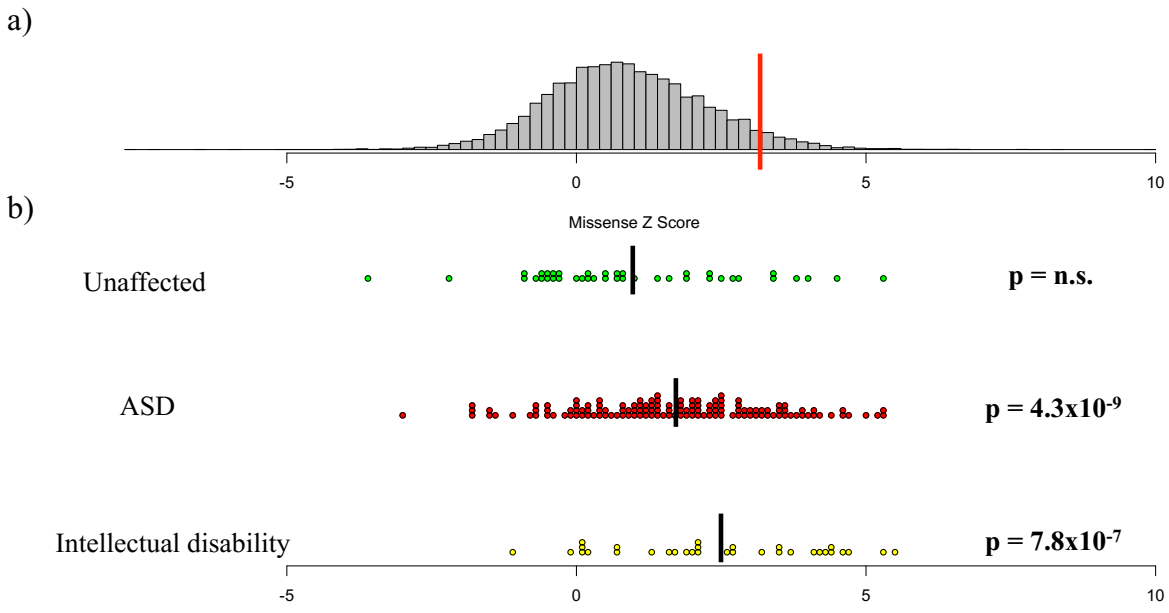


Figure 3.3. The distribution of missense Z scores and Z scores of genes containing *de novo* loss-of-function variants identified in unaffected individuals, autism spectrum disorder (ASD) cases, and intellectual disability cases. (a) The distribution of missense Z scores. The red line indicates a Z score of 3.09, or the threshold for inclusion into the set of 1,003 constrained genes. (b) The missense Z scores for genes containing *de novo* LoF in unaffected individuals, ASD cases, and intellectual disability cases[2-6,9,10,15]. Black bars indicate the mean Z score of each group: 0.94, 1.68, and 2.46 for unaffected

Figure 3.3 (Continued) individuals, ASD cases, and intellectual disability cases, respectively. While the missense Z scores of the *de novo* LoF variants found in unaffected siblings matched the overall distribution (Wilcoxon p = 0.8325, n.s. = not significant), *de novo* LoF variants found in both ASD and intellectual disability cases were significantly shifted towards more extreme constraint values ($p < 10^{-6}$ for both). All p-values for deviation from the overall distribution are listed on the right side of the figure. In addition, the distribution of missense Z scores fore each of the three *de novo* lists were all individually significant at $p < 0.05$.

When evaluating the ASD cases split by IQ group, we found no enrichment of *de novo* LoF-containing genes with either constrained genes and targets of FMRP in the group with IQ ≥ 100 ($p > 0.5$ for both sets of genes) but very strong enrichment in the set with IQ < 100 ($p < 0.0001$ for both sets of genes). These results underscore that phenotypically distinct subsets of ASD cases may have significantly different contributions from *de novo* variation.

Comparison of constraint metric with existing methods

Identifying constrained genes by comparing observed nonsynonymous sites to expectation is conceptually similar to the traditional approach of detecting selective pressure by comparing observed nonsynonymous sites to observed synonymous sites (e.g. d_N/d_S) that has been used extensively. Our approach should in principle achieve greater statistical power to detect constrained genes; comparison of an observation to expectation is statistically more powerful than contrasting that observation with a generally smaller second observation – the number of observed synonymous variants. In order to investigate this claim, we identified genes that had significant evidence for selective constraint using the d_N/d_S metric (i.e. their ratio of synonymous and

nonsynonymous sites deviated from the genome-wide average at $p < 0.001$, Materials and Methods). There were only 377 of these genes, over half of which overlapped with the constrained gene list defined by our method (n = 1,003; overlap 237 genes). The genes identified as significantly constrained by only our metric (the top 10 of which include *RYR2*, *KMT2A* (*MLL*), *KMT2D* (*MLL2*), and *SYNGAP1*) are still significantly enriched for known causes of autosomal and X-linked dominant forms of mendelian disease ($p = 5x10^{-4}$). We therefore conclude that the model-based approach to identifying constrained genes adds substantial power to traditional approaches. The importance of this increased power to detect constraint in further articulated in the ASD and intellectual disability analyses below.

Several groups have previously published approaches, and specific gene sets from them, that are also aimed at identifying genes under excessive purifying selection or generally intolerant of functional mutation. Bustamante *et al*[16] expanded on the McDonald-Kreitman framework[17] contrasting fixed differences in the primate lineage to polymorphic differences in humans to identify a set of genes under weak negative selection, while more recently Petrovski *et al*[8] utilized the excess of rare versus common missense variation within humans to flag genes intolerant of functional variation. We found a reasonable correlation between our metric of constraint and Petrovski's Residual Variation Intolerance Score (RVIS[8]; **Figure 3.4**). A comparison of these approaches as applied to prioritization of known haploinsufficient genes as well as the autism *de novo* LoF variants described here are provided in the Materials and Methods and demonstrates that the two human-only approaches (constraint and RVIS) perform better on these tasks of identifying medical genetic lesions of severe effect in

modern humans (**Table 3.8**). Intriguingly, both of these other approaches utilize independent information from each other and from our approach (which uses the absence of rare functional variation versus expectation within humans), raising the potential that composite scores employing all three sources of information pointing to which genes are most sensitive to heterozygous mutation could add further value.

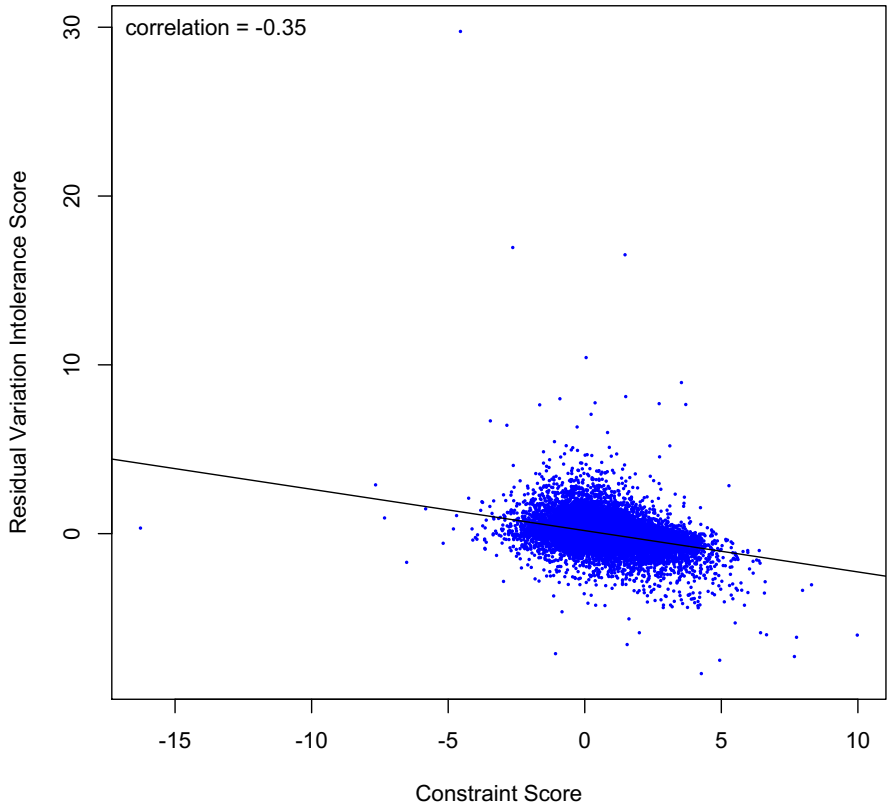


Figure 3.4. Correlation between the constraint score and RVIS. The constraint scores (missense Z scores) determined by our method and residual variation intolerance score from Petrovski *et al* [8] have a Pearson correlation of -0.35. The black line shows the linear regression between the two metrics.

Table 3.8. Comparison of the predictive ability of different sets of constrained genes for known haploinsufficient genes and those disrupted by a *de novo* LoF mutation in ASD patients. In (a), the ability of both constraint scores and lists of constrained genes were tested for their ability to predict known haploinsufficient genes, as listed in OMIM. The quantitative scores (constraint and RVIS[8]) were used in a linear regression with gene size added as a covariate. The gene lists (constrained, top 5.5% most intolerant genes using RVIS[8], and the genes identified in Bustamante *et al*[16]) were evaluated with a logistic regression with gene size as a covariate. In (b), the three gene lists were evaluated for their enrichment of *de novo* LoF mutations identified in ASD patients. To do this, the expected fraction of constrained genes to contain one of these *de novo* mutations was determined and then used to establish the fold enrichment and significance of the observed fraction.

a) Linear and logistic regressions

		Quantitative Scores			List-Based		
		Constraint score	**RVIS**		**Top Constrained**	**Top RVIS**	**Bustamante**
OMIM Haplo-insufficiency	t-value	10.011	-9.561	OR	4.909	5.490	1.307
	p-value	$< 10^{-16}$	$< 10^{-16}$	p-value	$< 10^{-16}$	$< 10^{-16}$	0.191

b) Enrichment of genes with those containing a *de novo* LoF in ASD patients

		Top Constrained	**Top RVIS**	**Bustamante**
ASD *de novo* LoF	Fold enrichment	2.282	1.904	0.836
	p-value	3.58×10^{-6}	5.36×10^{-5}	0.718

Discussion

We have developed a framework for evaluating excesses of *de novo* variants identified through exome sequencing. Even though this framework can be leveraged to evaluate excesses of variants study-wide and in gene sets, the key focus is to evaluate the significance for individual genes. Given the small number of observed *de novo* events per gene, simple case-control comparisons cannot achieve any meaningful level

of significance. For example, observing three *de novo* loss-of-function variants in a small gene in 1,000 case trios is perhaps quite compelling, especially if no such variants were identified in 1,000 control trios. However, a simple three to zero case/control comparison in this situation would yield no compelling statistical evidence (one-tailed p = 0.125). Incidence of such extremely rare events, however, can be evaluated if the expected rate of such events is known. Sequencing large numbers of control trios to gather empirical rate estimates on a per-gene basis that are accurate is infeasible and inefficient. The calibrated model and statistical approach described here can achieve a close approximation of this ideal. Our method, therefore, offers the ability to evaluate the rate of rare variation in individual genes in situations where burden tests would fail.

Other groups have developed similar statistical frameworks[11,18] – notably, the Epi4k consortium[18] used the same base model we begin with[3] to interpret event rates. Our model, however, has two primary strengths. First, our model of *de novo* mutation incorporates additional factors beyond sequence context that affect mutation rate. Both the depth of coverage – how many sequence reads were present on average – for each base and the regional divergence around the gene between humans and macaques independently and significantly improve the predictive value of our model (Materials and Methods). Second, given the high correlation between the number of rare synonymous variants in ESP and the probability of a synonymous mutation determined by our full model, we have a metric to evaluate the extent to which genes in the human genome show evidence of selective constraint. The list of 1,003 genes that we define as constrained contains an enrichment of genes known to cause severe human disease – an observation analogous to that recently made in using empirical comparison of

common and rare rates of functional variation to evaluate intolerance[8]. In fact, site count deficits and site frequency shifts each contribute independent information to the definition of constraint and can in principle be combined in a composite test.

The results of our metric were compared to both the scores created by Petrovski and colleagues[8] and loci identified as under negative selection by Bustamante *et al* [16] . Overall, our metric and the residual variation intolerance scores defined by the Petrovski worked similarly well, reinforcing the benefits that could come from combining the two approaches. It is unsurprising that these methods outperform the evolutionary ones on the specific matter of genes intolerant to heterozygous mutation. Evolutionary methods examining differences between polymorphism and fixed differences, which are more sensitive to weaker negative selection, require that mutations be tolerated well enough to become polymorphic in the first place. By contrast, approaches measuring the complete absence of variation will pick up the most strongly intolerant genes.

Ideally, we can conceptualize defining two metrics of genic constraint, one based on missense variants and the other based on LoF variants. With only 6,503 individuals in ESP, we are underpowered to determine significant deviations for most genes with respect to loss-of-function variants. As sample size increases, our ability to calculate constraint improves. For example, if the sample size were to increase by an order of magnitude, we would be able to evaluate approximately 66% of genes using LoF variants. We therefore view the constrained gene list as a work in progress, to be updated when larger exome sequencing data sets become available.

Applying our statistical framework to *de novo* variants from 1,078 ASD cases reveals that, while there is no global excess in *de novo* missense variants, there are

significantly more genes that contained multiple *de novo* missense variants than expected. We also see significant overlap between the list of genes with a *de novo* LoF in ASD cases and the set of constrained genes that we defined. In addition, there is a significant overlap between the genes with a *de novo* LoF variant and the targets of FMRP, as reported in Iossifov *et al*[2]. All of the significant signals in ASD – the global excess of *de novo* LoF variants, the excess of genes with multiple functional *de novo* variants, the overlap between the *de novo* LoF genes and both constrained genes and the targets of FMRP – are not found in the subset of ASD cases with IQ ≥ 100. The lack of signal in the IQ ≥ 100 indicates that genetic architecture among ASDs varies as a function of IQ. Overall, the probabilities of mutation defined by our full model and list of constrained genes can be used to critically evaluate the observed DNVs from sequencing studies and aid in the identification of variants and genes that play a significant role in disease.

Materials and Methods

De novo variant information

Published *de novo* variants were collected for both autism spectrum disorders (ASD)[2-6] and severe intellectual disability[9,10]. Updated *de novo* calls were provided from two of the ASD studies[3,5]. Details about sample collection, sequencing, and variant processing can be found in the separate studies.

Additional sequencing

Exome sequencing of the additional families (n = 129) was performed at the Broad Institute. Exons were captured using the Agilent 38Mb SureSelect v2. After capture, a round of ligation-mediated PCR was performed to increase the quantity of DNA available for sequencing. All libraries were sequenced using an IlluminaHiSeq2000. Data were processed with Picard (http://picard. sourceforge.net/), which uses base quality-score recalibration and local realignment at known indels[19] and BWA[20] for mapping reads to hg19. SNPs were called using GATK for all trios jointly[19,21]. The variable sites that we have considered in analysis are restricted to those that pass GATK standard filters. From this set of variants, we identified putative *de novo* variants and validated them as previously described[3]. Autism Consortium samples (n = 78 trios) were collected in Boston under IRB approval from Harvard Medical School, Massachusetts General Hospital, Children's Hospital Boston, Tufts-NEMC, Boston University Medical Center with ADI and ADOS assessment. Finnish autism samples (n = 51 trios) were collected under IRB approval at University of Helsinki with ADI and ADOS assessment and consented for autism research only. In both studies, all participants gave written informed consent, though as autism is classified as a childhood disorder, many subjects are children with informed consent provided by parents or guardians.

Mutational model

We wanted to create an accurate model of *de novo* mutation for each gene. The steps involved in the creation of the model are outlined in **Figure 3.1**. Briefly, we

determined the probability of a given base mutating into one of the three other possible bases as well as the coding impact of each possible mutation. We added probabilities across a gene to create per-gene probabilities of all mutation types under study: synonymous, missense, nonsense, and splice site.

The first, and most important, step of making a model based on sequence context is to establish the mutability of a given base. Krawczak and colleagues determined that the best context for determining the mutability of a single base is to include both the 5' and 3' bases[22]. Following the lead of other groups, we took this trinucleotide context as sufficient for determining mutability[23]. We used 1000 Genomes intergenic regions that are orthologous between humans and chimps as the basis for our mutation rate table. Across the sequence, we tallied the number of observations for each of the 64 possible trinucleotides and, for each SNP, considered the chimp allele to be ancestral and determined the trinucleotide (XY_1Z) to trinucleotide (XY_2Z) change that occurred. To determine the probability of a given trinucleotide mutating, we divided the number of mutations in that trinucleotide context by the number of occurrences of the trinucleotide. This probability is adjusted by a proportionality constant, λ, that gives the mutation rate of that trinucleotide for a single generation. The mutation rate for the given nucleotide is then proportionally divided between the three possible trinucleotides to which it could mutate. In the end, we have a mutation rate table that contains the probability of any of the 192 possible mutations.

We then use the mutation rate table and the sequence context to determine the per-gene probability of mutation based on the sequence of the gene. For a given base in the gene, the trinucleotide sequence context is determined. The probability of the

middle base mutating to one of the three other bases is queried in the mutation rate table and the type of change it would create is determined. The probability of mutation is added to a running total for the type of mutation it would cause. This is repeated for the two other possible mutations for every coding base in the gene as well as the bases in the conserved splice sites for all genes in the genome. In the end, there is a per-gene probability of each type of mutation under study: synonymous, missense, nonsense, and splice site. We determine the probability of a frameshift mutation by multiplying the probability of a nonsense mutation by 1.25, the relative rate of singleton frameshift to singleton nonsense mutations found in exome sequencing data from roughly 2,000 ASD cases and controls.

Adjustments to the model

In order to evaluate the predictive value of the model of *de novo* mutation probability, we extracted the number of synonymous singletons – seen only once in the data set – found in each gene from the National Heart, Lung and Blood Institute's Exome Sequencing Project (ESP). The number of these singletons in each gene was correlated to both gene length and the probability of synonymous mutation determined by our model. While gene length alone showed a high correlation with the number of synonymous singletons (0.835), the probability of a synonymous mutation was significantly higher (0.854, $p < 10^{-16}$)

Depth adjustment

We first investigated the role that depth of coverage could have on the predictions of mutation rates. The ability to call a *de novo* event is dependent on how well sequenced the location of the event is. Therefore, bases that are not covered at all should not contribute to the overall probability of mutation for the gene. In order to account for differences in sequencing coverage, we created a way to determine what fraction of a base's mutation probabilities should be added to the total for the gene based on the coverage. For each base, we looked up the number of trios in which all members had 10x coverage or greater and used that number to determine the appropriate discount. For bases with almost all trios having 10x coverage, the probability of mutation was not adjusted. However, as the number of trios with 10x coverage dropped, the probability of mutation was multiplied by an adjustment factor in between 0.9 and 1. To determine the endpoints of the adjustment, we compared the ratio of the observed number of synonymous singletons to the overall probability of a synonymous mutation for a high confidence set of bases to sets of bases with fewer trios passing at 10x. The depth adjusted probabilities of synonymous mutation showed a significantly greater correlation to the number of synonymous singletons in the ESP data set when compared to gene length alone (0.891, $p < 10^{-16}$).

Divergence adjustment

Divergence between humans and other primates is known to correlate with the relative number of SNPs in large regions[24]. We postulated that local divergence rates could be added to the model as a regional term that captured the local deviation from

the base mutation rate. We used human-macaque divergence information to determine the divergence score – defined as the number of divergent sites over screened sites for the region containing the gene as well as 1 MB upstream and downstream – for each gene. We used linear models to determine the best equation to adjust the per-gene probabilities of mutation to incorporate the divergence score. In the end, the probability of mutation is adjusted slightly for the divergence score. For genes with no divergence information, the average divergence score is used. This, however, lead to a global increase in the predicted rate of mutation, so all probabilities of mutation were modified so that the sum of all probabilities after divergence adjustment was equal to the sum of probabilities from before the adjustment. This adjustment of predictions significantly increased the correlation with the synonymous singletons in the ESP data (0.910, $p < 10^{-16}$).

Replication timing adjustment

Replication timing has also been associated with overall mutation rate, with later replicating DNA having a higher rate of mutation[25]. We used replication timing Z scores from Koren *et al* to create a replication timing score for each gene[26]. The replication timing score is defined as the average replication timing score across the length of the gene. The replication timing score was used in linear models. It did significantly add to the mutational model ($p = 0.005$), but had a very slight overall effect. Further investigation revealed that the model was predicting more synonymous changes as the average replication Z score increased, and thereby was already accounting for the

adjustments that the replication score was adding. We did not include the replication timing adjustment in any further analyses.

Using rare variants instead of singletons

To increase power for our definition of constrained genes, we extracted the number of rare (minor allele frequency < 0.01%) synonymous variants found in each gene in the ESP data set. The correlation between the number of rare synonymous variants and the gene length was 0.880; the probability of synonymous mutation as defined by our full model and the number of rare synonymous variants was 0.940. Due to the stochastic nature of small counts in the ESP data set, the maximum correlation we could achieve is 0.975, indicating that our model captured ~66% of the remaining correlation that we could achieve above gene length.

Definition of constrained genes

A traditional approach to identifying genes that appear to be under constraint is to compare the ratio of nonsynonymous to synonymous substitutions (known as the K_a/K_s or d_N/d_s). Given that the correlation between the probability of a synonymous mutation and the number of rare synonymous variants in a gene was high, we wanted to use our model to predict the number of rare missense variants as a way to evaluate genes under constraint in an approach similar to the K_a/K_s. We determined the expected number of variants by fitting a linear model based on the probability of mutation and the observed number of synonymous variants. The autosomes were fit separately from the

X chromosome. The equations were applied using the probability of a missense mutation to create an expected number of rare missense variants in the ESP dataset. For both synonymous and missense variants, we created a signed Z score of the chi-squared deviation of observation from expectation. Negative values indicate more variants than expected, while positive values are tied to fewer variants observed than expected.

In order to define the set of genes that appeared to be under excessive constraint, we used three filters: (1) the predicted number of rare synonymous variants should be 5 or greater, (2) the observed number of rare synonymous variants should not be significantly lower than expectation ($p > 0.001$), and (3) the observed number of missense singletons should be significantly lower than expectation ($p < 0.001$). The reason for restricting to genes with 5 or more expected synonymous singletons is so that true deviations from expectation can be separated from deviations caused by sampling problems. Using these filters, we identified 1,003 genes—which represent roughly 5% of the genes in the genome—that appeared to be under excessive constraint.

The genes in the constrained gene list are enriched for entries in the OMIM database, especially for entries associated with mental retardation and retinitis pigmentosa. 31% of the top 86 constrained genes – for which the observed number of missense rare variants is significant at $p < 10^{-6}$ – have entries in the Online Mendelian Inheritance in Man (OMIM) database with dominant or *de novo* inheritance patterns. None of them have recessive inheritance entries in OMIM. A comparison set was made to 111 genes for which the missense observations fell very closely around prediction

(-0.01 < Z < 0.01). This set of genes had 2 OMIM entries (1.8%) with dominant or *de novo* inheritance and 11 (10%) with recessive inheritance.

Removing potential false positive constrained genes

In order to identify genes that appeared to be significantly constrained, we used our probabilities of mutation to predict the expected amount of synonymous and nonsynonymous variation in the NHLBI's ESP data. Those genes that had the expected amount of synonymous variation, but were significantly ($p < 0.001$) deficient for missense variation were labeled as constrained. To ensure that genes were not nominated as being constrained erroneously, we excluded from all analyses 134 genes where the observed synonymous and nonsynonymous rates were both significantly elevated or significantly depressed (both $p < 0.001$). Upon inspection, this list contained a number of genes that contained an internal duplication (e.g. *FLG*), a nearby pseudogene (e.g. *AHNAK2*), and a number of cases where recent duplications and/or annotation errors have led to the same sequence being assigned to two genes (e.g. *SLX1A* and *SLX1B*). These are all scenarios where standard exome processing pipelines systematically under-call variation – reads are unmapped due to uncertainty of which gene to assign them to – or overcall false variants owing to read misplacement. This further suggests that a byproduct of this analysis framework is the identification of a residual set of challenging genes for current exome sequencing pipelines.

Evaluating the global excesses of *de novo* variants

To compare the observed rate of *de novo* variants by mutation type to the expected rate, we summed the total probability of the given type of mutation and adjusted for the number of individuals in the study. Poisson distribution probabilities were invoked to determine the significance of the observation.

Number of genes with multiple *de novo* variants

Even though there is a global excess in LoF variants in the ASD cases, the signal was spread over many genes, making it hard to determine which specific genes may be contributing to the etiology of ASD. One way to prioritize genes would be to focus on those genes that contain multiple *de novo* variants; we wanted to evaluate whether there was an excess of such genes. To do so, we simulated *de novo* events by extracting each gene's probability of mutation and then randomly drew the expected number of *de novo* variants based on weight (the probability). Using these simulations, we could determine an empirical p-value for the observed number of genes with multiple *de novo* variants. Results are presented in **Table 3.1b** for the unaffected siblings and ASD cases, and in **Table 3.4b** for intellectual disability cases. The "LoF+missense" category uses the combined probability of a LoF and missense mutation to evaluate genes that show two or more *de novo* mutations that are LoF, missense, or both. The lowest possible p-value is 0.001 since 1,000 simulations were run.

Single genes with multiple *de novo* variants

Since we generated a per-gene probability of *de novo* mutation, we can directly evaluate genes that contain multiple *de novo* variants for significance. To do so, each gene's probability of mutation is extracted and the predicted number of *de novo* variants by mutation type is determined by adjusting for the number of individuals in the study. The observed and expected numbers of *de novo* variants are compared and the Poisson is invoked to determine significance. We perform two comparisons: the LoF mutations alone and the LoF and missense mutations together. The first comparison is only made for those genes that contain multiple LoF *de novo* mutations; the second is performed for genes that have a combination of missense and LoF *de novo* mutations. Here, we have set the significance threshold at 10^{-6} since it conservatively accounts for both the number of genes under study and the number of tests using the Bonferroni correction.

Global *de novo* mutation rates separated by IQ group

Due to the significant role of *de novo* variation in intellectual disability, we wanted to investigate the overall rates of mutations for those ASD cases without intellectual impairments. Several intelligence tests were used to assess proband IQ across testing sites. The IQ analyses presented here include individuals whose IQ was measured using one of four standardized, commonly used tests to evaluate intelligence in children: the WISC-IV[27], the WASI[28], the WPPSI-III (preschool and primary school age)[29], and the DAS (early years and school age)[30]. These tests provide comparable assessments of full scale intelligence, using both verbal and nonverbal assessments[31]. Children who

did not complete one of these four tests (n = 95, 10.0%) were treated as missing without attempt. Probands who are missing IQ without attempt include those who were given an IQ test that does not assess intelligence comparably (n = 78, 8.2%), specifically the Mullen Scales of Early Learning or the Leiter International performance scale, which are strongly weighted towards nonverbal assessment[32,33].

We had access to phenotypic information for 954 of the sequenced probands. Of these, 859 had taken an IQ test that could be compared to other tests. We removed those individuals that had a 30-point or greater difference between their verbal and nonverbal IQs to avoid inclusion of excess measurement error or learning disabilities. Verbal and nonverbal IQ were correlated strongly with each other ($r = 0.70$, $p < 0.0001$) as well as with the full scale IQ score (verbal IQ: $r = 0.89$, $p < 0.0001$; nonverbal IQ: $r = 0.93$, $p < 0.0001$). We separated the remaining 801 probands into those with and without measured IQs above statistical average. It is common for individuals affected with ASDs to be unable to complete or be scored on an IQ test; this was the case for 14.3% (n = 115) of probands for whom a test was attempted in the Simons sample. In the Simons Simplex Collection, probands who attempted to complete an eligible IQ test, but did receive a score, had significantly lower scores on the Vineland Scales of Adaptive Behavior (IQ test scored mean = 76.0, IQ test not scored mean = 60.3; $t = 15.9$, $p < 0.0001$). A Vineland composite standard score of 60 reflects adaptive behavior (overall functioning and self care skills) scores nearly three standard deviations below the mean, or approximately in the lowest 1% of the general population, controlling for age. As the inability to complete an IQ test is associated with case severity, we were specifically interested in estimating the de novo rate among individuals with both IQ

above the general population mean and the behavioral capability to complete an IQ test—both indicators of higher functioning ASDs. The observed and expected *de novo* variants per exome are listed in **Table 3.5a-b**. The individuals with full scale IQ ≥ 100 matched expectation for *de novo* variants per exome. Those individuals without measured IQs over 100, on the other hand, showed a global excess in *de novo* LoF variants. The results were similar when verbal and nonverbal IQ were analyzed separately (**Table 3.5c**). There was no excess of *de novo* LOF mutation in individuals with verbal ($p = 0.19$) or nonverbal ($p = 0.48$) IQ greater than 100.

Overlap between gene sets of interest and *de novo* containing genes

A number of gene sets have been proposed as relevant to autism or descriptive of an ASD biochemical pathway. Given the global excess of *de novo* LoF variants, we wanted to evaluate whether or not the list of genes that contain such mutations overlap more than expected with several of the proposed gene sets.

In order to determine the significance of any observed overlap between a gene set of interest and the list of genes that contain *de novo* variants, we first determine the total probability of mutation for all genes on the gene set of interest. The set total is compared to the total probability of mutation for all genes. This percentage becomes the expected overlap of *de novo* variants with the gene set. Using the expected overlap and the number of variants on the *de novo* list, we evaluate the observed overlap between the *de novo* list and the gene set of interest by invoking the binomial. All p-values are one-tailed. The *de novo* variant list is broken down by mutation type (LoF, missense, and synonymous), as are the probabilities of mutation for the gene set of interest.

We evaluated the overlap between three *de novo* lists and four separate gene sets of interest (**Figure 3.2**). The gene sets of interest are a set of genes reported as disrupted in individuals with ASD or autistic features (Betancur)[12], the set of targets of FMRP identified by Darnell and colleagues (FMRP)[13], the set of significantly constrained genes that we defined earlier (Constrained), and the set of FMRP targets that are also constrained (Constrained FMRP). Significance was conservatively set at 0.01.

Phenotype of individuals with *de novo* LoF mutations in FMRP targets

Across the 1,078 individuals with ASD, there were 35 *de novo* LoF variants in targets of FMRP spread across 34 individuals (referred to as FMRP-I here)[13]. For those individuals for which we had access to phenotypic information, we extracted IQ and sex. We found that the FMRP-I group had significantly fewer individuals with IQ ≥ 100 than the rest of the sample set (**Table 3.6a**, Fisher's exact $p = 4.01 \times 10^{-4}$). As before, individuals who started an IQ test but were not given an IQ score due to being severely impaired are included in the IQ < 100 group. To ensure that the association was not driven by those probands with attempted but missing IQ values, we also tested the association using only those individuals with estimated full scale IQ scores (**Table 3.6b**, Fisher's exact $p = 0.0021$). The FMRP-I group also had a reduced male bias. Where the whole set of individuals is ~80% male, the FMRP-I group is only ~59%, which is a significant difference (**Table 3.7**, Chi-square $p = 0.02$).

Comparing the power of our constraint method to that of NS:S ratio

The ratio of nonsynonymous (NS) substitutions per NS site to synonymous (S) substitutions per S site in a gene has been often used to determine if that gene has evidence of selection acting on it. A high NS:S ratio would indicate positive selection, while a low NS:S ratio would be evidence for purifying selection. Theoretically, our method of comparing observed NS variants to expectation should achieve greater statistical power than the NS:S comparison. To support this claim, we used the number of NS and S rare variants (minor allele frequency < 0.1%) found in the NHLBI's Exome Sequencing Project (ESP) dataset and determined each gene's deviation in terms of their ratio of S to NS sites compared to the genome-wide average.

We removed the 134 genes where the observed synonymous and nonsynonymous rates were both significantly elevated or significantly depressed from expectation as determined by our model (both $p < 0.001$). These poorly sequenced or mapped genes – as mentioned in the main text – were also removed from our analysis to define constrained genes. We then identified the remaining genes that were as deviant from the genome-wide average as the constrained genes we defined with our model were from expectation ($p < 0.001$). Compared to the 1,003 genes defined as constrained by our model, this approach only identified 377 genes that showed evidence of purifying selection, 237 (~63%) of which were also identified as constrained by our method. Included in the 766 genes considered constrained only by our metric were a number of genes – the top ten of which include *RYR2*, *KMT2A* (*MLL*), *KMT2D* (*MLL2*), and *SYNGAP1* – that have already been established as causes of autosomal or X-linked dominant forms of Mendelian disease (OMIM enrichment $p = 5 \times 10^{-4}$).

Since our metric was able to identify more genes that showed evidence of selective constraint, and especially since some of those are known to be causes of Mendelian disease, we conclude that our method of identifying constrained genes adds substantial power to the traditional approach and is an appropriate metric.

Comparison of constrained genes to the RVIS metric

Recently, Petrovski *et al* published a similar method to search for genes that appeared to be intolerant of mutations[8]. Their method evaluates the shift in the allele frequency spectrum of variants identified in genes in the ESP dataset to identify genes that have more rare variation. Specifically, the number of common nonsynonymous variants found in each gene was regressed against the total number of variants to determine the intolerance score. Genes with an unusually high ratio of rare to common variation are more likely to be intolerant of mutations and are assigned a lower residual variation intolerance score (RVIS). This approach is orthogonal to our metric of constraint since we search for a deficiency of rare nonsynonymous variation.

We took the intersection between the two datasets to compare our metric with the scores provided in Petrovski *et al*[8]. This process eliminated some of the genes considered constrained by our metric, leaving 827 genes. Their score yielded a similar number of constrained genes (n = 842), which were defined as those genes with a residual variation intolerance score in the top 5%. 231 genes were considered constrained by both metrics, which is far greater than expected (0.25%, ~41 genes). Using a Wilcoxon rank-sum test, we found that the genes defined as constrained by our metric had significantly lower (more intolerant) RVIS values ($p < 10^{-16}$). Similarly, the

genes with the top 5% RVIS had significantly higher constraint scores (Wilcoxon rank-sum, $p < 10^{-16}$). We found a correlation of -0.35 between the two scores of constraint, which is illustrated in **Figure 3.4**.

Confirming the association between constraint and *de novo* variants

The power to determine if a gene is significantly constrained relies on gene size. As mentioned above, genes where we predicted fewer than 5 rare synonymous variants had to be removed. In order to confirm that the association we found between constraint and the *de novo* LoF variants identified in ASD patients, we first investigated the relationship between constraint and the *de novo* variants found in unaffected individuals. As depicted in **Figure 3.2a**, we found no enrichment of *de novo* LoF variants from unaffected individuals in constrained genes. Additionally, we included gene length as a covariate while performing regressions of ASD *de novo* LoF genes on constraint and found that the association remained. We also took the largest 10% of genes and performed the regression again; constraint was still significant, but the gene length – when included as a covariate – showed no association.

Our method of determining constraint generates the number of rare missense variants that are expected to be in each gene. As an alternative metric to constraint, we also evaluated the fraction of missense variation that was not seen, a metric that is completely independent of gene size. We found that, in a linear regression, the fraction of missing missense variation was significantly able to predict whether a gene was haploinsufficient ($p = 2.13 \times 10^{-12}$).

For our final analyses to confirm that our enrichment analysis was not biased towards bigger genes, we created a list of the largest 5% of genes and queried the *de novo* loss-of-function variants identified in unaffected individuals. We expect that there should be no significant relationship between *de novo* LoF variants in unaffected individuals and these large genes. When we use a simple logistic regression to explain the *de novo* LoF genes in unaffected individuals, we find an odds ratio (OR) of about 5.5, which describes a highly significant enrichment of big genes. Our method of determining enrichment, however, accounts for the expected mutation rate of each gene – thereby inherently incorporating gene size – and shows this set of mutations is not actually "enriched" at all ($p = 0.425$; fold enrichment/OR = 1.1). These *de novo* LoF mutations in unaffected individuals are occurring in exactly the chance proportion they should be in larger genes. We therefore conclude that the enrichment analysis central to our interpretation of ASD events is not affected by gene lists being non-random with respect to size.

Comparison of three different metrics of constraint

Our metric is one way of searching for genes that appear to be relatively intolerant of mutations in the human population. One approach is the residual variation intolerance score (RVIS) created by Petrovski and colleagues[8], which evaluates the relative excess of rare variants to common ones in genes. Since Petrovski *et al* did not define a list of intolerant genes in their paper, we defined such a list by taking the top 5.5% most intolerant genes according to their metric. 5.5% was selected since that is the percentage of genes that we define as constrained using our metric. An additional

alternative comes from Bustamante *et al*, who used both fixed and polymorphic synonymous and nonsynonymous sites to find genes that appear to be affected by selection, including 813 loci that appeared to be under negative selection[16].

We sought to compare both our constraint score and list of constrained genes with the results of these other approaches. To do this, we focused on the ability to predict known haploinsufficient genes (as defined in OMIM) and the enrichment of these genes with *de novo* LoF mutations identified in ASD patients. Our results are summarized in **Table 3.8**. For the quantitative metrics (our constraint score and the RVIS metric), we performed a linear regression between haploinsufficient genes and the score with gene size as a covariate. While both metrics have significant predictive ability, our constraint score outperforms RVIS slightly (t-value = 10.011 for constraint, -9.561 for RVIS). For the list-based comparison, we used a logistic regression with gene length as a covariate. In this comparison, the top 5.5% intolerant genes according to RVIS had an odds ratio (OR) of ~5.5, while the constrained gene set that we defined had an OR of 4.9, both of which were significant. The genes identified by Bustamante and colleagues showed no significance (**Table 3.8a**).

We also evaluated the fraction of these different sets of constrained genes that contained a *de novo* LoF in ASD cases. Our method, as explained above, determines the fraction of constrained genes that are expected to contain a *de novo* mutation by chance. We then evaluate the observed fraction and can determine both the fold enrichment and significance. When we evaluated the three previously mentioned lists of genes – our constrained, top 5.5% intolerant genes using RVIS[8], and the loci identified by Bustamante[16] – we found that our list of constrained genes had the greatest fold

enrichment of genes that contained a *de novo* LoF in ASD cases ($p = 3.58x10^{-6}$; **Table 3.11b**). The top 5.5% of genes identified using RVIS also performed well (fold enrichment of 1.9, $p = 5.36x10^{-5}$), but the loci from Bustamante *et al* showed no significant enrichment.

Author contributions

Kaitlin Samocha: conceived and designed mutational model and constraint methods, performed all analyses not listed below, writing

Elise Robinson: analyses of autism samples split by verbal and nonverbal IQ groups (Table 3.5c), phenotype analysis of autism cases (Tables 3.6 and 3.7), writing

Jack Kosmicki: obtained IQ information for autism samples and helped Elise with the IQ and phenotype analyses

Stephan Sanders: provided updated *de novo* calls from trios sequenced at Yale and manuscript comments

Andrew Kirby: additional indel calling and manuscript comments

Swapan Mallick: provided table with divergent sites between humans and macaques

Lauren McGrath: gave suggestions for IQ analysis of autism samples

Christine Stevens, Stacey Gabriel, Mark DePristo: data processing and sample tracking

Aniko Sabo, Karola Rehnström, Dennis Wall, Daniel MacArthur, Shaun Purcell, Aarno Palotie, Eric Boerwinkle, Joseph Buxbaum, Edwin Cook Jr, Richard Gibbs,

Gerard Schellenberg, James Sutcliffe, Bernie Devlin, Kathryn Roeder: provided sequencing data and manuscript comments

Benjamin Neale: conceived and designed mutational model and constraint methods, created the mutation rate table and depth of coverage file, writing

Mark Daly: conceived and designed mutational model and constraint methods, overall guidance, writing

Bibliography

1. Ng, S.B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature genetics* **42**, 790-3 (2010).

2. Iossifov, I. *et al.* De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* **74**, 285-299 (2012).

3. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-245 (2012).

4. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250 (2012).

5. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241 (2012).

6. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature genetics* **43**, 585-9 (2011).

7. Antonarakis, S.E. CpG Dinucleotides and Human Disorders. in *eLS* (John Wiley & Sons, Ltd, 2006).

8. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet* **9**, e1003709 (2013).

9. de Ligt, J. *et al.* Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *New England Journal of Medicine* **367**, 1921-1929 (2012).

10. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *The Lancet* **380**, 1674-1682 (2012).

11. O'Roak, B.J. *et al.* Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. *Science* (2012).

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

13. Darnell, J.C. *et al.* FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism. *Cell* **146**, 247-261 (2011).

14. Sanders, S.J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-85 (2011).

15. Xu, B. *et al.* De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature genetics* **44**, 1365-1369 (2012).

16. Bustamante, C.D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153-1157 (2005).

17. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652-4 (1991).

18. Epi, K.C. & Epilepsy Phenome/Genome, P. De novo mutations in epileptic encephalopathies. *Nature* **501**, 217-221 (2013).

19. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-8 (2011).

20. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).

21. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-303 (2010).

22. Krawczak, M., Ball, E.V. & Cooper, D.N. Neighboring-Nucleotide Effects on the Rates of Germ-Line Single-Base-Pair Substitution in Human Genes. *The American Journal of Human Genetics* **63**, 474-488 (1998).

23. Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R. Most Rare Missense Alleles Are Deleterious in Humans: Implications for Complex Disease and Association Studies. *The American Journal of Human Genetics* **80**, 727-739 (2007).

24. Hellmann, I. *et al.* Why do human diversity levels vary at a megabase scale? *Genome Research* **15**, 1222-1231 (2005).

25. Stamatoyannopoulos, J.A. *et al.* Human mutation rate associated with DNA replication timing. *Nature genetics* **41**, 393-395 (2009).

26. Koren, A. *et al.* Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation. *The American Journal of Human Genetics* **91**, 1033-1040 (2012).

27. Weschler, D. *Weschler Intelligence Scale for Children--4th Edition (WISC-IV).* (Harcourt Assessment, San Antonio, Texas, 2003).

28. Weschler, D. *Weschler Abbreviated Scale of Intelligence (WASI)*, (Harcourt Assessment, San Antonio, Texas, 1997).

29. Weschler, D. *Weschler Primary and Preschool Scale of Intelligence--Third Edition*, (Harcourt Assessment, San Antonio, Texas, 2002).

30. Elliott, C.D. *DAS Administration and Scoring Manual*, (The Psychological Corporation, San Antonio, Texas, 1990).

31. Elliott, C.D. *DAS Introductory and Technical Handbook*, (The Psychological Corporation, San Antonio, Texas, 1990).

32. Mullen, E.M. *Mullen Scales of Early Learning*, (American Guidance Service Inc., Circle Pines, MN, 1995).

33. Roid, G.H. & Miller, L.J. *Leiter International Performance Scale- Revised*, (Stoelting Co., Wood Dale, Illinois, 1997).

Chapter 4

Leveraging large reference populations to identify functionally constrained genes

Work presented in this chapter will be published as part of:

Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Under review*.

Abstract

Large-scale exome sequencing efforts of reference populations have greatly improved both clinical and functional interpretation of genetic variation. We analyzed the variation identified in 60,706 individuals included in the Exome Aggregation Consortium (ExAC) dataset to identify genes under strong selective constraint. Of particular interest is the set of 3,230 genes that are significantly depleted of loss-of-function variation. These constrained genes are enriched for established haploinsufficient and dominant disease genes, and represent core biological processes (e.g. spliceosome and proteasome). However, only 28% have been associated with a human disease phenotype; those that have not yet been associated promise to be a fruitful set to further investigate both within the clinic and in functional studies.

Introduction

One of the major challenges within the field of human genetics is determining which variant, or set of variants, is associated to disease. High-throughput DNA sequencing technologies have aided this effort by allowing researchers to investigate nearly all single nucleotide and small insertion and deletion (indel) variants within an individual's genome or exome (the 1% of the genome that codes for proteins). Unfortunately, each individual harbors tens of thousands of variants and examining each of these variants would be a long and laborious task.

To make the task of associating variation to disease, it is critical to be able to prioritize variants and define a subset for further analysis. There are many variant-level prioritization tools[1-3], but we have found that using gene's intolerance of

nonsynonymous variation can also aid in variant interpretation[4,5]. Identifying these constrained genes depends on the availability of exome sequencing datasets of large reference populations. While both the 1000 Genomes Project[6] and the National Heart, Lung and Blood Institute's Exome Sequence Project[7] publically released the protein-coding variation from thousands of individuals (n = 2,504 and 6,503, respectively), the size of these datasets restricted researcher's ability to identify genes that are intolerant of loss-of-function variation.

Here, we describe using the Exome Aggregation Consortium (ExAC), which is an order of magnitude larger than previously released datasets (n = 60,706), to evaluate missense and loss-of-function constraint.

Results

The deep ascertainment of rare variation in the Exome Aggregation Consortium (ExAC) allows us to infer the extent of selection against variant categories on a per-gene basis by examining the proportion of variation that is missing compared to expectations under random mutation. Conceptually similar approaches have been applied to smaller exome datasets[4,5] but have been underpowered, particularly when analyzing the depletion of loss-of-function (LoF) variants. We compared the observed number of rare (minor allele frequency [MAF] < 0.1%) variants per gene to an expected number derived from a selection neutral, sequence-context based mutational model[5] (Chapter 3). The model performs well in predicting the number of synonymous variants, which should be under minimal selection, per gene (r = 0.98; **Figure 4.1**).

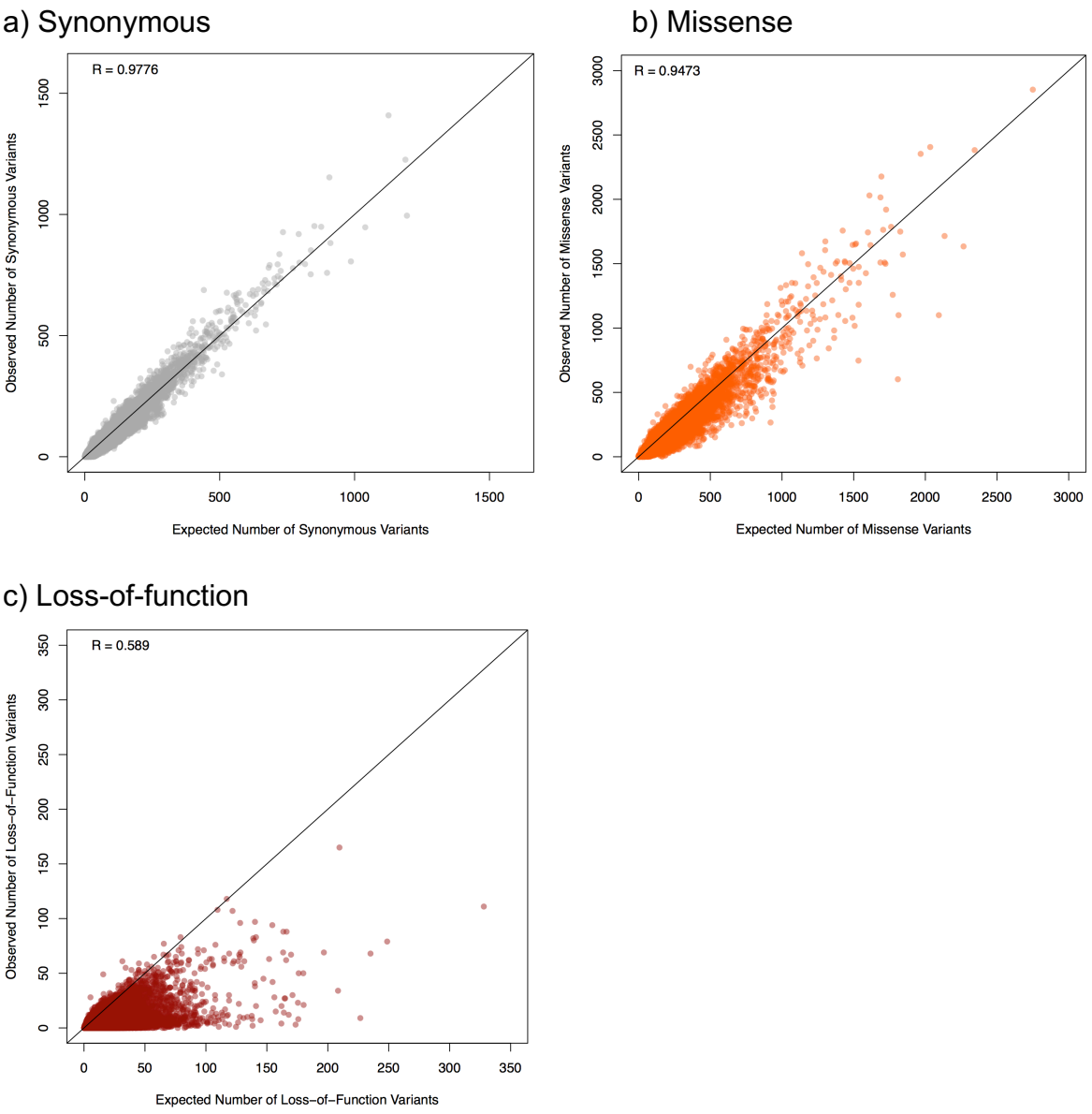


Figure 4.1. The correlation between observed and expected variant counts for synonymous (a), missense (b), and loss-of-function (c) variants. The line shows a perfect correlation (slope = 1). Axes have been trimmed to remove *TTN*.

We quantified deviation from expectation with a Z score[5], which for synonymous variants is centered at zero, but is significantly shifted towards higher values (greater constraint) for both missense and LoF (Wilcoxon $p < 10^{-50}$ for both; **Figure 4.2**). The genes on the X chromosome are significantly more constrained than those on the

autosomes for missense ($p < 10^{-7}$) and loss-of-function ($p < 10^{-50}$). The high correlation between the observed and expected number of synonymous variants on the X chromosome (r = 0.97 versus 0.98 for autosomes) indicates that this difference in constraint is not due to a calibration issue. To reduce confounding by coding sequence length for LoFs, we developed an expectation-maximization algorithm (see Materials and Methods) using the observed and expected LoF counts within each gene to separate genes into three categories: null (tolerant of homozygous LoFs), recessive (tolerant only of heterozygous LoFs), and haploinsufficient (intolerant of homozygous LoFs). This metric – the probability of being loss-of-function intolerant (pLI) – separates genes of sufficient length into LoF intolerant (pLI ≥ 0.9, n = 3,230) or LoF tolerant (pLI ≤ 0.1, n = 10,374) categories. pLI is less correlated with coding sequence length (r = 0.17 as compared to 0.57 for the LoF Z score), outperforms the LoF Z score as an intolerance metric (discussed more in Materials and Methods), and reveals the expected contrast between gene lists (**Figure 4.3**).

Additionally, pLI is positively correlated with a gene product's number of physical interaction partners ($p < 10^{-41}$). The most constrained pathways (highest median pLI for the genes in the pathway) are core biological processes (spliceosome, ribosome, and proteasome components; Kolmogorov-Smirnov [KS] test $p < 10^{-6}$ for all) while olfactory receptors are among the least constrained pathways (KS test $p < 10^{-16}$), demonstrated in **Figure 4.3** and consistent with previous work[8-12].

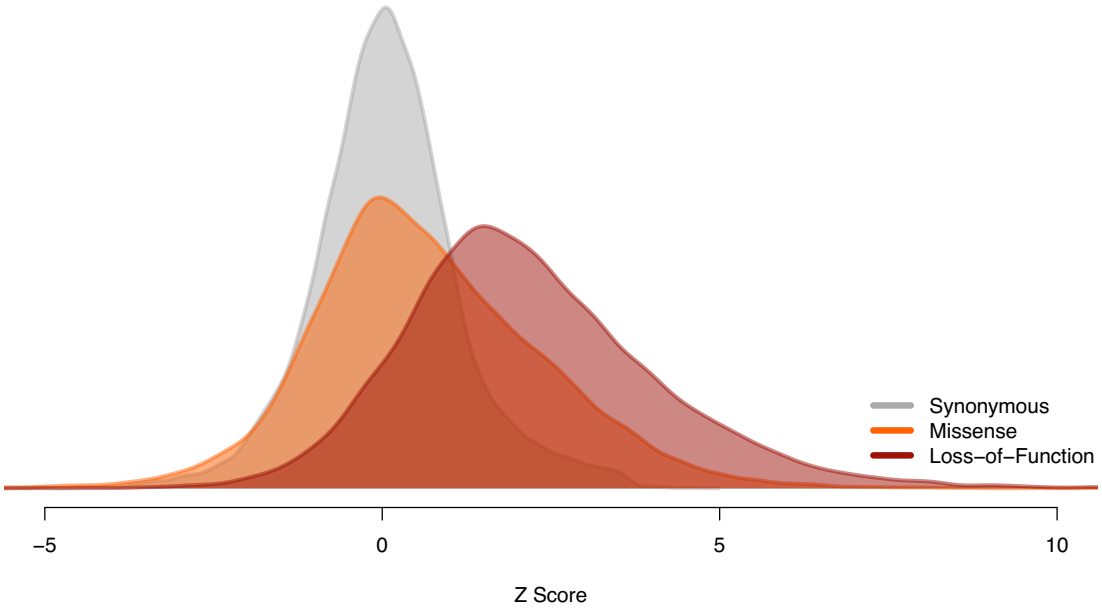


Figure 4.2. The distribution of Z scores for synonymous (gray), missense (orange), and loss-of-function (red) for 18,225 genes. This measure of departure of number of variants from expectation is normally distributed for synonymous variants, but right-shifted (higher constraint) for missense and loss-of-function variants, indicating that more genes are intolerant to these classes of variation.

Critically, we note that LoF-intolerant genes include virtually all known severe haploinsufficient human disease genes (**Figure 4.3**), but that 72% of these genes have not yet been assigned a human disease phenotype despite clear evidence for extreme selective constraint. Many of these genes (79%) specifically do not have a disease-associated variant in ClinVar[13] (a database that collects evidence for pathogenicity of variants). We note that this extreme constraint does not necessarily reflect a lethal disease, but is likely to point to genes where heterozygous loss-of-function confers some non-trivial survival or reproductive disadvantage.

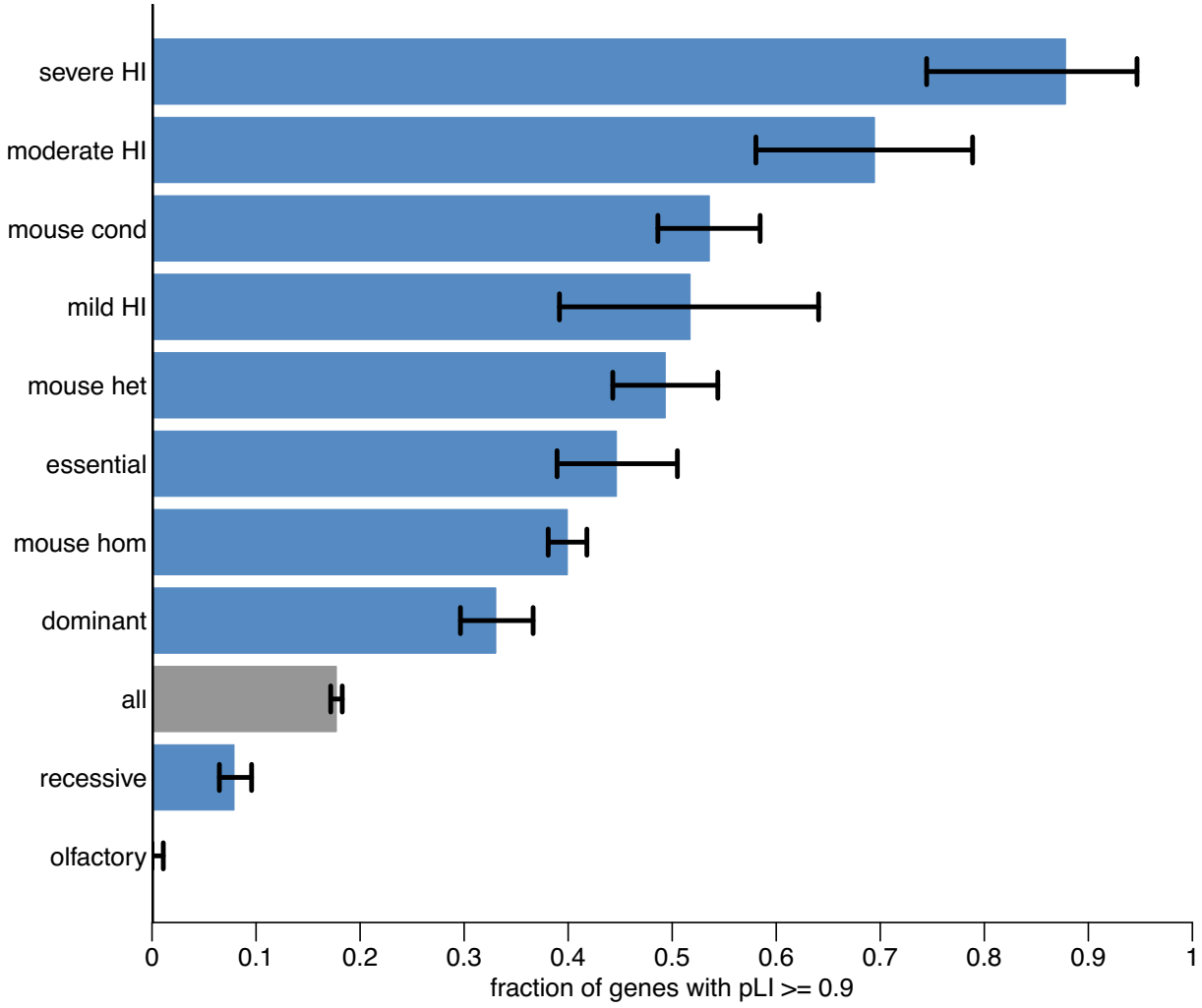
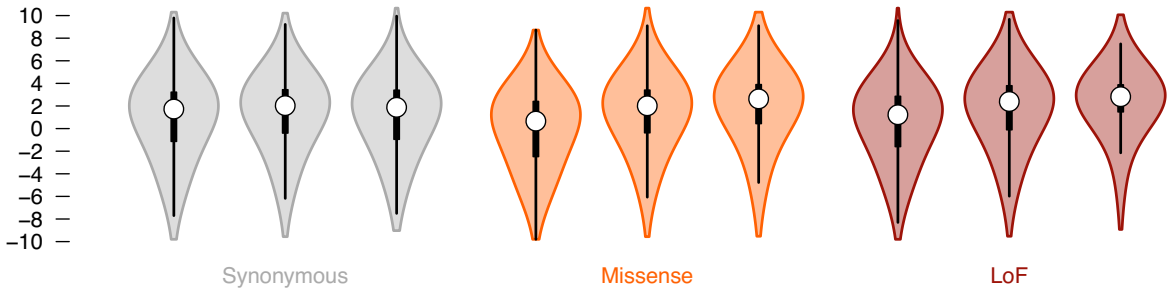


Figure 4.3. The proportion of genes in gene sets that are very likely intolerant of loss-of-function variation. pLI close to one indicates extreme intolerance to loss-of-function variation; we therefore take pLI ≥ 0.9 as the cut-off for extreme loss-of-function intolerance. The black error bars indicate a 95% confidence interval. olfactory = olfactory receptor genes (n = 371); recessive = recessive disease genes from Blekhman and Berg (n = 1,183); all (n = 18,225); dominant = dominant disease genes from Blekhman and Berg (n = 709); mouse hom = genes that are lethal in mice when both copies are knocked out (n = 2,760); essential = genes that are essential in cell culture as curated by Hart et al 2014 (n = 285); mouse het = genes that are lethal in mice when one copy is knocked out (n = 387); mild HI = haploinsufficient genes that cause a mild disease (n = 59); mouse cond = genes that are lethal in mice when conditionally knocked out in adult mice (n = 402); moderate HI = haploinsufficient genes that cause moderately severe disease (n = 77); severe HI = haploinsufficient genes that cause severe disease (n = 44).

The most highly constrained missense (top 25% missense Z scores) and LoF (pLI ≥0.9) genes show higher expression levels and broader tissue expression than the least constrained genes[14] **(Figure 4.4**). These most highly constrained genes are also depleted for eQTLs ($p < 10^{-9}$ for missense and LoF; **Figure 4.5a**), yet are enriched within genome-wide significant trait-associated loci (χ^2 $p < 10^{-14}$, **Figure 4.5b**). Intuitively, genes intolerant of LoF variation are dosage sensitive: natural selection does not tolerate a 50% deficit in expression due to the loss of single allele. Unsurprisingly, these genes are also depleted of common genetic variants that have a large enough effect on expression to be detected as eQTLs with current limited sample sizes. However, smaller changes in the expression of these genes, through weaker eQTLs or functional variants, are more likely to contribute to medically relevant phenotypes.

a) Median gene expression across all tissues for bins of constraint



b) Number of tissues where the gene is expressed for bins of constraint

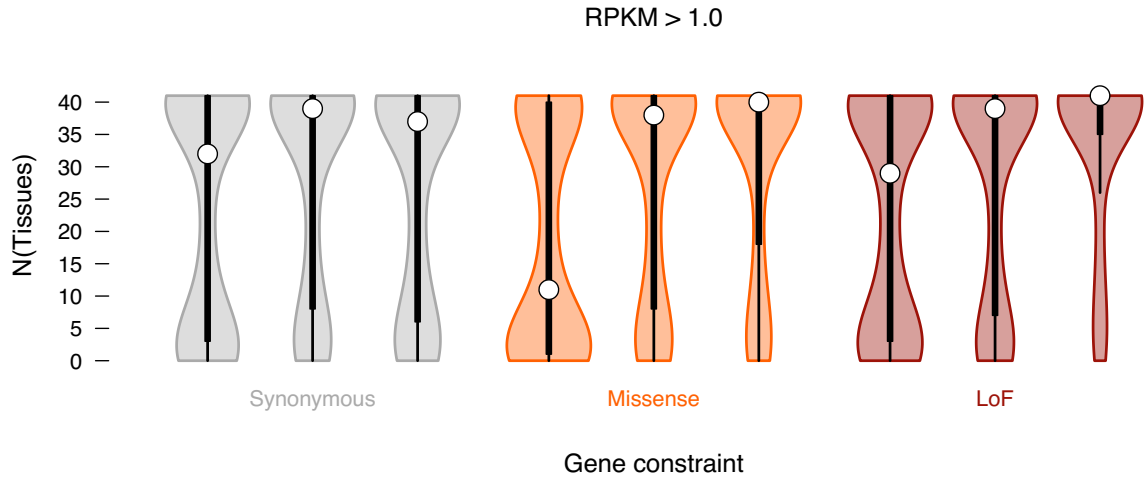


Figure 4.4. Expression patterns of genes for bins of constraint. For synonymous and missense Z, the bins are: bottom quartile (< 25%), two middle quartiles grouped together, and top quartile (> 75%). For pLI: pLI ≤ 0.1, 0.1 < pLI < 0.9, and pLI ≥ 0.9. Note pLI is the metric used for loss-of-function (LoF) intolerance. (a) The median gene expression, in $\log_2$(RPKM), across all tissues for bins of constraint. (b) The relationship between constraint and the number of tissues in which a gene is expressed at an RPKM > 0.1. Synonymous Z scores show no correlation with the number of tissues in which a gene is expressed, but the least missense- and LoF-constrained genes tend to be expressed in fewer tissues. Thick black bars indicate the first to third quartiles, with the white circle marking the median.

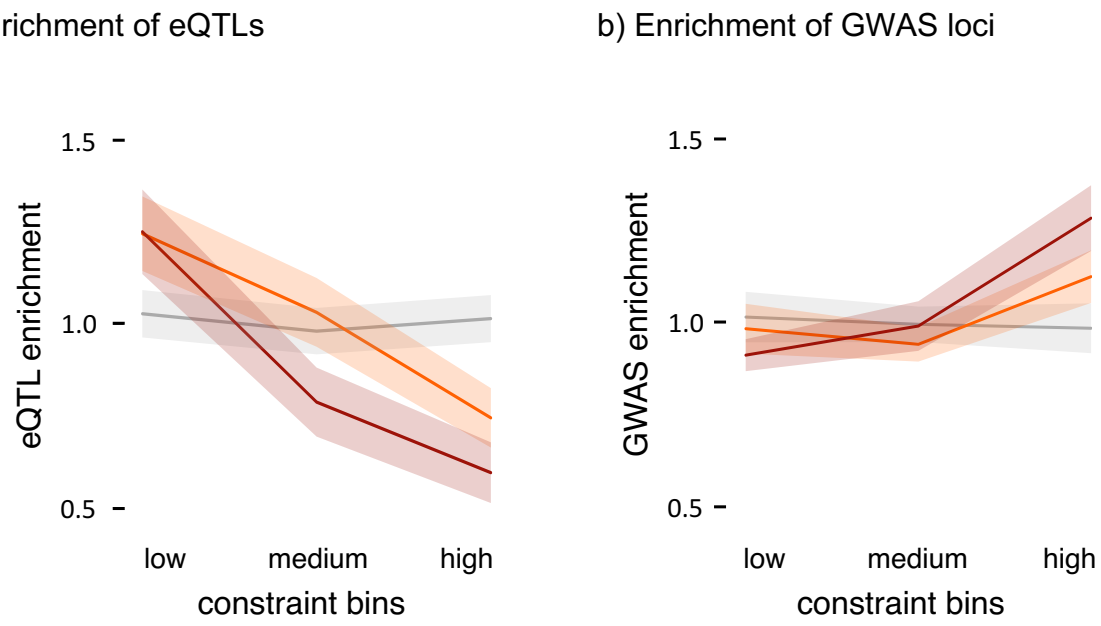


Figure 4.5. Signals of eQTL and GWAS loci enrichment for constraint bins. For synonymous and missense Z, the bins are: bottom quartile (< 25%), two middle quartiles grouped together, and top quartile (> 75%). For pLI: pLI ≤ 0.1, 0.1 < pLI < 0.9, and pLI ≥ 0.9. (a) The proportion of eGenes (a gene with a significant eQTL at a false discovery rate [FDR] of 5%) found in whole blood samples from GTEx[14] for each constraint bin. Highly missense- and LoF-constrained genes are less likely to have eQTLs as the average gene. No relationship between synonymous genes and eQTLs is observed. (b) Enriched of GWAS loci downloaded from the Catalog[15] for each constraint bin. Highly missense- and LoF-constrained genes are more likely to be adjacent to GWAS signals than the average gene, but no relationship is seen for synonymous Z bins. Shaded regions around the lines indicate 95% confidence intervals.

Discussion

The large sample size of the ExAC dataset provided the opportunity to analyze the sensitivity of human genes to nonsynonymous variation. While previous sample sizes have been adequately powered for the assessment of gene-level intolerance to missense variation[4,5], ExAC provides for the first time sufficient power to investigate genic intolerance to loss-of-function (LoF) variants.

We created pLI—the probability of being loss-of-function intolerant—to identify highly LoF constrained genes and highlighted 3,230 that were significantly depleted of

LoF variation. Comparing pLI to the LoF Z score revealed that pLI was better able to predict haploinsufficient genes and had a greater enrichment of *de novo* LoFs identified in 3,982 cases with an autism spectrum disorder[16,17]. We also compared pLI to a previous metric developed to predict haploinsufficient genes called p(HI)[18]. Our metric was able to identify twice as many genes at same cut off as p(HI)—indicating increased sensitivity of our metric—but a larger proportion of the genes in the high p(HI) tail are considered likely haploinsufficient by both metrics. The subset of genes that are considered likely haploinsufficient (≥ 0.8) by both metrics shows the greatest enrichment of ClinGen haploinsufficient genes when compared to genes uniquely flagged by each metric. Therefore, there would be benefit in combining the two metrics in a future measure of haploinsufficiency.

The 3,230 severely LoF constrained genes represent core biological processes and include many dominant and haploinsufficient disease genes. The established disease genes, however, do not explain the majority of the highly LoF-intolerant genes; only 28% of genes with a pLI ≥ 0.9 have a human disease phenotype listed in OMIM or ClinVar[13]. Further investigation will likely reveal genes that, when disrupted, cause embryonic lethality as well as additional disease genes that have yet to be tied to specific phenotypes. These results suggest that this set of genes will be able to aid in the interpretation of genetic variation identified in patients.

Materials and Methods

Establishing the expected number of variants per gene

Probabilities of a mutation

Our metrics to evaluate a gene's intolerance to variation—their level of constraint—rely on comparing the observed variant counts to an expectation. In order to determine the expected number of variants per gene, we modified a previous method described in detail in Chapter 3. We used the mutation rate table created for Samocha et al[5] to determine the probability of mutation, split by mutational class (synonymous, missense, nonsense, and splice site), for each exon in the canonical transcript. As before, we adjusted the probabilities of mutation for regional divergence between humans and macaques. Two major changes were made between the previous version of the method and the one used in this paper: (1) we now used GENCODE v19 annotations for transcripts instead of Refseq and (2) the expected number of variants, and not the probability of mutation, is adjusted for depth of sequencing coverage (see below). Here, we focused on the canonical transcript as defined by Ensembl v75 for each protein-coding gene and drop all transcripts that do not begin with a methionine, end with a stop codon, or whose length are not divisible by three. After all of these filters, there were 19,620 canonical transcripts that are used in all following analyses.

Determining the depth of coverage correction

We used the Exome Aggregation Consortium (ExAC; n = 60,706) dataset and extracted the number of rare (minor allele frequency < 0.1%) single nucleotide variants for every exon of the canonical transcripts. These variants were assigned functional

classes (synonymous, missense, nonsense, and splice site) based on the amino acid change or position in the splice site. We then needed a way to account for the depth of sequencing coverage since regions that are poorly sequenced will, by definition, have fewer variants than expected. To do this, we determined the median depth of coverage for each exon. Given that synonymous variants are most likely to be free of extreme negative selection, we focused on those variants. Using only those exons with a median depth ≥ 50, which we consider to be well sequenced, we regressed the number of rare synonymous variants on the probability of a synonymous mutation to determine the appropriate formula to predict the number of expected synonymous variants. This formula was applied to all exons (regardless of depth). To find the appropriate way to correct for sequencing coverage, we grouped exons by depth (bins of 2) and determined the sums of all observed and expected synonymous variants in these exons. The sum of observed synonymous variants divided by the sum of expected variants had a logarithmic relationship between depth bins of 0 and 50, where it then plateaued at ~1 (**Figure 4.6**). We fit the curve to determine the appropriate depth of coverage correction for exons with a median depth between 1 and 50.

$$depth\ adjusted\ count = \begin{cases} expected\ count, median\ depth\ \geq 50 \\ expected\ count * (0.089 + 0.217 * \ln(median\ depth), 1 \leq\ median\ depth < 50 \\ 0.089 * expected\ count, median\ depth < 1 \end{cases}$$

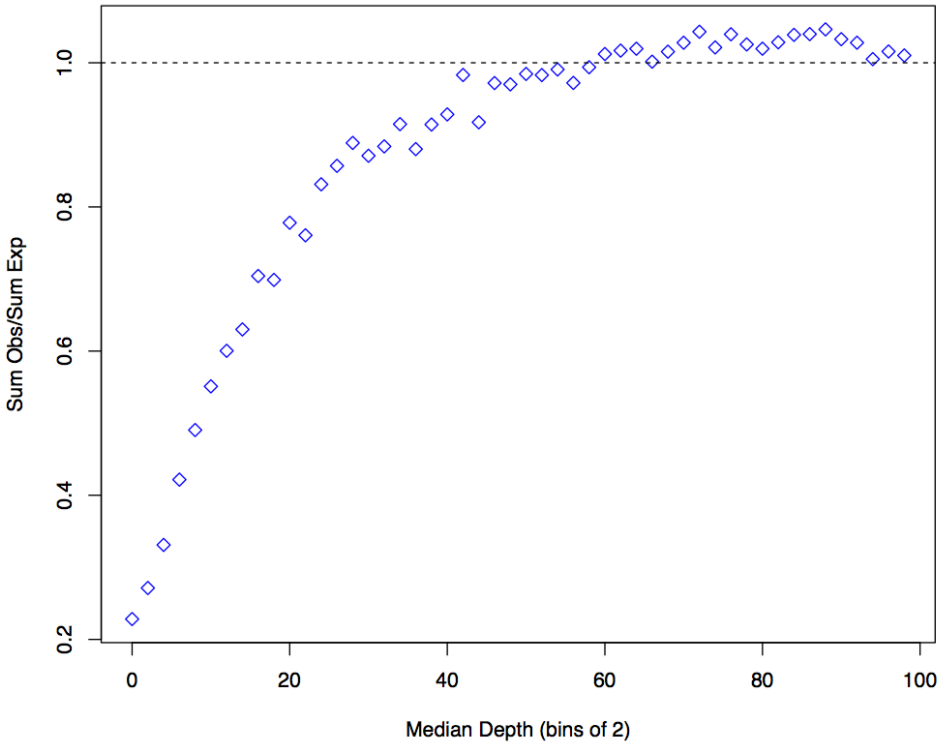


Figure 4.6. The relationship between the median depth of exons and the sum of all observed synonymous variants in those exons divided by the sum of all expected synonymous variants.

Expected number of variants

To determine the depth-corrected expected number of variants per exon, we used those exons with a median depth ≥ 50 and regressed the number of rare synonymous variants on the probability of a synonymous mutation. These regressions were done separately for the autosomes with the pseudo-autosomal regions (PAR) of the X chromosome, the non-PAR regions of the X chromosome, and the Y chromosome. The resulting formulas were used to predict the depth-uncorrected expected number of synonymous, missense, and loss-of-function variants (LoFs; nonsense and splice site) variants for all exons. The correlation between the observed and depth-uncorrected expected number of synonymous variants per exon was 0.8360.

We then corrected these expected numbers by the above equation and observed an increased correlation between observed and depth-corrected expected synonymous variants (r = 0.9283). Note that from this point forward, the expected number of variants always refers to the depth-corrected counts.

Creation of the constraint metric

Determining Z scores of the deviation of observation from expectation

We created a signed Z score to establish the significance of the deviation of observed variant counts per gene from expectation as in Chapter 3 with minor modifications. To start, we sum all exon level variant counts across canonical transcripts. Here, the observed count is the number of unique variants with a VQSLOD ≥ -2.632 and 123 or fewer alternative alleles (minor allele frequency cut off of ~0.1%). If an exon had a median depth < 1, the variant counts for that exon were not included in the total for the transcript. We then removed all transcripts where no variants were observed. For the remaining 18,466 transcripts, we calculated the chi-squared value for the deviation of observation from expectation for each mutational class: synonymous, missense, and loss-of-function (LoF). The square root of these values is multiplied by -1 if the number of observed variants is greater than expectation (or 1 if observed counts are smaller than expected) to create the Z score.

A critical next step is to correct the scores so that the synonymous Z scores followed an approximately normal distribution. For the synonymous Z scores, we used a subset of transcripts whose synonymous Z scores fell in between -5 and 5. All synonymous Z scores were divided by the standard deviation of this outlier-removed

subset to create the corrected Z scores. A slightly different approach was used for missense and LoF Z scores. We took all transcripts with a missense Z score between -5 and 0 and combined them with those same Z scores multiplied by -1 (to create a mirrored distribution). All missense Z scores were divided by the standard deviation of the mirrored distribution to create the corrected missense Z scores. The same procedure was applied to the LoF Z scores.

Removing outliers

We then used these corrected Z scores to define outlier transcripts—specifically those with significantly elevated synonymous and missense counts or significantly depleted synonymous and missense counts. These outliers were defined as transcripts with a synonymous Z < -3.71 and a missense Z < -3.09 or transcripts with a synonymous Z > 3.71 and a missense Z > 3.09. These filters removed a total of 241 transcripts, leaving 18,225 for all further analyses. The distribution of the synonymous, missense, and LoF Z scores are depicted in **Figure 4.2**. Note that a Z score of ~3.09 is equivalent to a p-value of 10^{-3} and is considered the significance threshold when splitting transcripts into constrained and unconstrained classes.

Correlation of observed and expected counts

For the set of 18,225 cleaned transcripts, the correlation between the number of observed rare (minor allele frequency < 0.1%) synonymous variants and the expected number of variants given the above model is 0.9776. This correlation is higher than simply regressing the observed synonymous variants against number of coding bases

in the gene (r = 0.9201), or the probability of a synonymous mutation (r = 0.9349). This relationship between observed and expected mutation counts can be seen for synonymous, missense, and LoF variants in **Figure 4.1**.

Power of the Z score analyses

To achieve a Z score of 3.09 (a p-value equivalent of 10^{-3}), the number of expected variants would need to be a minimum of 10. Following this criterion, 99.5% of transcripts could be evaluated for missense constraint. However, only 11,437 transcripts (62.8%) were mutable enough to have 10 or more expected LoFs in the ExAC dataset (see below).

Z score distributions for gene lists

We next investigated the synonymous, missense, and LoF Z score distributions for the following gene lists: autosomal recessive[19,20], autosomal dominant[19,20], essential in cell culture[21], ClinGen haploinsufficient, FMRP interactors[22], and olfactory receptors[23]. For the synonymous Z scores (**Figure 4.7**), most gene lists match the distribution of the full set of canonical transcripts (median Z = 0.05). The only notable exception is the list of olfactory receptors, which show 118% of the expected synonymous variation (Wilcoxon $p < 10^{-46}$).

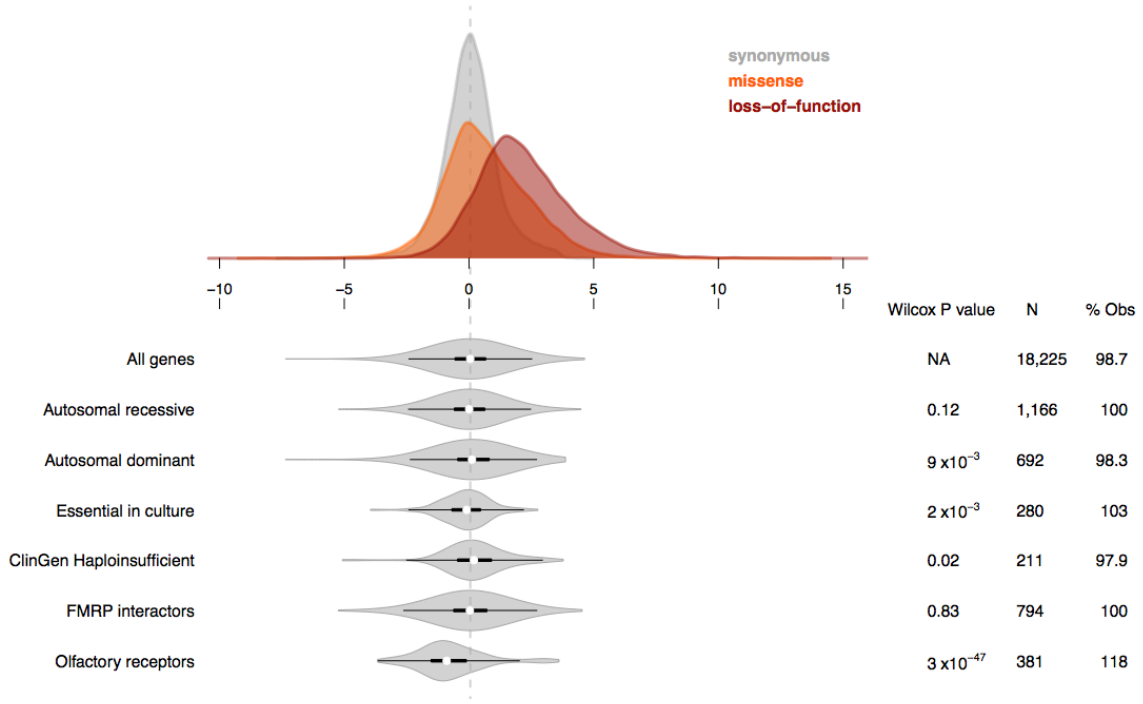


Figure 4.7. Distribution of synonymous Z scores for gene sets. Wilcoxon p-value for difference from the full distribution, the number of genes in the set, and the percentage of expected variation observed are reported on the right.

Across all canonical transcripts, ~89% of all missense variation is observed and the median missense Z score is 0.51. As a note, higher (more positive) Z scores indicate increased selective constraint, while negative Z scores are given for transcripts where more variation was seen than expected. All of the gene sets tested significantly differ from the overall distribution (**Figure 4.8**) with the recessive genes and olfactory receptors showing slightly lower missense Z scores. The rest of the gene sets have significantly higher missense Z scores (Wilcoxon $p < 10^{-28}$).

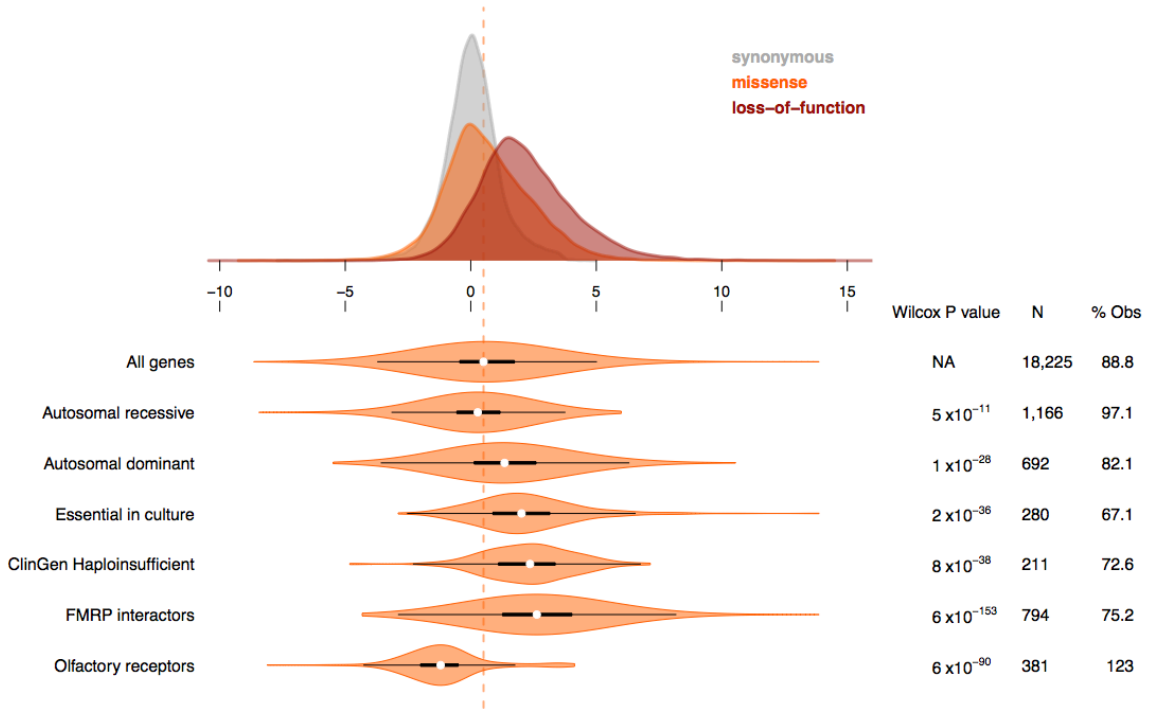


Figure 4.8. Distribution of missense Z scores for gene sets. Wilcoxon p-value for difference from the full distribution, the number of genes in the set, and the percentage of expected variation observed are reported on the right.

The LoF Z scores have the most skewed distributions (**Figure 4.9**). Overall, only 39% of the expected loss-of-function variation is observed, giving the full set of canonical transcripts a median LoF Z score of 1.97. The Z scores for the autosomal recessive genes match the overall distribution fairly closely (Wilcoxon p = 0.02, median = 2.09). The olfactory receptors, as before, have significantly lower LoF Z scores (Wilcoxon $p < 10^{-50}$, median = 0.16), but unlike with synonymous and missense do not have more loss-of-function variation than expected (95% observed).

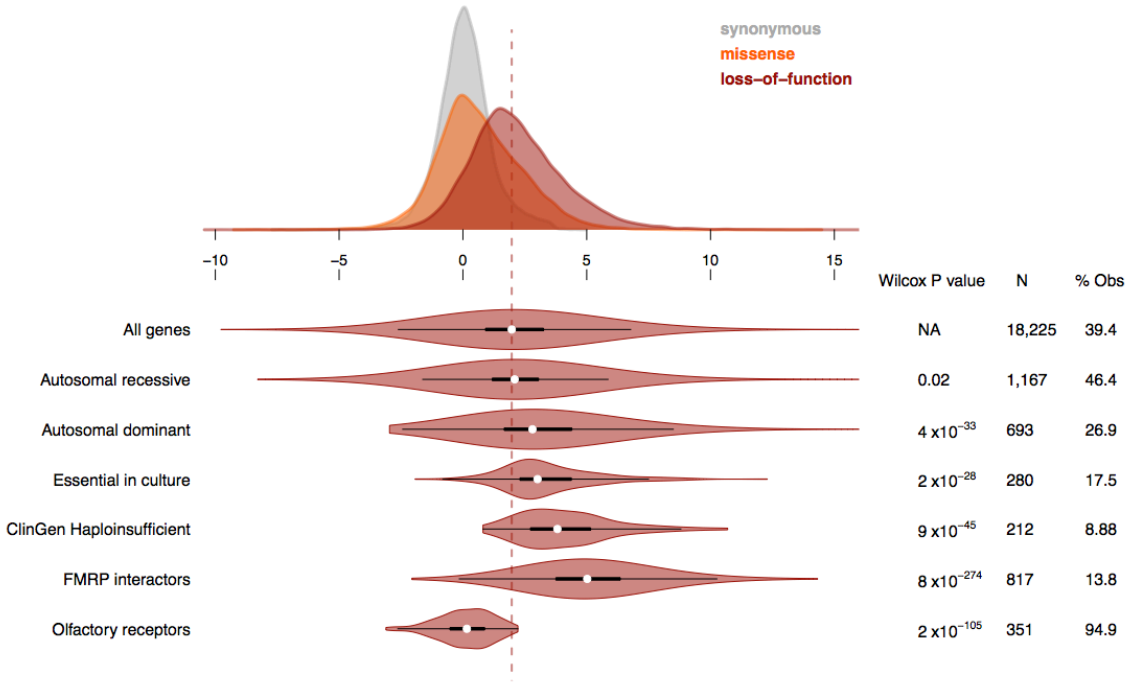


Figure 4.9. Distribution of loss-of-function Z scores for gene sets. Wilcoxon p-value for difference from the full distribution, the number of genes in the set, and the percentage of expected variation observed are reported on the right.

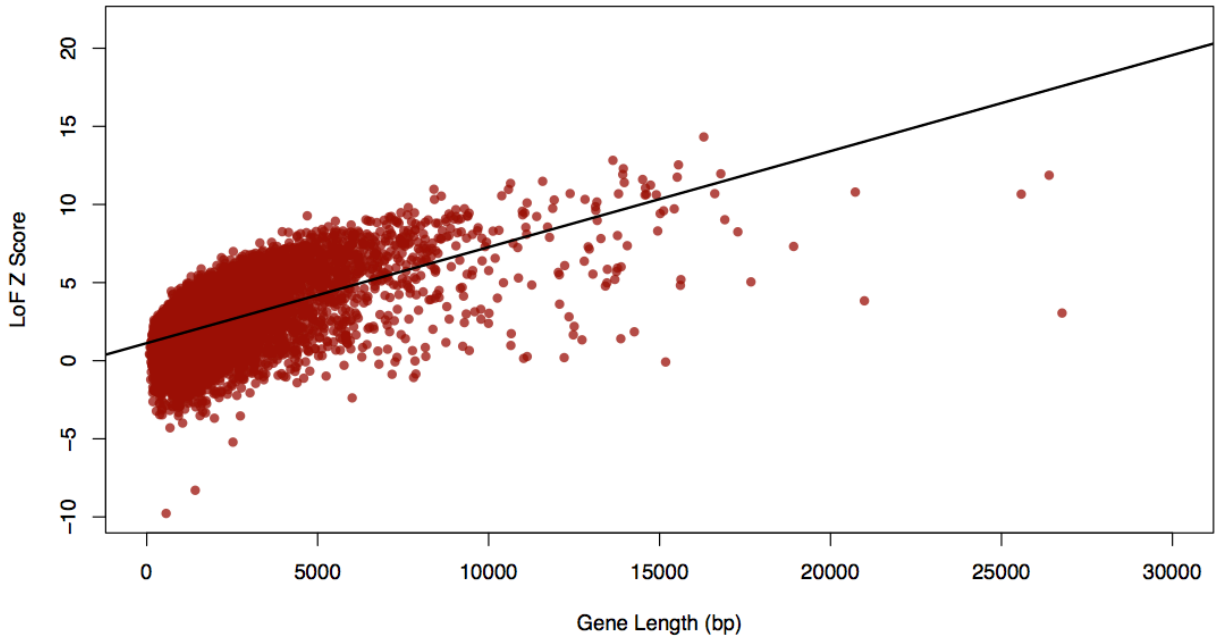
Creation of a new loss-of-function constraint score

The LoF Z score is correlated with gene length

The Z scores were created to evaluate the significance of the deviation of observed counts from expectation. Given this, it is sensitive to differences in power. For example, a gene with 0 observed variants would require ~10-11 expected variants to pass a significance threshold of 10^{-3} (Z score of 3.09). The expected number of variants per gene is based on the length and mutability of the transcript. Since the probability of having a loss-of-function mutation is small (roughly an order of magnitude less than the probability of a missense mutation), only 63% of the canonical transcripts are expected to have 10 or more LoFs in the ExAC dataset (59% if expecting 11 LoFs).

Due to this reliance on mutability, it is unsurprising that the LoF Z score is correlated with the coding length of the transcript (r = 0.5697; **Figure 4.10a**). This correlation is not seen for the missense Z score (r = 0.0566; **Figure 4.10b**). Therefore, larger transcripts will have more significant LoF deviations (and Z scores) than smaller transcripts and some transcripts that are truly intolerant of loss-of-function variation will be too small to achieve statistical significance. These results motivated the search for a better metric to capture LoF constraint (discussed below).

a) Loss-of-function



b) Missense

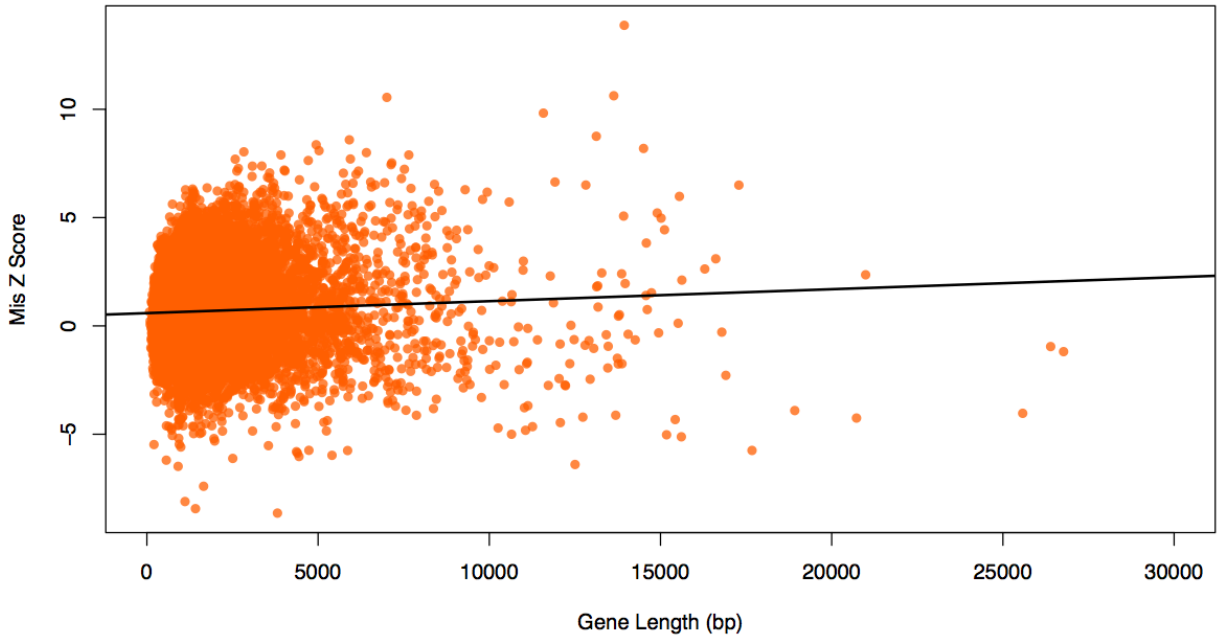


Figure 4.10. The correlation between the length of the gene and the Z score. (a) The correlation for the loss-of-function Z score. The Pearson's r between the two is 0.5697. (b) The correlation for the missense Z score. The Pearson's r between the two is 0.0566. The black line shows the linear relationship. Axes have been trimmed to remove *TTN*.

Evaluating the ratio of missing loss-of-function variation

A natural metric to evaluate intolerance to loss-of-function variation is the amount of expected variation that was not observed. Truly intolerant transcripts should be missing most of the expected variation, which is independent of the length of the transcript. We defined the ratio of missing variation as one minus the quotient of the observed counts divided by the expected counts.

The correlation between the length of the transcript and the ratio of missing loss-of-function variation is 0.1561 (**Figure 4.11**). The distributions of the ratio of missing synonymous, missense, and loss-of-function variation are depicted in **Figure 4.12a**. The majority of transcripts fall between 0 and 1 for the ratio of missing LoF variation, where 1 means the transcript is completely devoid of LoF variation. Both the synonymous and missense distributions shift towards transcripts having more of their expected variation.

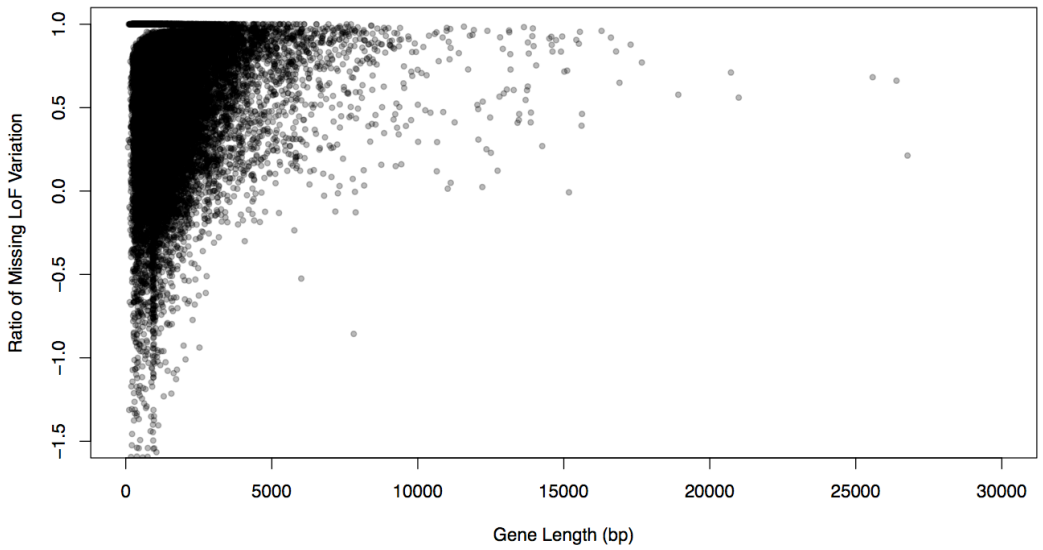
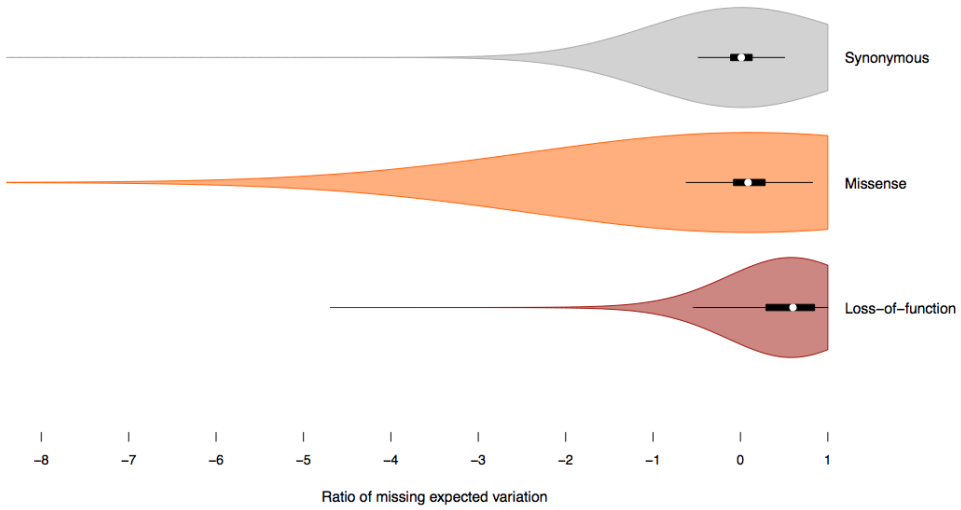


Figure 4.11. The relationship between gene length and the ratio of missing loss-of-function variation. The Pearson's r between the two is 0.1561. The x-axis was trimmed to remove *TTN* and the y-axis was cut at -1.5 (out of -4) to show pattern of the data.

a) Distribution of the ratio of missing expected variation for synonymous, missense, and loss-of-function



b) The ratio of missing loss-of-function variation for gene lists

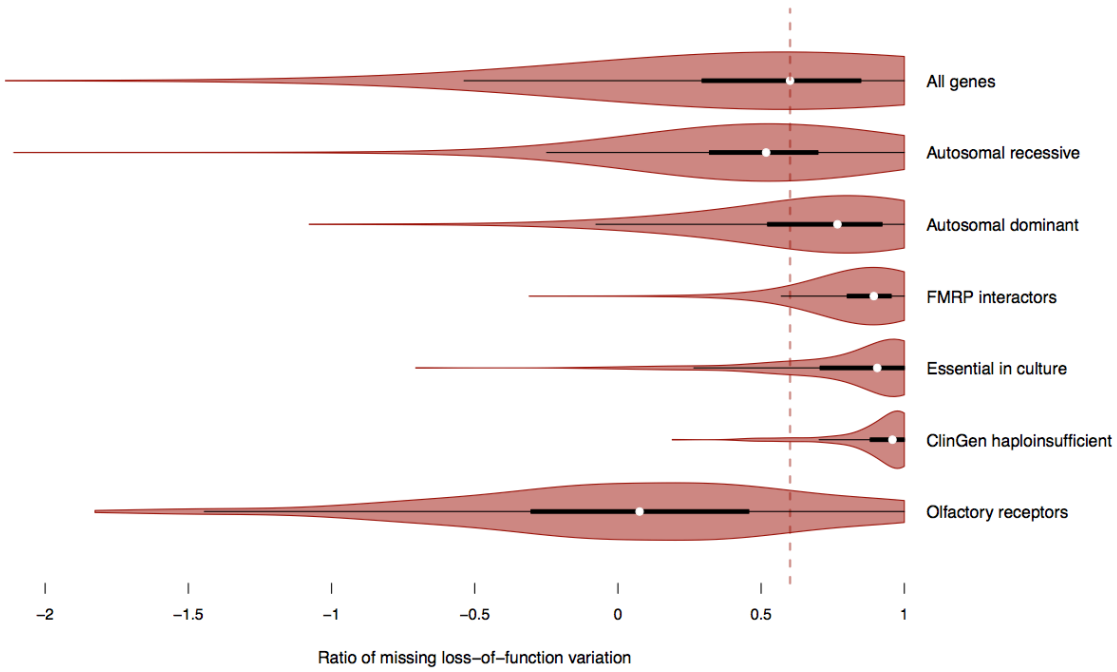


Figure 4.12. Distributions of the ratio of missing variation. Note that 1 means there were no variants observed and negative values indicate more variation observed than expected. (a) The distribution of the ratio of missing expected variation for synonymous, missense, and loss-of-function. The x-axis has been trimmed at -8 (out of -18) to

Figure 4.12 (Continued) highlight the patterns of the data. (b) The ratio of missing loss-of-function variation for gene sets. The median ratio of missing loss-of-function variation for all genes is indicated by the dashed red line. The x-axis has been trimmed at -2 (out of -5) to highlight the patterns of the data.

The ratio of missing LoF variation is depicted for the gene lists used above in **Figure 4.12b**. All gene sets are significantly different from the set of all canonical transcripts (referred to as "All genes" in the figure; Wilcoxon $p < 10^{-10}$ for all). Autosomal recessive genes and olfactory receptors have slightly more of their expected LoF variation than the set of all transcripts. The rest of the gene sets are significantly more depleted for the expected LoF variation than the full set of transcripts. The most striking signal comes from the haploinsufficient genes, none of which have more LoF variation than expected.

Creation of pLI

One of the main goals of this work was to identify genes that are intolerant of loss-of-function variation. Given the continuous nature of the ratio of missing loss-of-function variation, it is slightly challenging to do this. To address this challenge, we estimated the probability of being loss-of-function intolerant (pLI) using the expectation-maximization (EM) algorithm.

The underlying premise of this analysis is to assign genes to one of three natural categories with respect to sensitivity to loss-of-function variation: null (where loss-of-function variation – heterozygous or homozygous - is completely tolerated by natural selection), recessive (where heterozygous LoFs are tolerated but homozygous LoFs are not), and haploinsufficient (where heterozygous LoFs are not tolerated). We assume

tolerant (null) genes would have the expected amount of LoF variation and then took the empirical mean observed/expected rate of LoF variation for recessive disease genes (0.463) and severe haploinsufficient genes (0.089) to represent the average outcome of the homozygous and heterozygous intolerant scenarios respectively. These values (1.0, 0.463, 0.089) are then used as a three-state model to which we fit the observed/expected LoF variant rate of each gene via the following analysis.

Let $\pi := (\pi_{Null}, \pi_{Rec}, \pi_{HI})$ represent the proportion of all genes that fall into each of the three proposed categories: null, recessive, and haploinsufficient.

Let λ_{Null}, λ_{Rec}, and λ_{HI} denote the expected amount of loss-of-function depletion in each of the three categories. Based on the observed depletion of LoF variation in the autosomal recessive[19,20] and ClinGen dosage sensitivity gene sets, we use:

$$\lambda_{Null} = 1$$

$$\lambda_{Rec} = 0.463$$

$$\lambda_{HI} = 0.089$$

For each gene i, we model the observed data (LoF counts) as a function of the unobserved class labels (Z_i) as follows:

$$Z_i \mid \pi \sim Cat(\pi_{Null}, \pi_{Rec}, \pi_{HI})$$

$$LoF_i \mid Z_i \sim Pois(N\lambda_{Z_i})$$

Here, LoF_i represents the observed number of LoFs in gene i and N is sample size, such that $N\lambda_{Z_i}$ is the expected number of loss-of-function variants in a gene belonging to class Z_i in the ExAC data. Our goal is to find the maximum-likelihood estimate (MLE) for π (the mixing weights of the three gene classes), and to use this

estimate to obtain an Empirical Bayes maximum a posteriori (MAP) estimate for Z_i – the probability of gene assignment to each category – for all genes $i=1...M$.

We use an expectation-maximization (EM) algorithm to find the MLE for π and Z_i, treating π as the parameters and the Z_i as the latent variables. We initialize the EM algorithm by setting $\pi^0 = (1/3, 1/3, 1/3)$.

In the E-step, we evaluate the distribution of the latent variables (Z_i) given the values of the parameters (π) from the previous iteration. The E-step is

$$p(Z_i \mid \pi_i, LoF_i) = \frac{Pois(LoF_i \mid N\lambda_{Z_i})\pi_i}{\sum_i Pois(LoF_i \mid N\lambda_{Z_i})\pi_i},$$

where $Pois$ denotes the Poisson likelihood. In the M-step, we update the parameters π with a new expectation taken under the distribution of the latent variables (Z_i) computed in the M step. The update is

$$\pi^{new} := \sum_Z p\left(Z_i \mid LoF_{i,}\pi^{old}\right)/Ngenes$$

We repeat these steps until the convergence criteria are met (π_{HI} changes by less than 0.001 from one iteration to the next).

When the EM has converged, the final mixing weights are used to determine each gene's probability of belonging to each of the categories (null, recessive, haploinsufficient).

$$Z_{i,Null} = Pois(LoF_i \mid N\lambda_{Null})$$

$$Z_{i,Rec} = Pois(LoF_i \mid N\lambda_{Rec})$$

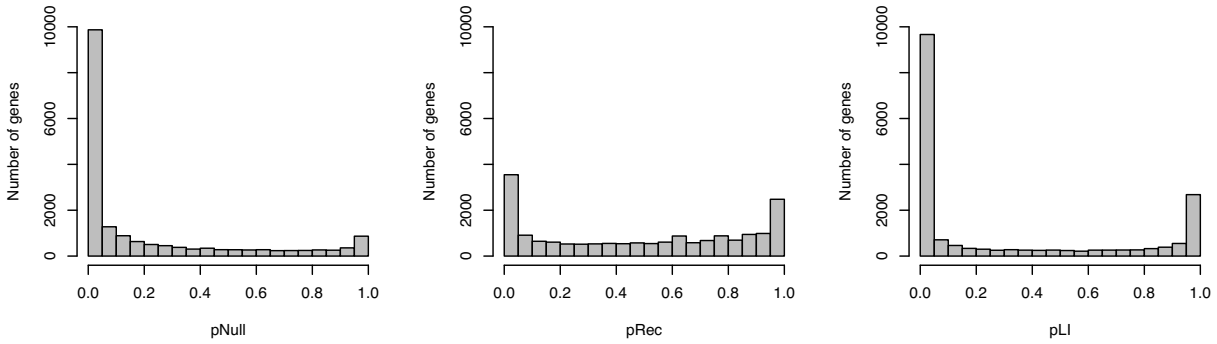
$$Z_{i,HI} = Pois(LoF_i \mid N\lambda_{HI})$$

The final metric, pLI (the probability of being loss-of-function intolerant):

$$pLI = \frac{Z_{i,HI}}{\sum Z_i}$$

The closer pLI is to 1, the more likely the transcript is loss-of-function intolerant. The overall distribution of pLI is fairly bimodal, with most genes looking either tolerant or intolerant of loss-of-function variation (Figure **4.13a**, right panel). Additionally, pLI is only modestly correlated with transcript length (r = 0.1668; **Figure 4.13b**). However, we find that the most highly LoF-intolerant genes (pLI ≥ 0.9) are significantly longer than all genes (Wilcoxon $p < 10^{-50}$; **Figure 4.14a**). The least intolerant genes are also significantly—but to a lesser extent—larger than all genes (Wilcoxon $p < 10^{-3}$; **Figure 4.14b**).

a) Distributions of pNull, pRec, and pLI



b) Relationship between transcript coding length and pLI

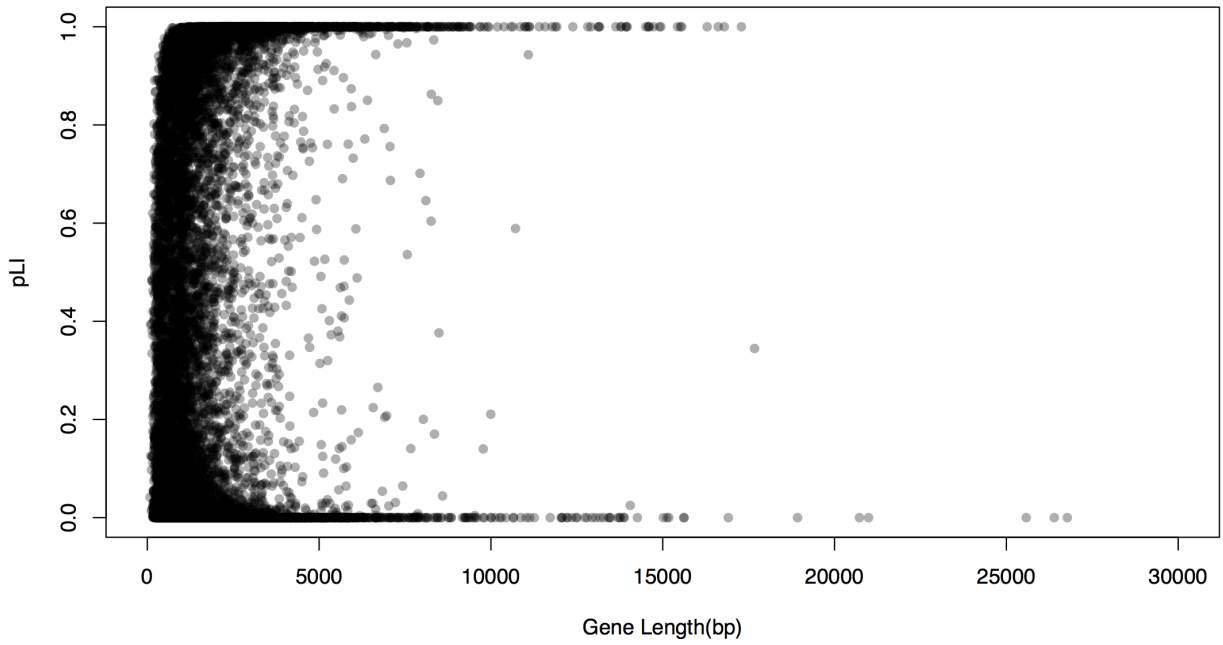
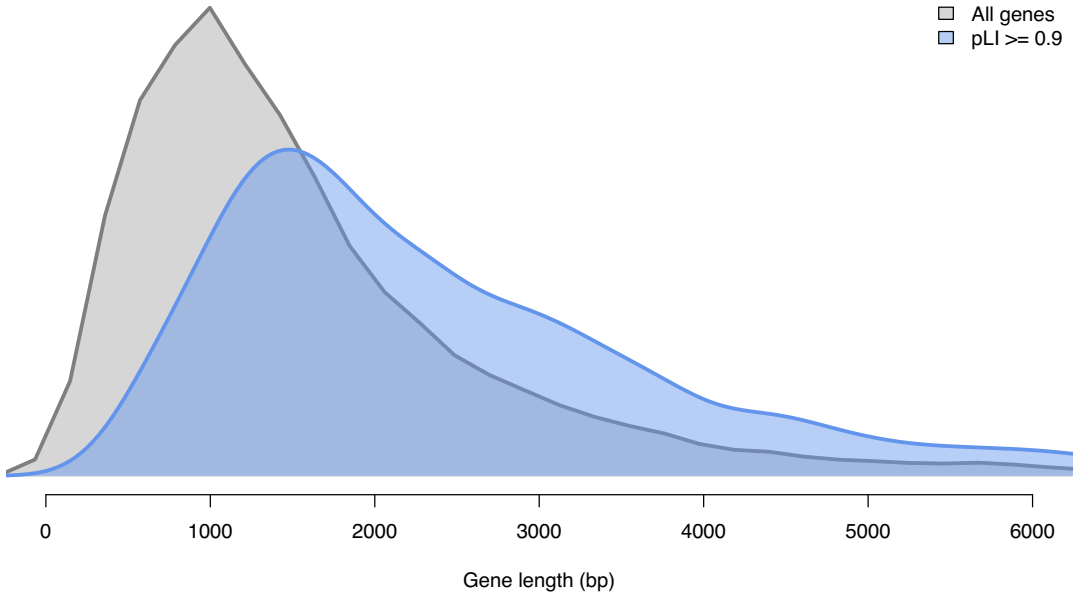


Figure 4.13. Properties of pNull, pRec, and pLI. (a) The distribution of pNull, pRec, and pLI across all transcripts. The distribution is roughly bimodal for each. (b) The relationship between pLI and the number of coding bases in each gene. The Pearson's r is 0.1668.

a) Highly loss-of-function intolerant genes (pLI ≥ 0.9**)**



b) Least loss-of-function intolerant genes (pLI ≤ 0.1)

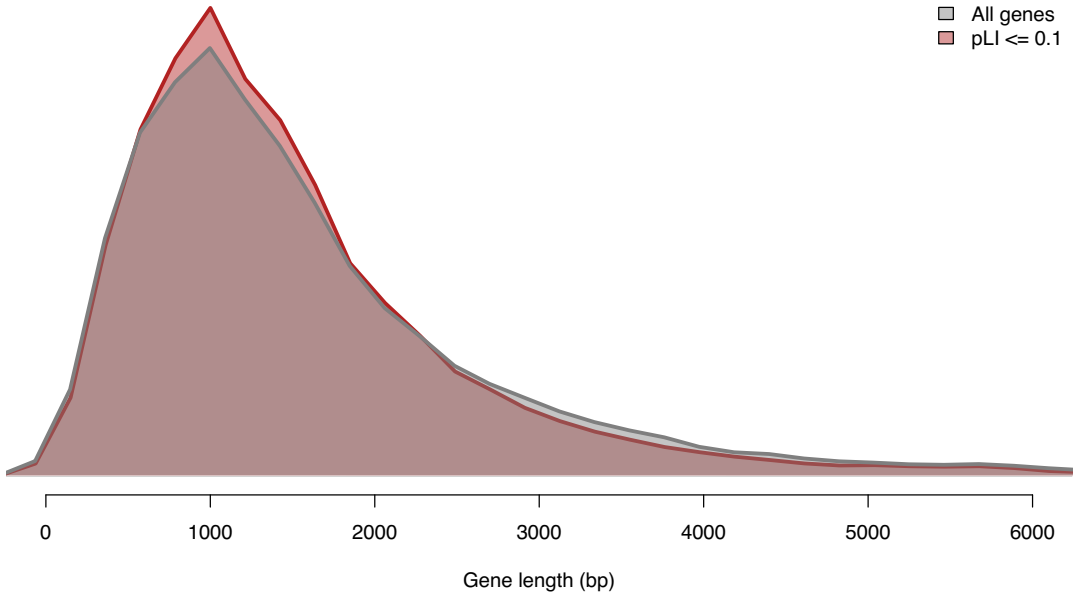


Figure 4.14. The distribution of gene length (bp) for all genes and those genes with high and low pLI values versus all genes. (a) The highly LoF intolerant genes (pLI ≥ 0.9) are significantly longer than all genes (Wilcoxon $p < 10^{-50}$). (b) The least LoF intolerant genes (pLI ≤ 0.1) are slightly significantly longer than all genes (Wilcoxon $p = 5 \times 10^{-4}$).

In order to additionally confirm that the pLI metric was free of confounding with gene length, we compare the gene size distribution of genes with a pLI ≥ 0.99 versus genes that had the pLI equivalent for falling into the recessive category (pRec) ≥ 0.99. pRec is determined by the equation below:

$$pRec = \frac{Z_{i,Rec}}{\sum Z_i}$$

We find no significant difference in the distribution of gene length between genes with pLI ≥ 0.99 (n = 1,803) and genes with pRec ≥ 0.99 (n = 1,145; p = 0.3032; depicted in **Figure 4.15**).

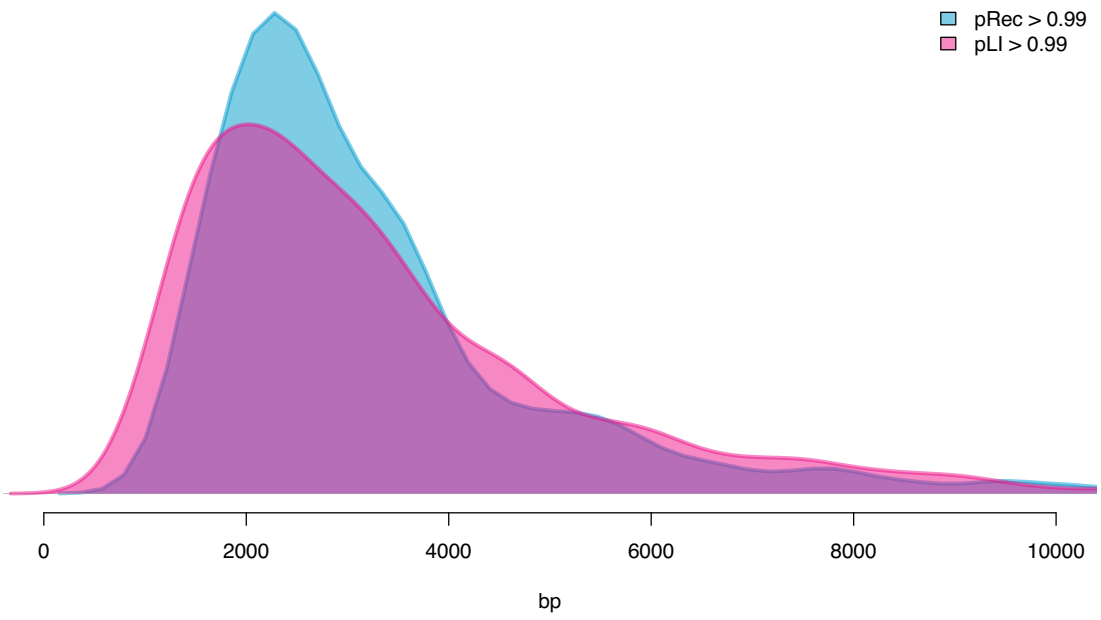


Figure 4.15. The distribution of gene length for high pLI and pRec genes. There is no significant difference between gene length (in base pairs [bp]) for genes with pLI ≥ 0.99 or pRec ≥ 0.99 (p = 0.3032).

We also show that longer genes are, in general, more depleted of LoF variation (observed/expected), which can explain the enrichment of long genes in the set of genes with pLI ≥ 0.9. There is a relationship between deciles of gene length (bins of increasing gene length) and the observed depletion of LoFs in that bin: longer genes (deciles closer to 1) have a significantly lower rate of observed/expected ($p < 10^{-50}$; **Figure 4.16**).

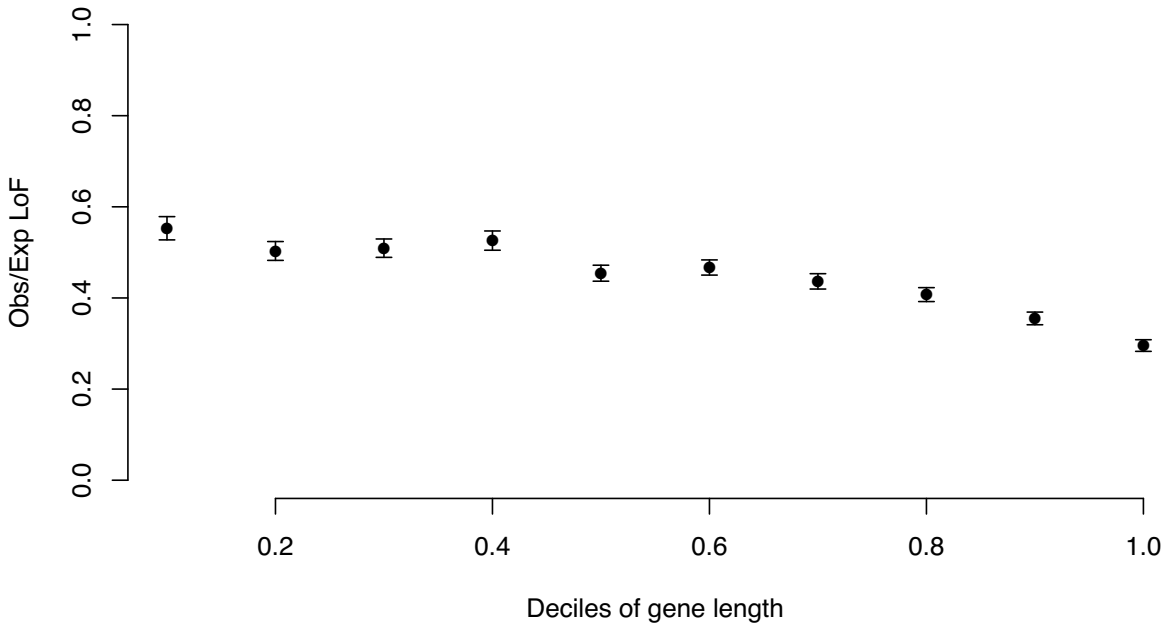


Figure 4.16. The relationship between deciles of gene length and the amount of expected variation observed. Longer genes (higher decile numbers) have a significantly lower rate of observed/expected loss-of-function (LoF) variation ($p < 10^{-50}$).

Given that the X chromosome is hemizygous in males, we expect that genes on the X would be more constrained than those on autosomes. As expected, we find the genes on the X chromosomes are significantly more constrained than those genes on the autosomes for missense and loss-of-function (synonymous p = 0.0223; missense p

= 4.43×10^{-8}; loss-of-function p = 2.50×10^{-75}). The high correlation between the observed and expected number of synonymous variants on the X chromosome (r = 0.9677 vs 0.9777 for autosomes) indicates that this difference in constraint is not due to a calibration issue.

We find that 3,230 (17.7%) of genes are confidently considered extremely loss-of-function intolerant since their pLI is 0.9 or greater. Similarly, there are 3,463 (19.0%) and 1,226 (6.7%) genes with pRec or pNull ≥ 0.9, respectively. pRec and pNull also show fairly bimodal distributions (**Figure 4.13**, middle and left panels, respectively). As a warning, while we consider pLI to be a valuable metric to identify genes that appear haploinsufficient, we caution against using pRec as a similar metric for recessive disease genes. An appropriate recessive disease gene metric would benefit from including information about the site frequency spectrum of variants observed in the gene, among other properties.

Comparison to a previous haploinsufficiency metric: p(HI)

Our metric to evaluate loss-of-function intolerance was designed to identify genes that are intolerant of heterozygous loss-of-function variants, which would mean that these genes are likely acting via haploinsufficiency. Previously, Huang et al (2010) designed p(HI)—the probability of being haploinsufficient[18]—to determine how likely each gene was to be haploinsufficient. Huang and colleagues made this metric by using properties of established haploinsufficient and haplosufficient genes to train a predictive model. The properties included in the final model were "d_N/d_S between human and

macaque, promoter sequence, embryonic expression and network proximity to known HI [haploinsufficient] genes"[18].

In order to compare pLI and Huang's p(HI), we took the 18,064 genes that had values for both metrics. Since p(HI) was trained on a set of haploinsufficient genes, we removed 64 genes that were part of their training data set and considered to be haploinsufficient by ClinGen's Dosage Sensitivity Map, which left 18,000 genes for analysis. While there are 3,175 genes in this set with pLI ≥ 0.9, there are only 613 with p(HI) ≥ 0.9. For this reason, we dropped the cut-off to 0.8, giving 3,878 genes for pLI and 1,061 for p(HI).

Within the 18,000 genes, 148 are considered haploinsufficient by ClinGen, 109 of which have a pLI ≥ 0.8. By contrast, only 51 of the 148 haploinsufficient genes have a p(HI) ≥ 0.8, and 80% of those (n = 41) also have pLI ≥ 0.8. Our metric identifies twice as many genes at the same cut off, but a larger proportion of the genes in the high p(HI) tail are considered likely haploinsufficient by both metrics.

Table **4.1a** and **b** depict the breakdown of all genes and ClinGen haploinsufficient genes, respectively, by their pLI and p(HI) values. We took those data and found the enrichment of ClinGen haploinsufficient genes in the high pLI and p(HI) tails by setting as baseline the fraction of ClinGen haploinsufficient genes with pLI and p(HI) < 0.8 compared to all genes in that category (n = 29 and 13,681, respectively). The fraction of each other category was compared to this baseline to determine the enrichment of genes that fall into each of the other categories (pLI < 0.8 and p(HI) ≥ 0.8, etc.) and is shown in **Table 4.1c**. Genes uniquely flagged by both metrics have similar

enrichments (10 for pLI versus 10.8 for p(HI)). The real enrichment, however, is found in the subset of genes that are considered likely haploinsufficient (≥ 0.8) by both metrics.

Table 4.1. Probability of Loss of Function (pLI) and Probability of Haploinsufficient (p(HI)) counts for all genes and ClinGen. The breakdown of all genes (a) and ClinGen haploinsufficient genes (b) by their pLI and p(HI) values. (c) The enrichment of ClinGen haploinsufficient genes that fall into the high pLI and p(HI) tails when taking the fraction of ClinGen genes with pLI and p(HI) < 0.8 compared to all genes.

a) Breakdown of all genes (n = 18,000) by their pLI and p(HI) values

	p(HI) < 0.8	**p(HI) ≥ 0.8**
pLI < 0.8	13681	441
pLI ≥ 0.8	3258	620

b. Breakdown of ClinGen haploinsufficient genes (n = 148) by their pLI and p(HI) values

	p(HI) < 0.8	**p(HI) ≥ 0.8**
pLI < 0.8	29	10
pLI ≥ 0.8	68	41

c. Enrichment of ClinGen haploinsufficient genes in each pLI and p(HI) category

	p(HI) < 0.8	**p(HI) ≥ 0.8**
pLI < 0.8	1.0	10.8
pLI ≥ 0.8	10.0	31.6

Evaluating loss-of-function constraint metrics

To determine which of the three protein-truncating constraint metrics (LoF Z, ratio of missing LoF variation, and pLI) is the most useful to use as a general LoF intolerance measure, we perform two tests: (1) the ability to predict known haploinsufficient genes and (2) enrichment of *de novo* LoFs found in autism spectrum disorder cases.

We perform a logistic regression using the three LoF constraint metrics to predict inclusion in the ClinGen haploinsufficient gene list. For all regressions, transcript length is included as a covariate. pLI has the highest Z-value (14.314), reflecting a more significant ability to predict haploinsufficient genes. The Z-value for LoF Z is 11.307 and is 12.164 for the ratio of missing protein-truncating variation.

For the enrichment of *de novo* LoFs, we use the published *de novo* variants from 3,982 cases with autism and 2,078 controls[16,17] and a previously described method that controls for the mutability of each gene (see Chapter 3)[5]. In brief, the probability of mutation (for a specific mutation type) is summed across all genes in a gene set and compared to the total probability of mutation (of the same type) for all genes. That fraction becomes the expected fraction of genes in the gene set that should harbor a *de novo* variant of the same type. We evaluate the observed overlap between the *de novo* list and the gene set of interest by invoking the binomial.

Since this method requires an established gene set, we took genes with pLI ≥ 0.9 (n = 3,230) and matched the set size using the genes with the highest LoF Z scores and ratio of missing LoF variation. While the fold enrichment is greatest for the ratio of missing LoF variation (enrichment = 1.9, $p < 10^{-21}$), pLI still outperforms the LoF Z score (**Table 4.2**). No significant enrichments are seen when using the control *de novo* LoFs (fold enrichments between 0.81 and 0.91).

Table 4.2. The enrichment of *de novo* loss-of-function variants (LoFs) from autism cases with the top loss-of-function intolerant genes as defined by LoF Z, the ratio of missing LoF variation, and pLI.

	LoF Z > 3.891 (n = 3,230)	**Ratio missing LoFs > 0.9061 (n = 3,230)**	**pLI ≥ 0.9 (n = 3,230)**
LoF fold enrichment	1.3656	1.9224	1.6290
p-value	5.07×10^{-12}	5.12×10^{-22}	8.31×10^{-20}

Applications of pLI

Given pLI's superior performance in predicting haploinsufficient genes and clearer interpretability than the ratio of missing LoF variation, we chose to use pLI as our main metric of LoF intolerance.

Established haploinsufficient genes are enriched in the high pLI tail (pLI ≥ 0.9, χ^2 $p < 10^{-50}$; **Figure 4.3**). Of note, the enrichment in pLI stratifies with the severity of the disease caused by the haploinsufficient genes with increasingly severe phenotypes showing increased enrichment in the highly LoF-intolerant genes (manually curated from the ClinGen dosage sensitivity list). Critically, we note that LoF-intolerant genes include virtually all known severe haploinsufficient human disease genes (**Figure 4.3**), but that 79% of these genes do not have a disease-associated variant listed in ClinVar[13] despite the clear evidence for extreme selective constraint.

The targets of FMRP[22] are also strongly enriched in the high pLI tail (pLI ≥ 0.9, χ^2 $p < 10^{-50}$; **Figure 4.3**). Dominant disease genes[19,20] and those essential in cell culture[21], however, are more evenly split between the two categories, but still enriched for pLI ≥ 0.9 (χ^2 $p < 10^{-30}$ and $p < 10^{-23}$, respectively). Olfactory receptors[23] and recessive disease genes[19,20] have low pLI scores overall, indicating that these sets are not likely

haploinsufficient. These results do not mean that recessive genes are not important to disease, but that they can on average tolerate a heterozygous LoF.

We also studied three gene lists that correspond to genes found in mice: those genes that are lethal as homozygous knock outs, genes that are lethal as heterozygous knock outs, and genes that are lethal when conditionally knocked out in adult mice (mouse gene lists were provided by Joanne Berghout from JAX). As depicted in **Figure 4.3**, the conditional lethal genes are the most enriched in the most LoF-intolerant genes, followed by the heterozygous lethal, and then the homozygous lethal genes.

Gene expression and eQTLs

To further understand the characteristics of constrained genes we investigate the association of the synonymous Z score, missense Z score, and pLI with various gene expression and regulation metrics utilizing the multi-tissue gene expression data from the Genotype-Tissue Expression (GTEx) project[14] (GTEx Analysis V4, dbGaP Accession phs000424.v4.p1) spanning 53 tissue types sampled from 212 post-mortem donors downloaded from the GTEx portal (http://www.gtexportal.org) on July 29, 2015.

The medians of log2-transformed RPKM values for each tissue are correlated with the constraint scores after excluding sex chromosomal transcripts and transcripts not expressed in the given tissue (i.e. median RPKM = 0). Given the high correlation in gene expression between the various brain regions sampled in GTEx, a composite measure for brain expression is created by taking the median expression values for each gene across these eleven brain tissue types (only one of the duplicate measurements for each cerebellum and cortex was included). This composite brain

expression measure is used instead of the individual brain regions when the per-gene median and maximum expression values across all tissues are calculated and similarly when the total number of tissues a given gene is expressed in is determined, therefore giving 41 as the maximum number of tissues in which a gene can be detected.

Consistently in each tissue, gene expression level is strongly and positively correlated with missense Z score and pLI, a result that is further strengthened after accounting for gene coding sequence length. The association with synonymous Z score, however, is non-significant or considerably subtler. Similar patterns of association are observed for the median and maximum gene expression across tissues (median gene expression is depicted in **Figure 4.4a**). Also, the total number of tissues a gene is expressed in is positively correlated with missense Z score and pLI at different RPKM cutoffs (**Figure 4.4b**; **Figure 4.17**).

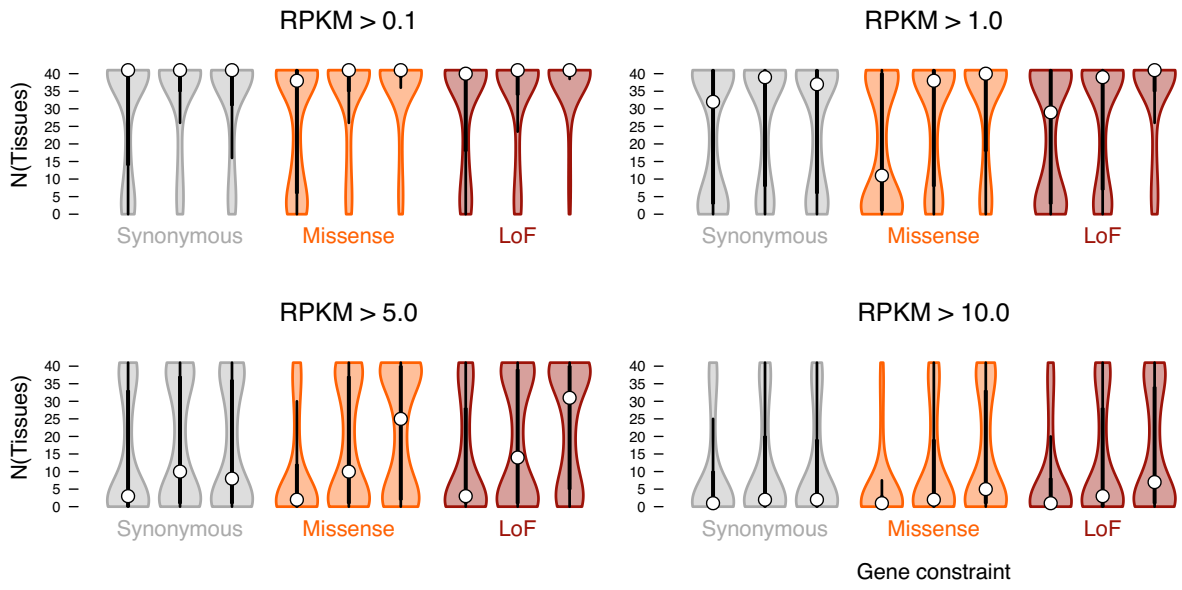


Figure 4.17. The relationship between constraint and tissue expression at different RPKM cutoffs for constraint bins. For synonymous and missense Z, the bins are: bottom

Figure 4.17 (Continued) quartile (< 25%), two middle quartiles grouped together, and top quartile (> 75%). For pLI: pLI ≤ 0.1, 0.1 < pLI < 0.9, and pLI ≥ 0.9.

The relationship between the constraint scores and gene regulatory variation detected in the GTEx dataset is investigated in the 13 tissues with the largest sample sizes (expression and genotype data available for >60 individuals) that were included in the GTEx V4 eQTL analyses (Adipose – Subcutaneous, Artery - Aorta, Artery – Tibial, Esophagus - Mucosa, Esophagus - Muscularis, Heart - Left Ventricle, Lung, Muscle – Skeletal, Nerve – Tibial, Skin - Sun Exposed (Lower leg), Stomach, Thyroid and Whole Blood). The eQTL analysis follows the steps described in detail in the GTEx pilot phase manuscript[14].

Dividing the analyzed transcripts into three subsets based on their constraint scores (for Z: bottom quartile (<25%), the two middle quartiles grouped, top quartile (>75%); for pLI: pLI ≤ 0.1, 0.1 < pLI < 0.9, pLI ≥0.9), we calculate the proportion of eGenes, i.e. a gene with a significant eQTL (FDR 5%), out of all genes included in the eQTL analysis (expressed in at least ten individuals at >0.1 RPKM) in each of the constraint subsets for each of the 13 tissues and for synonymous, missense and LoF constraint scores separately. The power for eQTL discovery varies widely from tissue to tissue given the sample sizes per tissue, which range from 74 (Artery - Aorta) to 168 (Whole Blood). Independent of the total number of eGenes discovered, in each tissue, the most missense and loss-of-function constrained group of genes are significantly depleted of eGenes compared to the least constrained group (e.g. in skeletal muscle, $p < 10^{-24}$ for pLI). Such pattern is not seen when grouping the genes based on their constraint for synonymous variation. To have a metric comparable between tissues, we

further normalize these eGene proportions by the total number of eGenes discovered in each tissue. **Figure 4.5a** shows the average proportion of eGenes in whole blood clearly demonstrating both the depletion (59.57% of the average for pLI) of eGenes among the most and enrichment (125.11% of the average for pLI) among the least missense and loss-of-function constrained genes.

Enrichment of GWAS signals

Next we investigate the same synonymous Z score, missense Z score, and pLI in the Genome-wide Association Studies (GWAS) Catalog[15] for the closest gene to signal; see Gene List table below) [Hindorff et al, Accessed 02/04/2015]. We filter results to include only those GWAS signals that had been reported with a $p < 5.0 \times 10^{-8}$. In order to categorize GWAS results by ontologies, we only include those signals that have been mapped in the "Experimental Factor Ontology" (EFO, http://www.ebi.ac.uk/efo). We find 2,792 unique genes that have been listed in the Catalog and for which we have Z scores and pLI.

As performed in previous analyses, we divide variants by functional categories: synonymous, missense and loss-of-function, and each category was further divided in three constraint groups: Lowest (0 - 25% quantile for Z; pLI ≤ 0.1), Middle (25 – 75% quantile for Z; 0.1 < pLI < 0.9) and Highest (75 – 100% quantile for Z; pLI ≥ 0.9). Then we estimate the enrichment of genes in the GWAS catalogue as:

$$E_q = P_q * S$$

$$P_q = {}^{GWAS_q}/_{GWAS}$$

$$S = {}^{N}/_{GWAS}$$

where:

P_q is the proportion of GWAS genes in the quantile q

and S is a scaling factor (number of evaluated genes divided by number of GWAS hits)

The standard error for the proportions are similarly scaled:

$$SE = \sqrt{\left(\frac{P_q(1-P_q)}{N_q}\right)} * S$$

We estimate the significance of the difference in the number of GWAS loci of highest versus the lowest constraint scores using a χ^2 test.

While only the loss-of-function category shows a clear and significant difference between the highest and the lowest constraint scores, we note a pattern in the missense category where the less constrained genes have higher, albeit not significant, proportion of GWAS hits than the middle category (**Figure 4.5b**).

To better characterize this pattern we divide the GWAS hits by major EFO categories: Cancer, Cardiovascular, Digestive, Immune, Metabolic, Nervous, Response to drug, Body measure and Others, and compare the least constrained genes versus the middle category as well as the most constrained genes versus the middle category (**Figure 4.18**). Again, we see that on average, GWAS hits are enriched in the most LoF constrained genes and depleted in the least constrained. In this sub-analysis we also identify an enrichment of Cardiovascular, Metabolic, and body measurement GWAS hits in the most missense constrained genes, while these categories with enrichments were non-significant in least missense constraint genes.

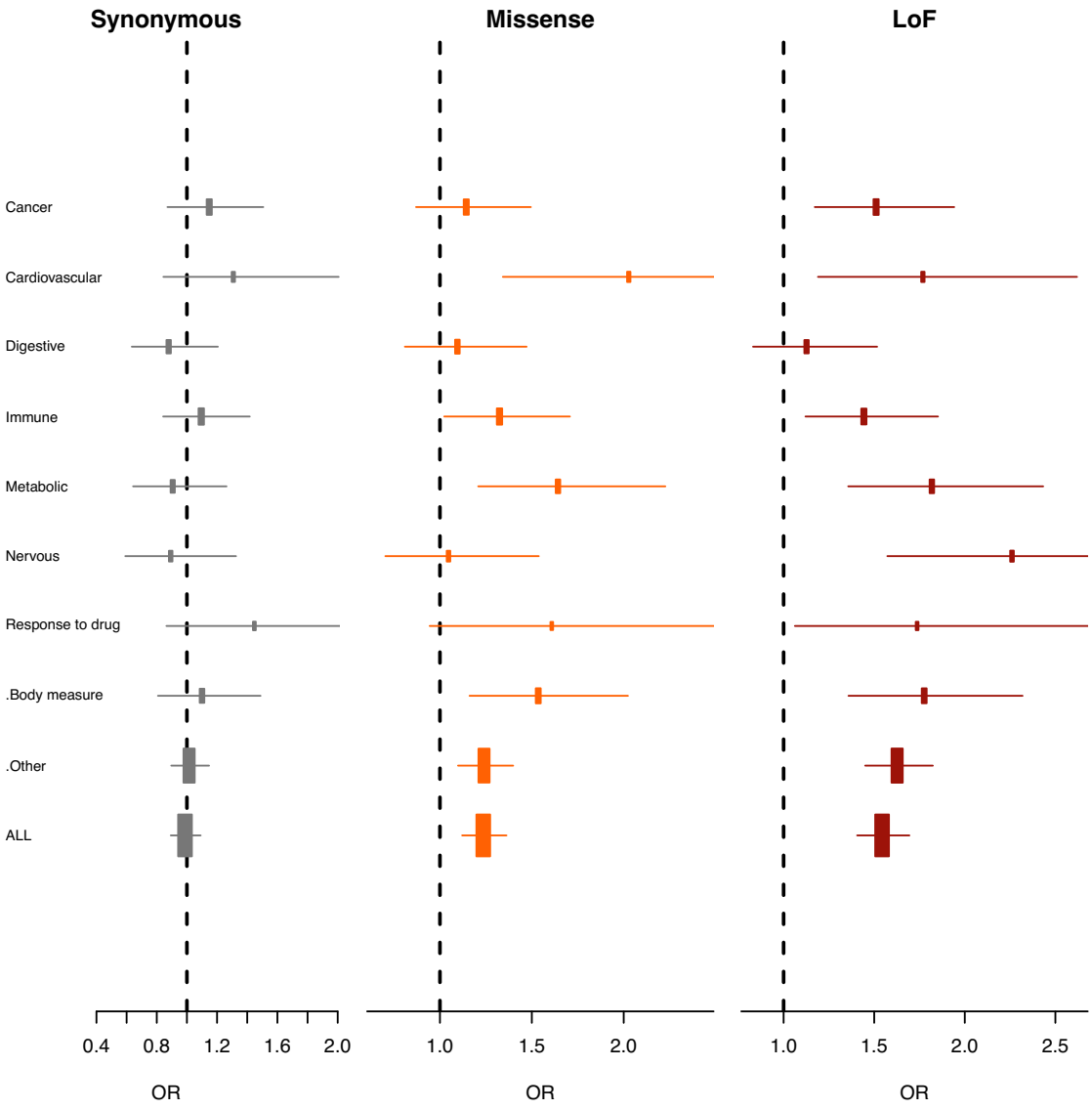


Figure 4.18. The odds ratio of being a GWAS hit for each Experimental Factor Ontology trait for the most constrained genes versus the middle bin. For synonymous and missense Z, the bins are: bottom quartile (< 25%), two middle quartiles grouped together, and top quartile (> 75%). For pLI: pLI ≤ 0.1, 0.1 < pLI < 0.9, and pLI ≥ 0.9.

Networks and pathway analysis

To better understand the set of genes considered intolerant of loss-of-function variation, we use the STRING database[24] to obtain a network of experimentally

supported protein-protein physical interactions. The network consists of 14,160 genes (nodes) and 712,137 physical interactions (edges). For each gene, we compute the number of neighbors it has in the network (degree of the node), which corresponds to the number of interaction partners its encoded protein has. We run a linear regression between the pLI score of a gene and its number of interaction partners and find that genes with more partners are more likely to have high pLI scores (t-test $p < 10^{-41}$). A weaker positive correlation is found between the number of interaction partners and the missense Z score of a gene (t-test $p < 10^{-8}$). A weak negative correlation is observed between the number of partners and the synonymous Z score (t-test $p < 10^{-6}$).

The list of 186 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were obtained from Broad Institute GSEA. Each pathway is represented by the list of pLI scores for each of the genes in the pathway. For each pathway, we compute the Kolmogorov-Smirnov (KS) statistic between its list of pLI scores and the pLI scores of all the genes to quantify the enrichment or depletion of pLI for this pathway. Fifty-eight pathways show significant deviations in pLI from the rest of the genes at multiple-testing adjusted p-value of 10^{-7} (**Table 4.3**).

Table 4.3. 58 KEGG pathways that show significant deviations in pLI.

Pathway name	p-value	Median pLI	Number of genes	Fraction of genes with duplication
KEGG_ACUTE_MYELOID_LEUKEMIA	1.13E-12	0.966868135	55	0.090909091
KEGG_SPLICEOSOME	3.22E-24	0.962186023	119	0.042016807
KEGG_ADHERENS_JUNCTION	9.98E-16	0.954642795	72	0.041666667
KEGG_PROSTATE_CANCER	4.65E-17	0.953884196	88	0.045454545
KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY	2.39E-12	0.953566812	74	0.067567568
KEGG_ENDOMETRIAL_CANCER	2.22E-09	0.945932366	50	0.04
KEGG_PROTEASOME	4.46E-07	0.939368296	44	0.068181818
KEGG_NON_SMALL_CELL_LUNG_CANCER	6.81E-11	0.937024402	53	0.056603774
KEGG_LONG_TERM_POTENTIATION	1.38E-09	0.934189798	69	0.115942029
KEGG_RENAL_CELL_CARCINOMA	9.26E-15	0.934189798	69	0.043478261
KEGG_CHRONIC_MYELOID_LEUKEMIA	8.58E-16	0.927847096	72	0.069444444
KEGG_PANCREATIC_CANCER	1.93E-12	0.914702778	69	0.057971014
KEGG_GLIOMA	3.49E-14	0.912772901	65	0.107692308
KEGG_SMALL_CELL_LUNG_CANCER	1.95E-10	0.912222953	83	0.108433735
KEGG_THYROID_CANCER	4.84E-06	0.907173398	28	0.035714286
KEGG_MTOR_SIGNALING_PATHWAY	1.08E-06	0.898304164	51	0.019607843
KEGG_WNT_SIGNALING_PATHWAY	5.43E-20	0.894082823	142	0.077464789
KEGG_AXON_GUIDANCE	3.44E-15	0.889436552	125	0.168
KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	4.10E-17	0.879812545	132	0.060606061
KEGG_MELANOMA	2.06E-10	0.86533156	69	0.057971014
KEGG_ERBB_SIGNALING_PATHWAY	1.54E-12	0.855133233	86	0.046511628
KEGG_NEUROTROPHIN_SIGNALING_PATHWAY	4.18E-18	0.833673302	124	0.10483871
KEGG_GAP_JUNCTION	1.50E-07	0.832476903	86	0.174418605
KEGG_COLORECTAL_CANCER	1.96E-11	0.82592172	60	0.033333333
KEGG_PATHWAYS_IN_CANCER	4.59E-31	0.817693684	315	0.082539683
KEGG_RIBOSOME	3.00E-22	0.791160924	86	0.046511628
KEGG_FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS	2.49E-10	0.787536053	94	0.106382979
KEGG_TGF_BETA_SIGNALING_PATHWAY	3.23E-09	0.774529936	83	0.120481928
KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY	5.22E-11	0.774064994	106	0.075471698
KEGG_MAPK_SIGNALING_PATHWAY	1.11E-20	0.726173344	257	0.062256809
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	2.92E-14	0.702854984	209	0.081339713
KEGG_HEDGEHOG_SIGNALING_PATHWAY	6.08E-06	0.68038415	55	0.163636364
KEGG_BASAL_CELL_CARCINOMA	3.18E-06	0.68038415	51	0.117647059
KEGG_PROGESTERONE_MEDIATED_OOCYTE _MATURATION	1.04E-08	0.657005772	84	0.107142857
KEGG_ENDOCYTOSIS	3.16E-11	0.656574472	175	0.08
KEGG_ALDOSTERONE_REGULATED_SODIUM _REABSORPTION	5.41E-06	0.641391597	41	0.146341463
KEGG_OOCYTE_MEIOSIS	6.30E-10	0.634553384	110	0.045454545
KEGG_FOCAL_ADHESION	1.84E-12	0.629287802	196	0.12755102
KEGG_CELL_CYCLE	1.18E-10	0.618667023	122	0.024590164
KEGG_MELANOGENESIS	5.89E-07	0.596680215	97	0.18556701
KEGG_CHEMOKINE_SIGNALING_PATHWAY	3.53E-14	0.42745098	185	0.210810811
KEGG_CARDIAC_MUSCLE_CONTRACTION	9.89E-06	0.324238693	71	0.070422535
KEGG_HUNTINGTONS_DISEASE	2.45E-07	0.301784519	170	0.070588235
KEGG_ALZHEIMERS_DISEASE	1.36E-06	0.218440721	155	0.096774194
KEGG_CYTOKINE_CYTOKINE_RECEPTOR _INTERACTION	8.60E-06	0.09785415	257	0.249027237

Table 4.3 (Continued).

pathway name	p-value	median pLI	number of genes	fraction of genes with duplication
KEGG_OLFACTORY_TRANSDUCTION	4.42E-17	0.005113489	376	0.909574468
KEGG_ARACHIDONIC_ACID_METABOLISM	2.04E-06	2.36E-05	58	0.431034483
KEGG_STEROID_HORMONE_BIOSYNTHESIS	3.81E-07	3.67E-06	54	0.685185185
KEGG_METABOLISM_OF_XENOBIOTICS_BY _CYTOCHROME_P450	9.30E-10	3.49E-06	69	0.753623188
KEGG_PENTOSE_AND_GLUCURONATE _INTERCONVERSIONS	7.42E-09	2.34E-06	27	0.703703704
KEGG_DRUG_METABOLISM_CYTOCHROME_P450	1.64E-09	9.11E-07	71	0.774647887
KEGG_RETINOL_METABOLISM	9.25E-12	2.93E-07	63	0.714285714
KEGG_DRUG_METABOLISM_OTHER_ENZYMES	2.12E-09	2.93E-07	51	0.588235294
KEGG_LINOLEIC_ACID_METABOLISM	1.37E-06	2.82E-07	29	0.586206897
KEGG_OTHER_GLYCAN_DEGRADATION	8.57E-06	2.77E-07	16	0
KEGG_ABC_TRANSPORTERS	1.65E-07	4.44E-08	44	0.272727273
KEGG_ASCORBATE_AND_ALDARATE _METABOLISM	1.51E-08	3.50E-08	25	0.8
KEGG_STARCH_AND_SUCROSE_METABOLISM	4.60E-09	3.06E-08	50	0.52

For each pathway, we quantify the degree of its redundancy by computing the fraction of its genes with a duplication in the human genome[25]. Among the highly constrained pathways (highest median pLI for the genes in the pathway) are core biological processes (spliceosome, ribosome, and proteasome components; KS test $p < 10^{-6}$ for all) while olfactory receptors are among the least constrained pathways (KS test $p < 10^{-16}$). More surprisingly, we identify multiple metabolic pathways, such as starch and sucrose metabolism (KS test $p < 10^{-9}$), as being highly unconstrained. Members of these pathways are also likely to have paralogous genes in the human genome.

Author contributions

Kaitlin Samocha: conceived and designed experiments, performed analyses not listed below, writing

Monkol Lek: data processing including variant calling, made depth of coverage file, writing edits

Taru Tukiainen: expression and eQTL analyses using GTEx, wrote the section “Gene expression and eQTLs”

Karol Estrada: GWAS analyses, wrote the section “Enrichment of GWAS signals”

James Zou: pathway and networks analyses, wrote the section “Networks and pathway analysis”

Konrad Karczewski: data processing and annotation, compiling gene lists, analysis suggestions, writing edits

Eric Minikel: compiling gene lists, analysis suggestions, writing edits

Anne O’Donnell Luria: split haploinsufficient disease genes into severe, moderate, and mild categories

Joanne Berghout: provided the mouse knockout lists

Matthew Hurles: providing the most updated p(HI) scores

Jon Bloom, Brendan Bulik-Sullivan: help with the mathematical annotation

Benjamin Neale: help with the depth of coverage correction, analysis suggestions, general guidance

Daniel MacArthur: analysis suggestions, writing edits, general guidance

Mark Daly: help with the depth of coverage correction, analysis suggestions, writing edits, general guidance

Bibliography

1. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-9 (2010).

2. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).

3. Schwarz, J.M., Cooper, D.N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361-2 (2014).

4. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet* **9**, e1003709 (2013).

5. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).

6. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216-20 (2013).

7. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

8. Goh, K.I. *et al.* The human disease network. *Proc Natl Acad Sci U S A* **104**, 8685-90 (2007).

9. Itan, Y. *et al.* The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc Natl Acad Sci U S A* **112**, 13615-20 (2015).

10. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41-2 (2001).

11. MacArthur, D.G. *et al.* A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* **335**, 823-828 (2012).

12. Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159**, 1212-26 (2014).

13. Landrum, M.J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-5 (2014).

14. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).

15. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-6 (2014).

16. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).

17. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-21 (2014).

18. Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. Characterising and predicting haploinsufficiency in the human genome. *PLoS genetics* **6**, e1001154 (2010).

19. Blekhman, R. *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr Biol* **18**, 883-9 (2008).

20. Berg, J.S. *et al.* An informatics approach to analyzing the incidentalome. *Genet Med* **15**, 36-44 (2013).

21. Hart, T., Brown, K.R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol Syst Biol* **10**, 733 (2014).

22. Darnell, J.C. *et al.* FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism. *Cell* **146**, 247-261 (2011).

23. Mainland, J.D., Li, Y.R., Zhou, T., Liu, W.L. & Matsunami, H. Human olfactory receptor responses to odorants. *Sci Data* **2**, 150002 (2015).

24. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **39**, D561-8 (2011).

25. Ouedraogo, M. *et al.* The duplicated genes database: identification and functional annotation of co-localised duplicated genes across genomes. *PLoS One* **7**, e50653 (2012).

Chapter 5

Investigating patterns of regional missense constraint within genes

Abstract

The identification of constrained genes has already proven useful when analyzing genetic variation, particularly within a clinical context. Treating the whole gene as a unit does not take advantage of the known function of elements within the gene, but was necessary when using smaller exome sequencing datasets. The size of the recently released Exome Aggregation Consortium (ExAC; n = 60,706 individuals) now permits the evaluation of constraint of regions within genes. Loss-of-function variants typically have the same effect no matter where they occur in the gene, but the deleteriousness of missense variants varies depending both on the location of the variant in the gene and the specific amino acid substitution. In this work, we use the ExAC dataset to identify patterns of regional missense constraint within genes and show that these constrained regions are enriched for both established pathogenic variants and *de novo* missense variants found in patients with a neurodevelopmental disorder. We additionally created a metric—which includes information about local missense depletion and amino acid substitution deleteriousness among other features—to aid in the prioritization of missense variants. Compared to multiple other metrics, it is the best predictor of missense variant pathogenicity and will ultimately improve variant interpretation of clinical exomes.

Introduction

The availability of large-scale exome sequencing datasets has provided the opportunity to better understand patterns and rates of variation within the human population. These resources permit the identification of genetic sequences that are

intolerant of nonsynonymous variation (constrained) and therefore more likely to be associated to disease. One signature of strong selective constraint is the depletion of nonsynonymous variation within reference populations of individuals. There is also a shift in the allele frequency spectrum of the remaining variants to increasingly rare variation. Both signatures have previously been evaluated in a set of 6,503 individuals from the National Heart, Lung and Blood Institute's Exome Sequencing Project (ESP)[1] to identify genes that are significantly missense constrained (Chapter 3)[2,3]. More recently, similar methods have been applied to the Exome Aggregation Consortium dataset (n = 60,706) and found genes intolerant of loss-of-function variation (Chapter 4). The constrained genes identified in all studies were enriched for known disease genes and harbored significantly more *de novo* loss-of-function variants identified in cases with severe neurodevelopmental disorders, establishing their medical relevance.

Identifying constrained genes has already proven to be useful in the interpretation of patient variation[4]. However, it is well known that missense variants can have dramatically different effects, depending on where they occur in the gene and the specific amino acid substitution. While the ESP dataset was not well powered to evaluate missense intolerance of sub-genic regions, the ExAC dataset permits such investigations. Determining a domain's intolerance to variation would highlight the functional components that are most sensitive to perturbation. Unfortunately, protein domain information is not known for all genes. We therefore use the exon as a basis to evaluate regional patterns of missense constraint within genes so that the method may be applied globally. In this work, we describe a method to perform this analysis and find that 15% of genes show evidence of variability in missense constraint.

We also sought to use the depletion of missense variation in the region where a variant resides to aid in variant interpretation. There are many tools to predict the deleteriousness of missense variants[5-7] and to evaluate specific amino acid substitutions[8,9]. We create a score that measures the increased deleteriousness of amino acid substitutions when they occur in missense-constrained regions. We then combine information from orthogonal deleteriousness into one metric (MPC), which outperforms all other metrics at separating pathogenic and benign missense variants.

To evaluate the usefulness of our metric outside of established disease-associated variants, we study newly arising (*de novo*) missense variants identified in cases with a neurodevelopmental disorder. Over the last 5 years, there have been many large-scale sequencing projects of parent-child trios to evaluate the role of *de novo* variation and identify genes and pathways relevant to disease etiology. These studies have focused primarily on neurodevelopmental disorders, such as intellectual disability[10,11], developmental delay[12], and epileptic encephalopathy[13]. These studies have established an important, but modest, role of *de novo* variation in these diseases. The largest excesses were seen for *de novo* loss-of-function variation, which have become the main focus for follow up research. However, there is also a significant enrichment of *de novo* missense variants in these patients, but it is modest (1.2 fold), indicating that a subset of the variants are disease-related but the majority are not. We find that the most missense constrained genes and regions contain nearly all of the excess of *de novo* missense variation in the neurodevelopmental cases and additionally show that MPC promises to be a powerful way to prioritize missense variants.

Results

Searching for regional missense constraint within transcripts

We used a set of 18,225 transcripts (see Materials and Methods for transcript filtering) and, for every exon, extracted rare (minor allele frequency [MAF] < 0.1%) missense variants from the Exome Aggregation Consortium (ExAC; n = 60,706) dataset and predicted the expected number as described previously (Chapters 3 and 4)[3].

To define regions within transcripts that were specifically missense constrained, we applied a likelihood ratio test to determine the break in between neighboring exons that most significantly (by χ^2) splits the transcript into two regions with varying levels of missense depletion. If the largest (most significant) χ^2 was above our significance threshold (≥ 10.8; $p < 10^{-3}$), we then similarly searched for a way to continue to split the transcript into regions until the best χ^2 fell below our significance threshold. If the transcript did not have strong enough evidence to be split into two regions, we tested two breaks at a time to recover transcripts that have a depleted region in the middle. We only accepted the two-break model if the χ^2 was 13.8 ($p < 10^{-4}$) or larger. The method is depicted in **Figure 5.1**.

Applying this method to 18,225 transcripts, we found evidence of regional differences in missense depletion in 2,671 transcripts (14.7%) with 1,700 having one significant break (being split into 2 segments), 919 with 3 breaks, and 52 with three or more breaks (**Table 5.1**).

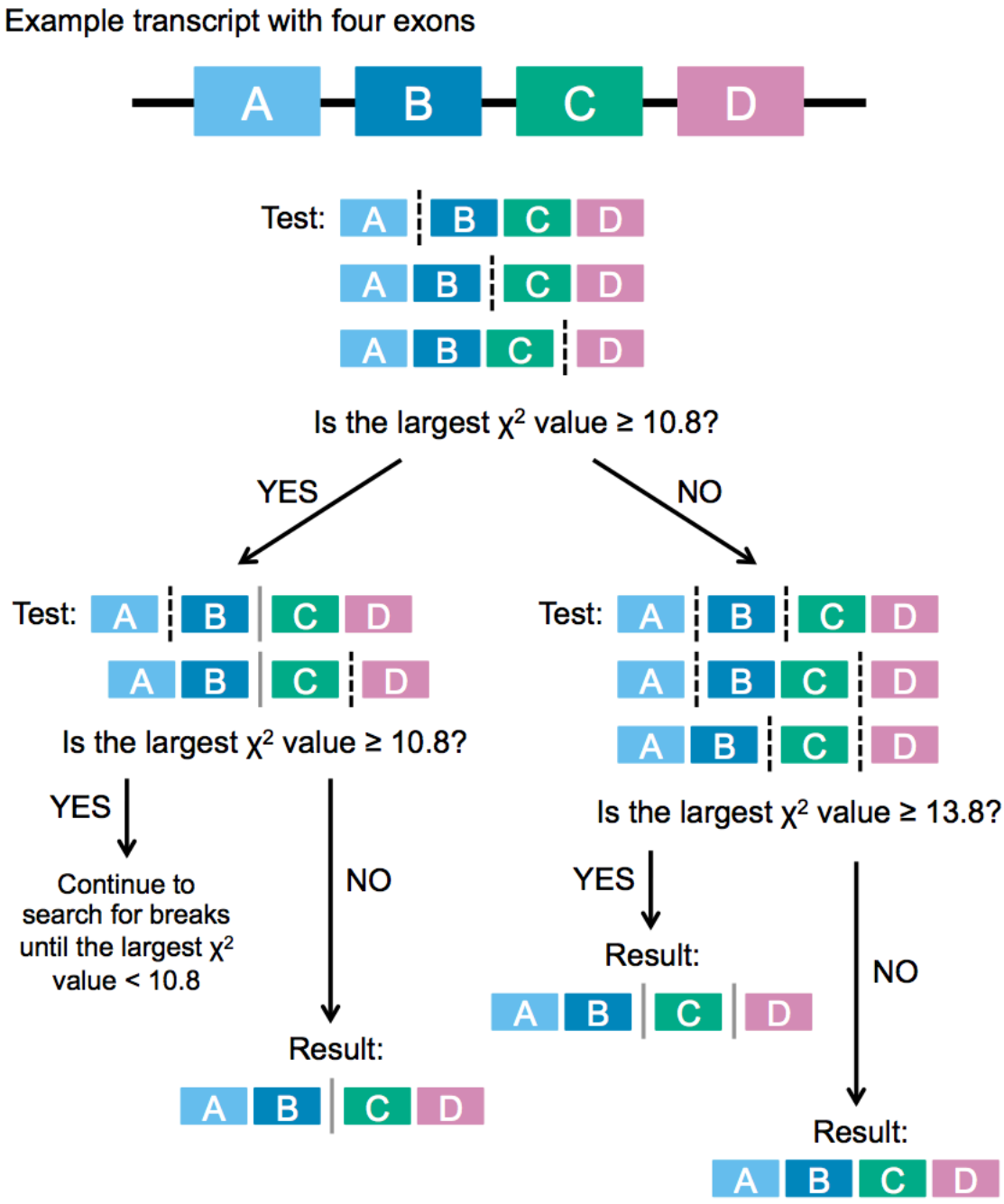


Figure 5.1. Visual of the method to find regional constraint within transcripts. The example transcript has four exons. First, all possible breaks in between exons are tested and the χ^2 are collected. If the largest $\chi^2 \geq 10.8$ ($p < \sim10^{-3}$), the method searches for a second significant break while keeping the first break set (here, the break between exons B and C). This process continues until the largest χ^2 obtained is less than 10.8 and, at that point, the last significant model is kept. If a transcript does not have

Figure 5.1 (Continued) evidence of a significant single break, the method searches for two breaks at a time. If the largest $\chi^2 \geq 13.8$ ($p < \sim 10^{-4}$), then that two break model is kept as the result. Otherwise, the transcript is considered to have no evidence of regional missense constraint.

Table 5.1. Distribution of significant breaks for all canonical transcripts.

Number of breaks	Number of transcripts	Percentage of transcripts
0	15,554	85.3
1	1,700	9.3
2	919	5.0
3	35	0.2
4	14	0.1
5	2	< 0.1
6	1	< 0.1

We plotted the fraction of expected variation observed (γ) for all full transcripts and the regions of transcripts that were split by our method (**Figure 5.2**). While most transcripts and regions of transcripts have the expected amount of missense variation, there is an excess of missense-depleted regions, particularly when $\gamma < 0.8$. All coding sequence above 0.8 does not appear to be missense constrained, so we focus our future analyses on those transcripts and regions with $\gamma < 0.8$. Within the missense constrained transcripts and regions, we further subdivided into four quartiles: [0-0.2], (0.2-0.4], (0.4-0.6], and (0.6-0.8].

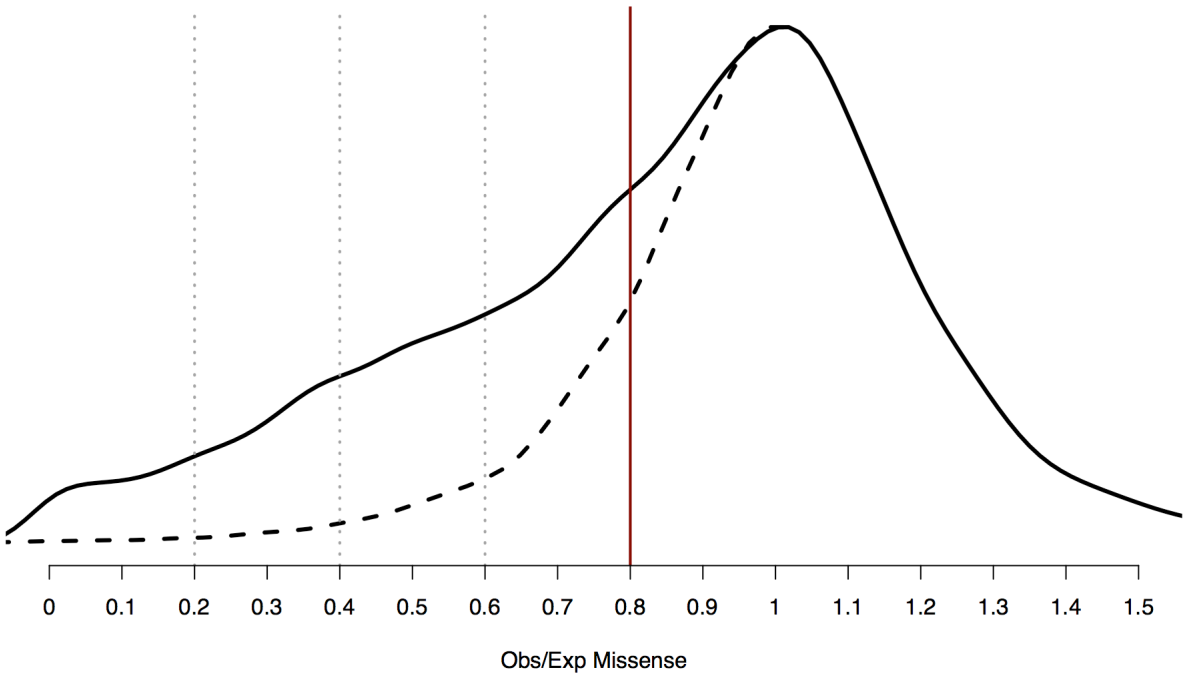


Figure 5.2. The distribution of the fraction of expected missense variation observed (observed/expected, also referred to as γ) for all transcripts and regions of transcripts. The dashed black line represents the mirror of the distribution above one. The solid red line indicates the threshold between likely missense constrained regions (γ ≤ 0.8) and regions that show no evidence of regional missense constraint (γ > 0.8). The dashed gray lines demarcate the γ quartiles used in later analyses: [0-0.2], (0.2-0.4], (0.4-0.6], and (0.6-0.8].

ClinVar variants and regional depletion

Given that the transcripts and regions with γ ≤ 0.8 are depleted of missense variation, we hypothesized that they would be enriched of disease-associated missense variants. We therefore extracted pathogenic variants from ClinVar[14] to evaluate any potential enrichments. Since our method is focused on finding regions that are intolerant of heterozygous missense variants, we selected only those variants that disrupt haploinsufficient genes known to cause severe disease (n = 440 variants).

While the missense-constrained regions (γ ≤ 0.8) represent about a third of all coding bases, they contain the great majority of the pathogenic ClinVar variants (2.7 fold enriched; $p < 10^{-50}$; **Table 5.2**). However, almost all of this enrichment is found in those transcripts and regions that have γ ≤ 0.6: 82.7% of ClinVar variants vs 15.6% of coding bases (5.3 fold enriched; $p < 10^{-50}$). These data indicate that the transcripts and regions in between 0.6 and 0.8 have a similar signature as the missense unconstrained transcripts and regions and are therefore less likely to harbor pathogenic variation that causes severe disease when disrupted.

Table 5.2. Shown for each bin of missense depletion is the count (N) and percentage (%) of coding base pairs (in megabase pairs [Mbp]), pathogenic or likely pathogenic variants from ClinVar[14] in haploinsufficient genes that cause severe disease (ClinVar). The range of missense depletion (fraction of expected missense variation observed) is provided in the first column (γ).

γ (obs/exp)	N Mbp	% Mbp	N ClinVar	% ClinVar
[0, 0.2]	0.7	2.21%	25	5.68%
(0.2, 0.4]	1.4	4.34	141	32.05
(0.4, 0.6]	2.9	9.06	198	45.00
(0.6, 0.8]	5.1	15.97	8	1.82
> 0.8	22.0	68.42	68	15.45

Of the 44 severe haploinsufficient genes, 24 (55%) have evidence of regional variability in missense constraint, and of this subset 18 (75%) contain both unconstrained and constrained regions. As an example, the first 9 exons of *CDKL5* have only 25% of their expected variation ($\chi^2 = 52.5$), but the last 11 have 81% ($\chi^2 = 6.4$). ClinVar lists 43 pathogenic or likely pathogenic missense variants in *CDKL5*, 39 (91%) of which are found in the constrained regions (**Figure 5.3**). Three of the

remaining variants are in the first 50 base pairs (bp) of exon 10 and lie in the kinase domain that extends 66 bp into that exon.

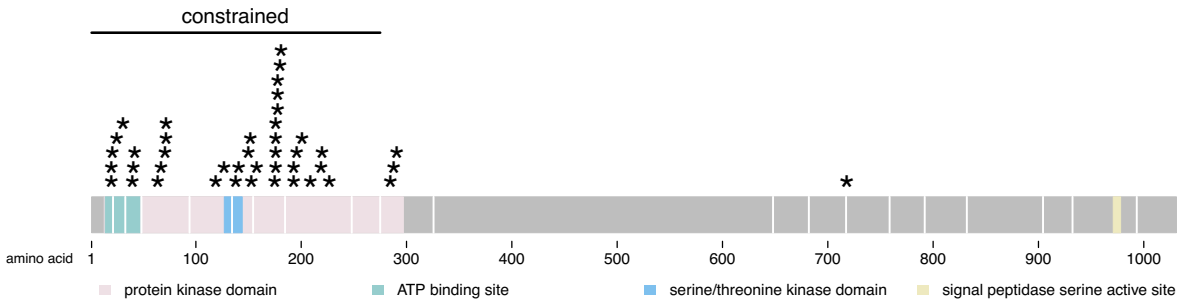


Figure 5.3. Distribution of ClinVar[14] pathogenic and likely pathogenic variants in *CDKL5*. Variants are indicated with a star. 91% of the variants (39/43) fall into the first 9 exons, which are significantly constrained (gamma = 0.25, χ^2 = 52.5). The constrained region is marked with a bar.

Using regional constraint to interpret *de novo* variation

The ClinVar variants have been established as pathogenic, but we wanted to test if our regional missense depletion results of the regions could aid in prioritization of variants identified in patients. We chose to study *de novo* missense variants from cases with a neurodevelopmental disorder (n = 1,640)[10-13] due to the significant, but modest, excess of *de novo* missense variants in these cases (1.2 fold enriched; $p = 2.3 \times 10^{-11}$; **Table 5.3**). The *de novo* missense variants from 2,078 unaffected siblings of autism cases were used as controls[15,16].

Table 5.3. Counts, fold enrichment, and significance of *de novo* variants. The observed counts (Obs), expected counts (Exp), fold enrichment (Fold), and p-value for synonymous (Syn), missense (Mis), and loss-of-function (LoF; nonsense, essential splice site, and frameshift) variants are presented for control trios[15,16], developmental delay (DDD)[12], intellectual disability (ID)[10,11], epileptic encephalopathy (EE)[13], and all neurodevelopmental cases (a combination of DDD, ID, and EE; all neuro).

		Control	DDD	ID	EE	All neuro
	N trios	2078	1133	151	356	1640
Syn	Obs	506	263	28	89	380
	Exp	582.68	317.70	42.34	99.82	459.87
	Fold	0.8684	0.8278	0.6613	0.8916	0.8263
	p-value	0.0013	0.0018	0.0254	0.3007	0.0001
Mis	Obs	1215	868	106	278	1252
	Exp	1308.86	713.64	95.11	224.23	1032.98
	Fold	0.9283	1.2163	1.1145	1.2398	1.2120
	p-value	0.0046	1.23×10^{-8}	0.1437	0.0003	2.30×10^{-11}
LoF	Obs	184	233	36	59	328
	Exp	181.71	99.07	13.20	31.13	143.41
	Fold	1.0126	2.3518	2.7265	1.8953	2.2872
	p-value	0.5868	1.91×10^{-30}	1.69×10^{-7}	5.57×10^{-6}	8.05×10^{-40}

As depicted in **Figure 5.4**, the distribution of control *de novo* missense variants between bins of missense depletion follows the distribution seen for coding base pairs. For example, 71.4% of the control variants are in regions with $\gamma > 0.8$, which represent 68.4% of all coding bases. By contrast, the *de novo* missense variants identified in patients with a neurodevelopmental disorder are enriched in the most missense-depleted regions. This is seen most strongly, as for the ClinVar variants, in regions with $\gamma \leq 0.6$ (2 fold enriched, $p < 10^{-17}$).

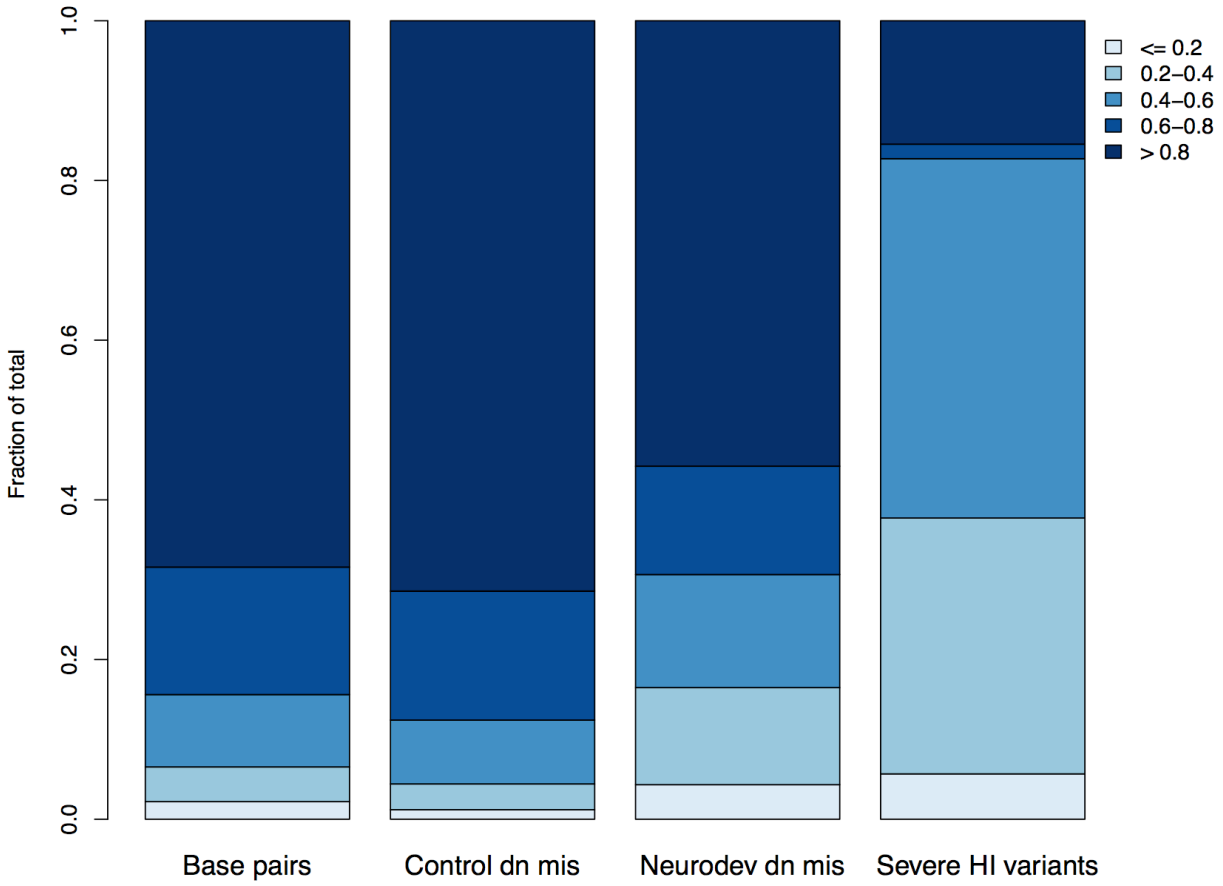


Figure 5.4. Fraction of base pairs and variants for each constraint bin. Shown for each bin of missense depletion (e.g. γ > 0.8) is the fraction of coding base pairs (base pairs), *de novo* missense variants from 2,078 control trios (control dn mis)[15,16], *de novo* missense variants from 1,640 cases with a neurodevelopmental disorder (neurodev dn mis)[10-13], and pathogenic or likely pathogenic missense variants from ClinVar[14] in haploinsufficient genes that cause severe disease (severe HI variants). Lighter blues indicate greater missense depletion.

We then compared the rate of *de novo* missense variants in cases to the rate in controls across missense constraint bins. If a region or transcript is tolerant of missense variation, we expect it to have the same rate of *de novo* variation in cases as in controls, reflecting the background rate of mutation (1:1). However, if the region is intolerant of missense variation—and therefore more likely to be associated to disease—we expect

to find a higher rate of *de novo* variants found in cases compared to the rate in controls (>1:1). As expected, the least missense-constrained bin (γ > 0.8) is indistinguishable from one (**Figure 5.5; Table 5.4**). While the most depleted two bins (γ ≤ 0.4) show a much higher rate of *de novo* missense variants in cases than in controls (OR > 4.5), there is no difference in the fourth bin (0.6 < γ ≤ 0.8). Regions and genes with more modest missense depletion (0.4 < γ ≤ 0.6) have an intermediate OR of 2.3, supporting that there is power is using the quantitative depletion of missense variation and not solely a threshold.

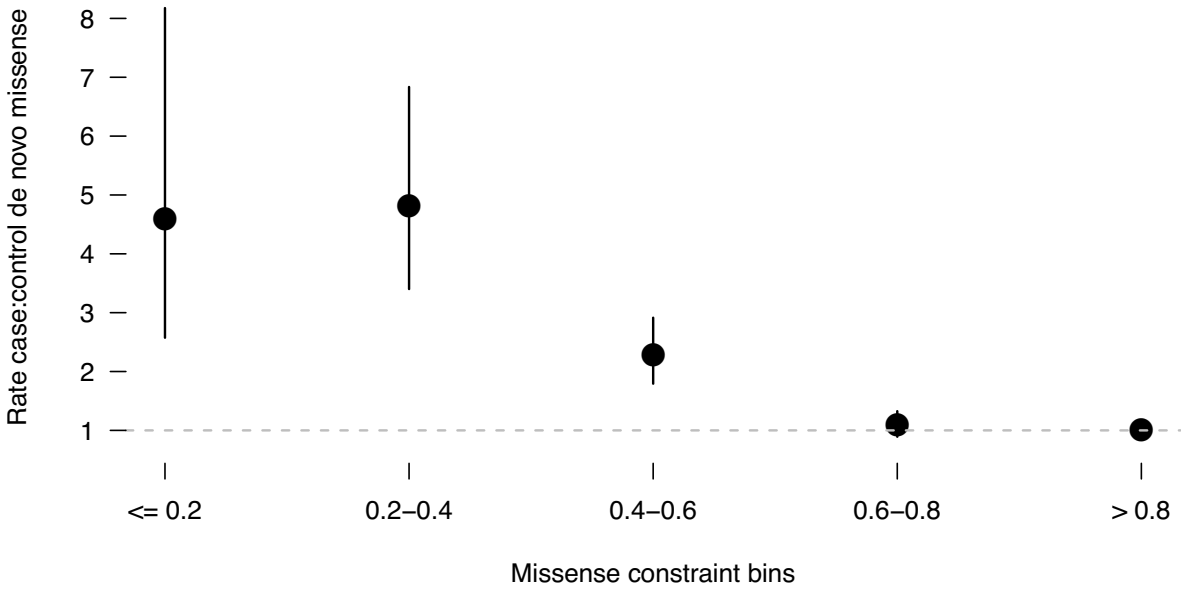


Figure 5.5. Comparison of the rate of case *de novo* missense variants to control *de novo* missense variants by bins of missense depletion. The case variants come from 1,640 trios with a neurodevelopmental disorder[10-13] and the control variants were identified in 2,078 control trios[15,16]. The dashed gray line indicates a ratio of one. 95% confidence intervals are depicted around each point estimate.

Table 5.4. Shown for each bin of missense depletion is the count (N) and percentage (%) of coding base pairs (in megabase pairs [Mbp]) for *de novo* missense variants found in 1,640 trios with a neurodevelopmental disorder (case dn)[10-13] and those from 2,078 control trios (control dn)[15,16]. The last column (C:C dn rate) provides the ratio of the neurodevelopmental case to control *de novo* missense rate. The first column lists the range of missense depletion (fraction of expected missense variation observed; γ).

γ (obs/exp)	N Mbp	% bp	N case dn	% case dn	N control dn	% control dn	C:C dn rate
(0, 0.2]	0.7	2.21%	52	4.33%	14	1.19%	4.5877
(0.2, 0.4]	1.4	4.34	146	12.16	38	3.23	4.8215
(0.4, 0.6]	2.9	9.06	170	14.15	94	7.99	2.2861
(0.6, 0.8]	5.1	15.97	163	13.57	190	16.16	1.0875
> 0.8	22.0	68.42	670	55.79	840	71.43	1.0179

Combining the three most depleted bins together (γ ≤ 0.6), there are 0.21 *de novo* missense variants per case exome and only 0.05 per control exome. However, this enrichment disappears when γ > 0.6 (0.51 events per case exome versus 0.50 in controls). It is important to note, however, that a majority (56%) of the *de novo* variants found in cases are in transcripts and regions are not considered missense constrained (γ > 0.8). These analyses have refined the signal of *de novo* variant enrichment and have shrunk the number of candidate pathogenic variants from 1,201 to 368.

Taken together, these analyses indicate that the signal for both established pathogenic variants as well as the excess of *de novo* missense variants in cases with a neurodevelopmental disorder can be found in those transcripts and regions with 60% or less of their expected missense variation.

Measuring the increased deleteriousness of amino acid substitutions

While the gene or region disrupted by a missense variant is important to consider, it is also critical to consider the specific type of amino acid substitution that occurred. Major changes in the physiochemical properties of the side chain are expected to have larger effects on the protein than more subtle changes. The deleteriousness of these changes has been quantified in a variety of metrics, the two most common of which are BLOSUM[9] and Grantham[8]. Here, we postulated that there may be specific amino acid substitutions that are preferentially eliminated when they occur in the most missense depleted regions of the exome.

To measure the increased deleteriousness of amino acid substitutions when they occur in the constrained regions of the exome, we tabulated all possible amino acid-to-amino acid substitutions that could occur in the exome via a single nucleotide mutation as well as the number observed in ExAC (with MAF < 0.1%). The rate of possible substitutions observed was determined for constrained ($\gamma \leq 0.8$) and unconstrained ($\gamma > 0.8$) regions separately; in almost all instances, we observed a higher rate in the unconstrained regions, including for synonymous variants. The fold difference between the rate in the unconstrained and constrained regions clusters for synonymous changes around one and is in the 2.5-3 range for nonsense, with missense values falling primarily in between the two (**Figure 5.6**).

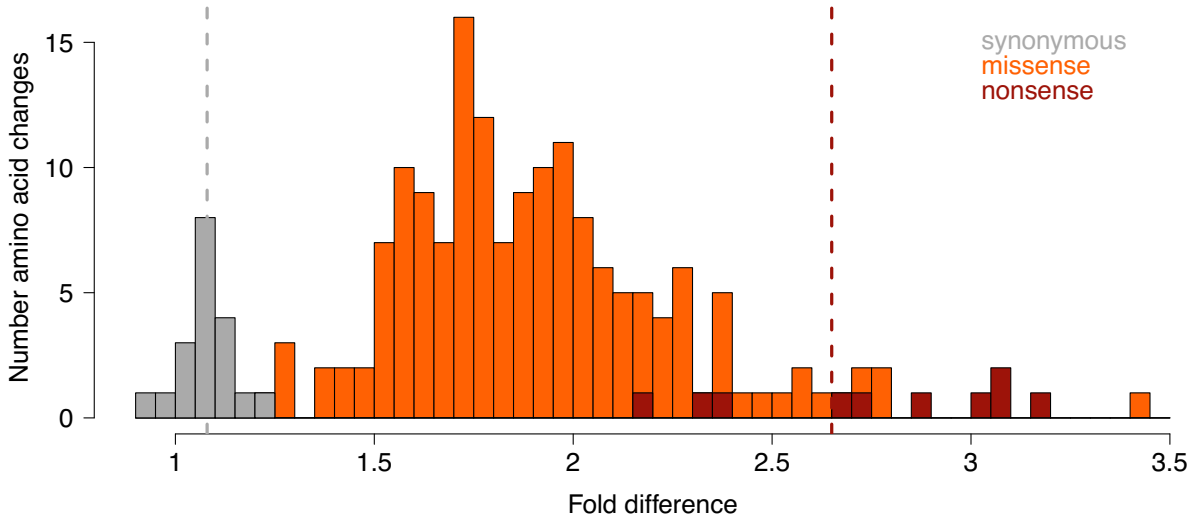


Figure 5.6. The fold difference between the rate of possible amino acid substitutions observed in unconstrained versus constrained regions. All possible amino acid substitutions that could be created by a single nucleotide mutation were tallied for unconstrained ($\gamma > 0.8$) and constrained ($\gamma \leq 0.8$) regions of the exome. The observed rate of the possible substitutions was calculated and the fold difference between that observed in the unconstrained regions versus the constrained regions is plotted. Synonymous substitutions are in gray; missense in orange; and nonsense in red. The dashed lines indicate the median of the fold differences for all synonymous substitutions (gray) and nonsense substitutions (red).

We used the normalized fold difference of missense substitutions ("missense badness") as a measure of the increased deleteriousness of amino acid substitutions when they occur in constrained genes and regions. As expected, this score has a high correlation with BLOSUM and Grantham scores (r = -0.6327 and 0.5255, respectively; **Figure 5.7**). Interestingly, we find that leucine to isoleucine substitutions are not among the most tolerant amino acid substitutions based on missense badness (missense badness = 0.42) even though the two are isoforms of each other. By contrast, both the BLOSUM and Grantham scores for this substitution indicate tolerance of the substitution (BLOSUM = 2; Grantham = 5). On the other side, serine to leucine substitutions—which

is a change from a hydrophilic to a hydrophobic side chain—are considered deleterious by BLOSUM (-2) and Grantham (145), but not by missense badness (0.20). Further investigation into these differences may reveal properties of the constrained transcripts and regions.

a) BLOSUM

b) Grantham scores

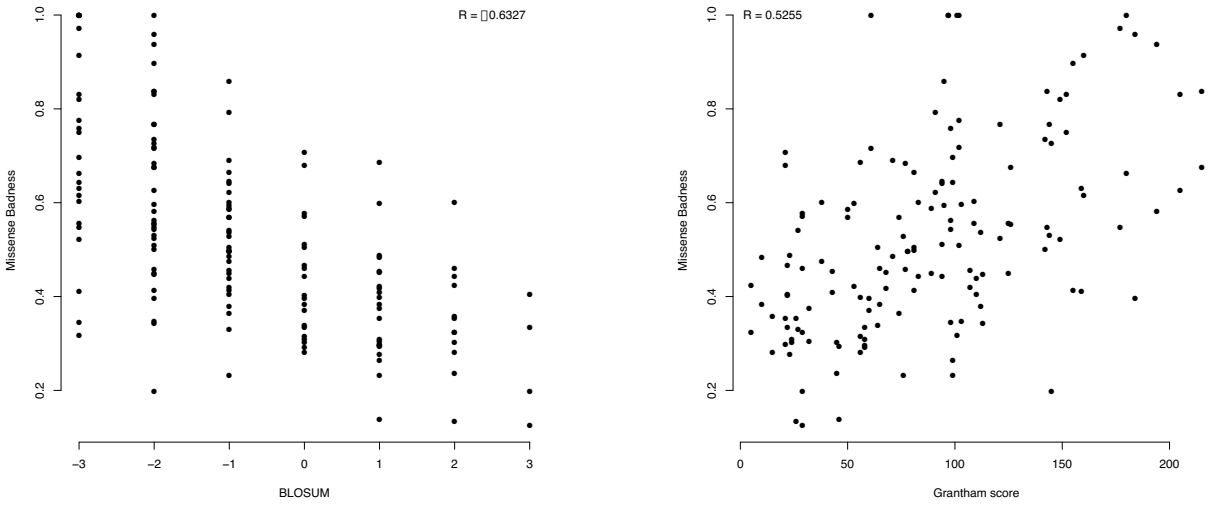


Figure 5.7. The correlations between missense badness and other metrics of amino acid substitution deleteriousness. Missense badness shows a high correlation to both BLOSUM (r = -0.6327, a) and Grantham scores (r = 0.5255, b).

Combining variant level deleteriousness scores

We wanted to determine which variant deleteriousness metric, or combination of metrics, was best at differentiating benign from pathogenic missense variants. We selected missense variants with a MAF > 1% in ExAC as our benign set (n = 93,238 variants) and used the ClinVar missense variants found in haploinsufficient genes that cause severe disease as our set of pathogenic variants (n = 1,674). The metrics we compared were: missense depletion of the region in which the variant was found (γ),

missense badness, Polyphen2[5], BLOSUM[9], and Grantham scores[8]. Using logistic regressions, we found that the best predictor of missense deleteriousness was the missense depletion (γ) of the region in which the variant was located (**Table 5.5**).

Table 5.5. Comparing the ability of various metrics to differentiate between benign and pathogenic variants. Logistic regressions were performed to determine which score could best separate benign from pathogenic missense variants. Missense variants in ExAC with a MAF > 1% were considered benign (n = 93,238). Pathogenic variants were those missense variants in ClinVar that were found in haploinsufficient genes that cause severe disease (n = 1,674). Lower AIC indicates a better predictor.

Score	AIC
Missense depletion (γ)	13967.06
Polyphen2	14615.62
Missense badness	15218.00
Grantham	15233.18
BLOSUM	15239.38

The metrics can provide complementary information, so we sought to create a composite predictor. Given that γ was by far the best score, we tested nested models and found that both missense badness and Polyphen2 significantly added to the composite predictor, but that neither BLOSUM nor Grantham did. Therefore, the best model included γ, missense badness, and Polyphen2 (**Table 5.6**), and we take the predictions as our final score, known as MPC.

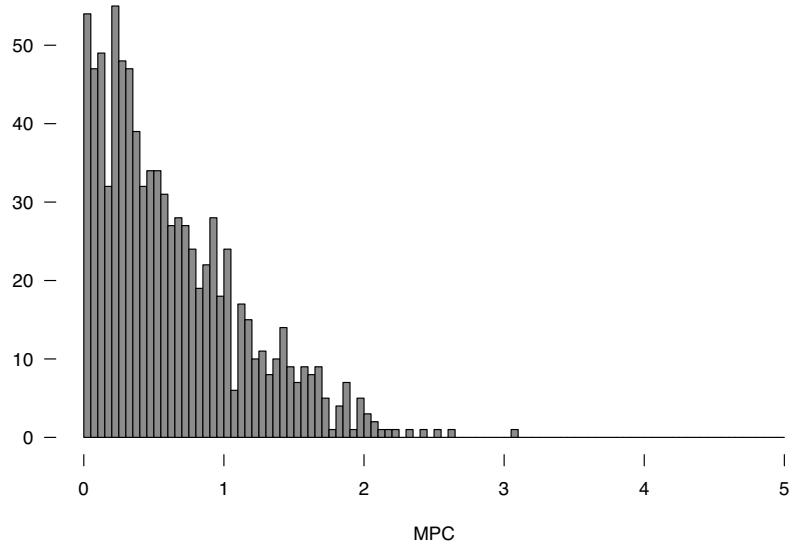
Table 5.6. The models tested combining missense depletion (obs_exp), missense badness (mis_badndess), and Polyphen2 (polyphen2). Note that when BLOSUM is added back, the predictor works less well.

Model	AIC
obs_exp + mis_badness + polyphen2	13286
obs_exp * mis_badness * polyphen2	13174
obs_exp + mis_badness + obs_exp:mis_badness + polyphen2 + obs_exp:polyphen2	13172
obs_exp + mis_badness + obs_exp:mis_badness + polyphen2 + obs_exp:polyphen2 + blosum	13176

Using MPC to evaluate the deleteriousness of *de novo* variants

We tested the usefulness of MPC by analyzing the *de novo* variants from cases with a neurodevelopmental disorder[10-13] and from controls[15,16]. The number of benign variants limits the range of MPC from 0 to 5, with increasing large numbers indicating increased deleteriousness. The distribution of MPC for the control *de novo* variants is made primarily of scores below 1 (**Figure 5.8a**). The MPC distribution for the *de novo* missense variants identified in cases with a neurodevelopmental disorder, on the other hand, appears to be made of two distributions: one following the distribution of the control *de novo* variants and the other with a peak at an MPC of 2 (**Figure 5.8b**), reinforcing that these variants are a mix of signal and noise.

a) MPC distribution for control *de novo* missense variants



b) MPC distribution for *de novo* missense variants found in cases

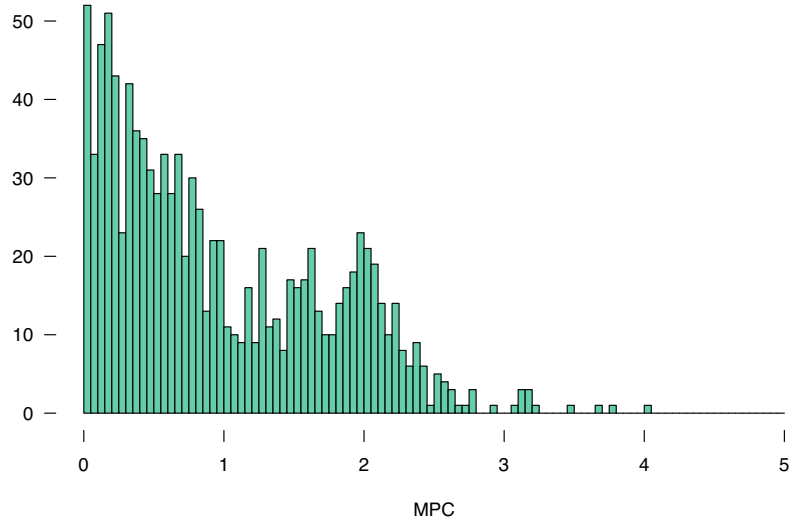


Figure 5.8. The MPC distributions for *de novo* variants in cases and controls. The MPC scores for the 1,254 *de novo* missense variants identified in control trios (a) and the MPC scores for the 1,234 *de novo* missense variants found in cases with a neurodevelopmental disorder (b).

Discussion

We have developed a method to locate regions within genes that are specifically intolerant of missense variation. Across all genes, 15% have evidence of regional variability in missense constraint, most of which are only split into two regions. We find that the genes and regions that have 60% of less of their expected missense variation—while only representing a small fraction of all coding sequence—contain 85% of pathogenic variants[14] in haploinsufficient genes known to cause severe disease. These genes and regions also contain nearly all of the excess of *de novo* missense variation that is seen in cases with a neurodevelopmental disorder[10-13].

Ideally, constraint would be calculated per base, but even the ExAC dataset is not large enough to provide sufficient power to do this. We therefore need to aggregate variant counts. While there are many options, we chose to aggregate counts across exons. Aggregating across protein domains would potentially be more informative functionally, but domain information is unfortunately unavailable for many genes. Since exons are natural biological units transferred between transcripts and are available for all genes, we believe they are currently the best option.

Moving forward, it will be important to not only include protein domain information but to consider non-linear sequences. Binding pockets are critical aspects of proteins, but are made up of amino acids scattered across the gene. Other 3D structural aspects of the protein (internal versus external residues, etc.) would also be important to consider when evaluating variant deleteriousness. Therefore, future work would greatly benefit from being able to evaluate disparate amino acids.

Since the missense constrained regions are depleted of variation due to selective pressures, we proposed that including information about the local missense depletion could improve variant deleteriousness metrics. We first created a measure of the increased deleteriousness of amino acid substitutions when they occur in missense constrained genes and regions, which outperformed similar amino acid substitution matrices (BLOSUM[9] and Grantham[8]) at separating pathogenic[14] from benign variants. The best predictor of variant deleteriousness, however, was the combination of regional missense constraint, the amino acid substitution score we developed (missense badness), and Polyphen2[5]. The MPC scores—the joint metric—for the *de novo* missense variants from neurodevelopmental cases[10-13] appeared to be a mixture of two distributions (benign and pathogenic), which matches what would be expected given the modest enrichment of such variants in the cases.

We predict that MPC will be most informative for those variants that are found in regions with intermediate missense depletion (40-60% of expected variation) since this set of variants has a lower signal to noise ratio than the variants found in the more missense depleted genes and regions. We also hope to test MPC on the *de novo* missense variants from 3,982 cases with an autism spectrum disorder (ASD)[15,16]. We previously found a relationship between the IQ of an ASD case and the rate of *de novo* loss-of-function variants, with lower IQ individuals having a higher rate[3,17]. We recently discovered a similar relationship with IQ for *de novo* missense variants that were predicted to be damaging by Polyphen2[5] and fell into one of the 1,003 missense constrained genes discussed in Chapter 3[18].

As the number of sequenced individuals increases in reference datasets such as ExAC, we will be able to further refine our analyses of regional constraint. Additionally, the aggregation of whole genome sequencing data from reference populations will allow similar analyses of noncoding regions and promises to empirically highlight genetic regions intolerant of variation. The knowledge gained from our work and similar studies will continue to improve our ability to interpret genetic variation and, therefore, understanding of the genetic basis of disease.

Materials and Methods

Transcript and exon definitions

In order to have one representative transcript for each gene, we used the canonical GENCODE (v19) transcript as defined by Ensembl 75, for protein-coding genes. We removed transcripts that lacked a methionine at the start of the coding sequence, a stop codon at the end of coding sequence, or were indivisible by three, which left 19,621 transcripts. Additionally, 795 transcripts that had zero observed variants—when dropping counts in exons with a median depth < 1 (explained below)—were removed, leaving 18,466 transcripts for analysis. The exon boundaries were defined by UCSC's annotation for GENCODE v19 (downloaded on June 16th, 2014).

Observed variant counts

We consider intolerance to loss-of-function variation to primarily be a property of a gene. We therefore searched for regional constraint to missense variation alone. To obtain the observed number of missense variants per exon, we extracted variants from

the Exome Aggregation Consortium's dataset (ExAC; n = 60,706) that met the following criteria:

(1) Defined as a missense change by the predicted amino acid substitution. Variants that would be considered "initiator_codon_variants" and "stop_lost" by annotation programs such as VEP[19] are therefore included in the total.

(2) Caused by a single nucleotide change.

(3) Had an adjusted allele count ≤ 123, corresponding to a minor allele frequency (MAF) < 0.1%. The adjusted allele count only includes individuals with a depth (DP) ≥ 10 and a genotype quality (GQ) ≥ 20.

(4) Had a VQSLOD ≥ -2.632.

Due to the VQSLOD threshold, variants were not required to have a PASS in their FILTER column. The observed counts represent the unique number of qualifying variants and not the aggregate allele count of all qualifying variants within the exon.

Expected variant counts

Expected missense variant counts were determined as described in Chapter 4. Briefly, we used a model of mutation based on sequence context and corrected for regional divergence between humans and macaques to define the probability of a mutation per exon in all canonical transcripts (as discussed in Chapters 3 and 4)[3]. We used exons with a median depth ≥ 50 and regressed the number of rare, synonymous variants on the probability of a synonymous mutation. Note that regressions were run separately for the autosomes with the pseudo-autosomal regions (PAR) of the X chromosome, the non-PAR regions of the X chromosome, and the Y chromosome. The

expectations produced by these regressions were then corrected for the median depth of coverage of the exon using the following equation:

$$depth\ adjusted\ count = \begin{cases} expected\ count, median\ depth\ \geq 50 \\ expected\ count * (0.089 + 0.217 * \ln(median\ depth), 1 \leq\ median\ depth < 50 \\ 0.089 * expected\ count, median\ depth < 1 \end{cases}$$

As mentioned above, for exons with a median depth < 1, we set both the observed and expected counts to 0.

Likelihood ratio tests to define regional constraint

Using the observed and expected counts for the 18,466 canonical transcripts, we searched for significant breaks between exons that would split the transcript into two or more regions with varying levels of missense depletion. We chose to use exons in these analyses for three main reasons: (1) the size of ExAC does not allow for base pair resolution so we must aggregate variant counts; (2) exons are a natural biological unit which are transferred between transcripts; (3) protein domain information, while ideal, is missing for many genes and we wanted an approach that would be applicable to all genes.

We assume that observed counts should follow a Poisson distribution around the expected number. We defined the null model—no regional variability in missense depletion—as the model where the overall fraction of expected missense variation observed (γ) for the transcript is used as the expectation for all segments. We then employed a likelihood ratio test to compare the null model with an alternative model where expectation was γ for each specific segment. Given that the alternative model should always have a better fit than the null, we require a χ^2 above a given threshold to establish significance.

We used the following general formula to determine the significance of a break that would split a transcript into segments *A* and *B*:

$$p_0 = Pois(obs_A, exp_A * \gamma) * Pois(obs_B, exp_B * \gamma)$$

$$p_1 = Pois(obs_A, exp_A * \gamma_A) * Pois(obs_B, exp_B * \gamma_B)$$

$$\chi^2 = 2(log(p_1) - \log(p_0))$$

Where γ is the fraction of expected variation observed across all segments in the transcript; obs_A is the observed number of missense variants in segment *A*; exp_A is the expected number of variants in segment *A*; γ_A is the fraction of expected variation observed only for segment *A*; obs_B is the observed number of missense variants in segment *B*; exp_B is the expected number of variants in segment *B*; γ_B is the fraction of expected variation observed only for segment *B*; and $Pois$ denotes the Poisson likelihood.

For the purposes of this method, all exons or sections with more observed variants than expected were assigned $\gamma = 1$ since we were looking for variation in missense depletion. In addition, exons or sections with zero observed variants were considered to have one variant to prevent $\gamma = 0$.

We first searched for a single break in between exons that would significantly ($\chi^2 \geq 10.8$, $p < \sim10^{-3}$) better model the transcript's data than the null model. If multiple significant breaks between exons were found, we took the best break as defined by the χ^2 value. If a significant break was found, we searched for a second break. This process was repeated until the best break between exons did not significantly improve on the model ($\chi^2 < 10.8$). If a transcript had no significant single break, we searched for two breaks at a time, requiring a $\chi^2 \geq 13.8$ ($p < \sim10^{-4}$) to indicate significance. Those

transcripts with $\chi^2 < 13.8$ were considered to show no evidence of regional variability in missense depletion, and were left intact. The general process is depicted in **Figure 5.1**.

Excess of missense depleted coding sequence

For all coding segments (both full transcripts and the regions of transcripts), we plotted the fraction of expected variation observed (γ; **Figure 5.2**). There is a peak at one, indicating that most transcripts and regions have the expected amount of missense variation. We expect that natural stochasticity in counts will lead to a distribution of γ around 1. Even given this, we see an excess of transcripts and regions that are depleted of missense variation. To aid in visualization, we took the distribution of transcripts and regions above one and mirrored it (displayed as a dashed line). The excess of transcripts and regions with low γ over the mirrored distribution occurs when $\gamma < 0.8$, particularly below 0.6. We therefore took 0.8 as an arbitrary cut-off between regions that are likely missense constrained ($\gamma \leq 0.8$) and those that have no evidence of missense constraint ($\gamma > 0.8$). Within the missense constrained regions and transcripts, we further subdivided into four quartiles: [0-0.2], (0.2-0.4], (0.4-0.6], and (0.6-0.8].

ClinVar pathogenic variants

To test if the genes and regions we identified as missense constrained were enriched for established disease-associated variants, we extracted variants from the July 9, 2015 release of ClinVar[14] that were labeled as “pathogenic” and “likely pathogenic”. We specifically focus on those missense variants that fell into a set of 44 haploinsufficient genes that cause severe disease (n = 440 variants). The

haploinsufficient genes were those with sufficient evidence for dosage pathogenicity (level 3) as determined by the ClinGen Dosage Sensitivity Map (www.ncbi.nlm.nih.gov/projects/dbvar/clingen/; downloaded on May 5, 2015); the severity of disease caused by variants in the genes was manually curated.

De novo variants from cases with a neurodevelopmental

Over the last five years, there have been a large number of exome sequencing studies, particularly of neurodevelopmental disorders. We collected the *de novo* variants found in 151 trios with intellectual disability[10,11], 1,133 with developmental delay[12], and 356 with an epileptic encephalopathy[13]. In these studies, there is a large excess of *de novo* loss-of-function variants (> 2 fold enriched; **Table 5.3**) but also a significant, but more modest, excess of *de novo* missense variants (1.1-1.3 fold enriched). The modest enrichment indicates that there is a set of variants contributing to disease (signal), but many neutral variants (noise). *De novo* variants from the unaffected siblings of autism cases were used as controls (n = 2,078)[15,16].

Confidence intervals around the ratio of case:control *de novo* variant rates

We compared the rate of *de novo* missense variants in cases compared to the rate in controls for the five constraint bins. To determine confidence intervals around the point estimates of the ratio of *de novo* variant rates, we took the natural logarithm of the point estimate

$$\hat{\theta} = \frac{x_1/n_1}{x_2/n_2} \ ,$$

and found the standard error

$$SE(\log\ \hat{\theta}) = \sqrt{\frac{(n_1 - x_1)/x_1}{n_1} + \frac{(n_2 - x_2)/x_2}{n_2}}$$

using the delta method. The upper and lower bounds are then transformed back to obtain the 95% confidence interval

$$\hat{\theta}\exp\left[\pm 1.96 SE\left(log\,\hat{\theta}\right)\right] \ ,$$

where x_1 is the number of case *de novo* variants; n_1 is the number of case trios; x_2 is the number of control *de novo* variants; and $n_2$is the number of control trios.

Creation of missense badness

We created a metric (missense badness) of the increased deleteriousness of specific amino acid substitutions when they occur in constrained regions to identify those substitutions that are preferentially eliminated when they occur in missense depleted sequence. We identified all possible amino acid-to-amino acid substitutions that could occur via a single nucleotide mutation and then tallied the number of these substitutions in ExAC with a MAF < 0.1%. The observed and possible were then split by whether they occurred in a gene or regions with γ ≤ 0.8 (constrained) or γ > 0.8 (unconstrained) and we determined the rate of possible substitutions observed for both groups. While we observed a higher rate of possible substitutions observed in the unconstrained regions, we noticed that synonymous changes in isoleucine and those in phenylalanine did not follow this pattern.

We used the median fold difference of all synonymous substitutions as a floor (set to 0) and the median of all nonsense substitutions as a ceiling (set to 1) and normalized the missense fold differences to create missense badness. We find a high

correlation between missense badness and other amino acid substitution matrices (r = -0.6327 for BLOSUM and 0.5255 for Grantham scores (**Figure 5.7**).

Creation of MPC, a composite missense deleteriousness score

We used logistic regressions to determine which of five deleteriousness metrics was best at separating benign from pathogenic missense variants. The metrics we compared were the missense depletion of the region in which the variant was found (γ), missense badness, Polyphen2[5], BLOSUM[9], and Grantham scores[8]. Our benign variants were missense variants with a MAF > 1% in ExAC (n = 93,238 variants). The pathogenic variants were ClinVar[14] missense variants found in haploinsufficient genes that cause severe disease (n = 1,674). The best single predictor of missense deleteriousness was the missense depletion (γ) of the region in which the variant was located (**Table 5.5**).

As the metrics provide complementary information, we used nested models to determine the best composite score starting with missense depletion (γ). Missense badness and Polyphen2 significantly added to the composite predictor, but BLOSUM and Grantham did not. We therefore tested the combination of the three significant metrics and all possible interactions between them. The best model included all three scores and the interaction between γ and missense badness as well as the interaction between γ and Polyphen2 (**Table 5.6**).

We used the best regression to predict scores for all benign and pathogenic variants. In order to make more easily interpretable numbers, we transformed the raw score (*RS*)

$$-log_{10}\left(\frac{n_{benign} < RS}{N_{benign}}\right),$$

where n_{benign} is the number of benign variants with a raw score less than *RS* and N_{benign} is the total number of benign variants. We refer to the final composite score as MPC. Since there are ~91k benign variants that had information for all three metrics, the highest MPC is ~5.

MPC contains three mostly orthogonal pieces of information for each missense variant: the missense depletion (γ) of the region in which the variant is found; the deleteriousness of the specific amino acid substitution; and the Polyphen2 score, which incorporates multiple lines of evidence (phylogenetic, structural, etc.) to determine deleteriousness of the variant.

Author contributions

Kaitlin Samocha: conceived and designed experiments, performed analyses, writing

Jack Kosmicki: helped curate *de novo* variants

Konrad Karczewski: provided the BLOSUM matrix, analysis suggestions

Eric Minikel: provided the ClinVar variant list

Anne O'Donnell Luria: split haploinsufficient disease genes into severe, moderate, and mild categories

Alex Bloemendal: help with mathematical annotation

Daniel MacArthur: analysis suggestions, guidance

Benjamin Neale: conceived of experiments, analysis suggestions, guidance

Mark Daly: conceived of experiments, analysis suggestions, guidance

Bibliography

1. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216-20 (2013).

2. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet* **9**, e1003709 (2013).

3. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).

4. Zhu, X. *et al.* Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. *Genet Med* **17**, 774-81 (2015).

5. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-9 (2010).

6. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).

7. Schwarz, J.M., Cooper, D.N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361-2 (2014).

8. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862-4 (1974).

9. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-9 (1992).

10. de Ligt, J. *et al.* Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *New England Journal of Medicine* **367**, 1921-1929 (2012).

11. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *The Lancet* **380**, 1674-1682 (2012).

12. Deciphering Developmental Disorders, S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-8 (2015).

13. Epi, K.C. & Epilepsy Phenome/Genome, P. De novo mutations in epileptic encephalopathies. *Nature* **501**, 217-221 (2013).

14. Landrum, M.J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-5 (2014).

15. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).

16. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-21 (2014).

17. Robinson, E.B. *et al.* Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *Proc Natl Acad Sci U S A* **111**, 15161-5 (2014).

18. Robinson, E.B. *et al.* Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nat Genet* (2016).

19. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).

Chapter 6

Discussion

The main goal of this thesis was to develop methods and tools to better understand rare protein-coding variation, especially within the context of interpreting such variation in disease. To that aim, we built a pipeline to robustly identify *de novo* variants from sequencing data; created a sequence-context based model of mutation; identified genes that were intolerant of missense and loss-of-function variation; and found regions of missense intolerance within genes.

Summary of results

Identifying *de novo* variation

In the study of *de novo* variation, it is especially important to be confident in the genotype calls of all members of the parent-child trio. As described in Chapter 2, we determined a set of key parameters to consider when identifying *de novo* variation: (1) the genotype likelihoods provided by the genotyping software, (2) the relative number of reference and non-reference sequencing reads, and (3) the depth of sequencing coverage. The second parameter was particularly critical as we found that the most likely explanation for a falsely called "*de novo*" variant was missing a heterozygous genotype in one of the parents.

As an additional improvement to our *de novo* identification pipeline, we used the allele frequency in a reference population of a potential *de novo* variant to compare the probability that the variant was truly *de novo* versus the probability of missing a heterozygous genotype in one of the parents. The probability of being a true *de novo* variant, in combination with the aforementioned parameters, was used to separate variants into three categories in terms of the likelihood of validating as *de novo*. We

found that the class with the strongest evidence of being *de novo* had a high rate of molecular validation (97.3% for single nucleotide changes and 92.3% for indels[1]; **Table 2.3**). Our final workflow is a sensitive and specific method to identify *de novo* variation from sequencing data of trios.

Creating a mutational model

We created a sequence-context based mutational model in order to rigorously evaluate the observed burden of *de novo* variants within cases with an autism spectrum disorder (ASD; Chapter 3). We first created a mutation rate table using intergenic single nucleotide polymorphisms (SNPs) from the 1000 Genomes Project[2] and applied it to the coding region of the genome to create a per gene probability of mutation, which we split by mutational class. The raw probabilities of mutation were corrected for only two factors: the depth of coverage at the site and the regional divergence between humans and macaques. The final probabilities of mutation formed the basis of a statistical framework to evaluate *de novo* variant burden globally, for sets of genes, and on a per-gene basis.

We also used the mutational model to predict the expected number of rare (minor allele frequency (MAF) < 0.1%) variants in the National Heart, Lung and Blood Institute's Exome Sequencing Project (ESP; n = 6,503)[3]. The high correlation between the observed and expected number of rare synonymous variants per gene (r = 0.940) supported that predictions of both missense and loss-of-function variants would also be accurate. We created a signed Z score to evaluate any deviation of observed from expected counts. While we were underpowered to analyze loss-of-function variation, we

found 1,003 genes that were significantly depleted of the expected amount of missense variation (missense Z score > 3.09). Given that the model is selection neutral, these deficits are consistent with evolutionary constraint. These constrained genes were enriched for established dominant and haploinsufficient disease genes.

We then used the statistical framework to analyze the *de novo* variants identified in 1,078 trios where the child had an autism spectrum disorder (ASD)[4-8]. We found both a global excess of *de novo* loss-of-function variants (1.57 fold enriched; $p = 2.1 \times 10^{-7}$; **Table 3.1a**) and far more transcripts harboring loss-of-function variants than expected ($p < 0.001$). An important aspect of our model was to determine the significance of burden within single genes: in this dataset, we found two genes (*DYRK1A* and *SCN2A*) had more *de novo* loss-of-function variants than expected at a significance threshold of 10^{-6} (**Table 3.2**). The targets of FMRP[9] and the missense constrained genes defined above were two gene sets that were significantly enriched for *de novo* loss-of-function variation in ASD cases (>2 fold; $p < 10^{-4}$ for both). By contrast, the *de novo* variants from 343 unaffected siblings had no significant enrichments in any category.

All analyses were repeated using the *de novo* variants found in 151 trios with intellectual disability[10,11]. The global enrichment of *de novo* loss-of-function variants was greater for intellectual disability (0.24 *de novo* loss-of-function events per exome; $p = 6.5 \times 10^{-7}$; **Table 3.4a**) and, even though there were fewer cases, there were three genes with a significant burden of *de novo* loss-of-function and missense variants (**Table 3.4c**). Given these results, we separated the ASD samples with IQ ≥ 100 from the rest of the cases. All of the significant signals in ASD—global enrichment of *de novo* loss-of-function variants, excess of genes with multiple such variants, and the enrichment of

such variants in the targets of FMRP and constrained genes—were not observed for the ASD cases with IQ ≥ 100, indicating that the genetic architecture of ASD varies between low and high IQ cases.

Finally, we found that the distributions of missense Z scores of genes harboring a *de novo* loss-of-function variant in ASD or intellectual disability cases were significantly shifted towards higher constraint (Wilcoxon $p < 10^{-6}$ for both; **Figure 3.3**). The distribution for genes with a *de novo* loss-of-function variant in an unaffected individual was no different from the overall distribution of missense Z scores. Together, these results indicated a significant role of *de novo* loss-of-function variation in ASD etiology, and that the constrained genes we identified were medically relevant.

Identifying genes intolerant of loss-of-function variation

The Exome Aggregation Consortium (ExAC) dataset, which contains protein-coding variation for 60,706 reference individuals, provided us the opportunity to investigate loss-of-function constraint (Chapter 4) and intolerance to missense variation within transcripts (Chapter 5). To identify constrained genes using the ExAC dataset, we slightly modified the mutational model to incorporate an empirically defined, and ExAC-specific, depth of coverage adjustment. While the Z score was well powered for studying missense constraint, the loss-of-function Z score was highly correlated with the number of coding bases in a transcript (r = 0.5697; **Figure 4.10a**). We therefore created pLI—the probability of being loss-of-function intolerant—which identified 3,230 genes that are extremely depleted of loss-of-function variation (pLI ≥ 0.9). Established haploinsufficient disease genes are enriched in the high pLI tail, as are dominant

disease genes[12,13], and genes found to be essential in cell culture[14] (χ^2 $p < 10^{-50}$, 10^{-30}, and 10^{-23}, respectively; **Figure 4.3**).

The most loss-of-function intolerant genes compromise core biological processes, such as members of the spliceosome and proteasome complexes. The missense Z score and pLI also show a relationship with the number of protein-protein interaction partners associated with the gene: those genes with many protein-protein interaction partners are more likely to be constrained (t-test $p < 10^{-8}$ for missense Z and $p < 10^{-41}$ for pLI). Additionally, we found that the most highly constrained missense and loss-of-function genes are expressed at higher levels and in more tissues, are depleted of eQTLs[15], and are enriched for GWAS loci[16].

Searching for patterns of missense constraint within genes

The size of the ExAC dataset also allowed us to investigate patterns of regional missense constraint within genes given the large expected number of missense variants per genes (average 170; median 127). We used the observed and expected missense variant counts per exon and applied a nested likelihood ratio test to identify significant breaks in between exons that split the gene into regions with varying levels of missense depletion. Overall, 2,738 genes (14.8%) had evidence of regional missense constraint with the majority of these being split into only two regions (**Table 5.1**).

Across all genes and regions of genes, those with 60% of less of their expected missense variation contained the majority (85%) of the ClinVar[17] pathogenic variants in severe haploinsufficient disease genes. These regions were also enriched for *de novo* missense variants in cases with a neurodevelopmental disorder[10,11,18,19], but not for *de*

novo variants found in control individuals[1,20] (**Figure 5.4**). The importance of these regions was further supported by the fact that the rate of *de novo* missense variants in cases with a neurodevelopmental is significantly higher than the rate seen in controls (2-4 fold enriched; **Figure 5.5**). Overall, we find 0.22 *de novo* missense variants per case exome and 0.07 per control exome in these missense-depleted regions. By contrast, all other regions show no difference in the number of events per exome (0.51 for cases compared to 0.50 in controls).

We used the total number of observed and possible amino acid substitutions in constrained and unconstrained regions to create missense badness, a measure of the increased deleteriousness of specific amino acid substitutions when they occur in the constrained regions of the exome. Missense badness is correlated with both BLOSUM[21] and Grantham[22] scores ($r = -0.6327$ and 0.5255, respectively; **Figure 5.7**) and was able to separate pathogenic variants from ClinVar[17] from benign variants (MAF > 1% in ExAC) better than the two other metrics.

The most accurate single predictor of whether a variant was pathogenic or benign, however, was the missense depletion of the region. Given that missense badness and missense depletion are capturing orthogonal pieces of information, we chose to find the best combination of a number of scores. The best joint metric included missense depletion, Polyphen2 score[23], and missense badness. MPC worked better than all other single metrics or combinations at separating pathogenic and benign missense variants.

Improvements and future directions

Better processing of challenging variants

Our *de novo* variant identification workflow has proven to be both sensitive and specific, but it currently does not process sites with more than three alleles (one reference and two non-reference). As more individuals are sequenced and included in the same datasets, the number of multi-allelic sites will increase and therefore the script should be updated. We are also limited by the quality of variants provided by the genotyping software. In particular, variant calls on the Y chromosome as well as indels could be much improved. Therefore, our *de novo* results are less reliable for both chromosome Y variants and indels.

Accounting for indels and methylation status of CpG sites

Our inability to reliably identify indels in sequencing data has also limited the field's ability to model indel mutation rates. A major limitation of the mutational model used throughout this thesis is that it lacks the ability to predict the expected number of indels—specifically frameshift variants—per gene. In order to study frameshift variants in our *de novo* data, we estimated the rate based on the rate of nonsense variants. While this estimate was useful for the *de novo* variant studies, we knew it was not accurate enough to predict the expected number of frameshift variants in reference populations such as ExAC, and thus we had to exclude all frameshift variants from our calculations of loss-of-function constraint.

While our mutational model accurately predicts the number of rare synonymous variants per transcript in ExAC ($r = 0.9776$), we are also aware that there are other

improvements that could be made to the mutational model itself. Two other factors that could influence mutation rate that we did not incorporate into the mutational model are the methylation status of CpG sites in the male germline and the effects of transcription-coupled repair (TCR). Cytosines in CpG dinucleotides are sometimes methylated and can then deaminate, leading to a C>T (G>A) transition. Transitions at methylated CpGs occur at a much higher rate than all other mutations, including transitions at unmethylated CpGs. Our model of mutation could therefore be improved by splitting CpGs by their methylation status in the male germline (where *de novo* variants are most likely to arise) and using separate mutation rates for the two types.

Another potential improvement to the model would be accounting for TCR, which is a DNA damage repair mechanism that corrects mutations on the template strand of transcribed genes via nucleotide excision. A signature of TCR is strand asymmetry for mutations, which is especially prominent when studying transitions that result from CpG methylation and then deamination[24]. Our early investigations into TCR indicated that it did not have a large influence on the predictions of our model, but strand asymmetry has been seen in *de novo* variants from whole genome sequencing[25], indicating that it may be important to revisit.

Probabilities of mutation for further split by mutational class

It would be useful to have the probabilities of mutation per gene split by more than simply mutation type. As an example, we could split the probability of a missense mutation by the three Polyphen2[23] categories (benign, possibly damaging, and probably damaging). We would specifically like to have the breakdown of high confidence versus

low confidence loss-of-function variants, as defined by LOFTEE (http://www.github.com/konradjk/loftee). This work is currently underway and will hopefully be released with the second release of the ExAC dataset.

Incorporating allele frequency information for loss-of-function constraint

The next wave of the ExAC dataset is predicted to have nearly 100,000 individuals as part of the reference population (D.G. MacArthur, personal communication) and would provide greatly increased power to detect constrained genes. For pLI, there are 4,621 (25%) transcripts that have a pLI between 0.1 and 0.9 that we consider to be uninterpretable due to their low expected loss-of-function counts (mean of 11.47; median of 8.25). Incorporating information from LOFTEE and removing low confidence variants, such as those that occur in the last 5% of a transcript, would also improve our loss-of-function constraint analyses. Additionally, a few of our high pLI genes have common (MAF > 0.1%) loss-of-function variants. A future improvement to the method may also include the combined allele frequency of all loss-of-function variants in the transcript. The largest drawback of this potential addition would be adding in common variants that appear to be loss-of-function, but do not have the predicted effect on the protein. This issue may be mitigated, in part, by only using high confidence LOFTEE variants.

Moving regional constraint beyond exon boundaries

We have many more analyses planned for the regional missense constraint work. We know that our method to search for regional constraint is limited by exon

boundaries. As depicted in **Figure 5.3**, 39 of the 43 pathogenic missense variants from ClinVar[17] in *CDKL5* are found in exons 1-9, which we considered constrained. While one of the remaining variants falls in the middle of the unconstrained exons (10-20), there are three pathogenic variants that lie within 50 base pairs of the beginning of exon 10, which is part of the unconstrained region in *CDKL5*, and are all within the kinase domain that extends 66 base pairs into that exon. We will be updating our method to detect regional constraint so that, once it finds a significant break in between two exons, we search amino acid by amino acid in the two nearby exons to find the best way to split the gene.

We are also working on a way to search for constraint of non-linear sequences. The current sample size of ExAC does not permit the evaluation of constraint on single bases and would require many to be combined to achieve the necessary power. However, the non-linear approach would allow us to interrogate constraint of 3D structural features of the protein, such as the amino acids in binding pockets.

Continued testing of MPC

Finally, our score of missense deleteriousness, MPC, that accounts for regional missense depletion, Polyphen2[23], and missense badness still needs to be tested against other variant prioritization tools, such as CADD[26]. We would also like to test how well it separates pathogenic from benign variants specifically in regions that have 40-60% of their expected missense variation. This is an interesting set of coding sequences to investigate since it contains a lower signal to noise ratio than sequence with < 40% of its expected missense variation.

Final thoughts

Throughout this thesis, we have sought to understand the rate and distribution of rare protein-coding variants. Our sequence-context based mutational model proved useful both to analyze the burden of *de novo* variation in trio sequencing studies and to identify genes and regions within genes that are intolerant of nonsynonymous variation. Overall, we have established methods to prioritize medically relevant variation with the goal of separating it from the vast amounts of relatively neutral variants also identified in sequencing studies.

The tools and metrics we created have become widely adopted within the field. The *de novo* identification pipeline and framework to rigorously evaluate *de novo* variation have been used in studies of schizophrenia, congenital heart disease[27], and in the children of testicular cancer survivors[28], among others. Beyond *de novo* studies, the probabilities of mutation we generated are being used outside of the context of *de novo* variation[29,30] and the missense Z scores created from the ESP dataset are being applied as a metric of genic intolerance to variation[31-33]. Finally, the constraint and pLI scores are available on the ExAC web browser (http://www.exac.broadinstitute.org) and for free download in order to aid the community in variant prioritization.

Bibliography

1. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).

2. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

3. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216-20 (2013).

4. Iossifov, I. *et al.* De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* **74**, 285-299 (2012).

5. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-245 (2012).

6. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature genetics* **43**, 585-9 (2011).

7. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250 (2012).

8. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241 (2012).

9. Darnell, J.C. *et al.* FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism. *Cell* **146**, 247-261 (2011).

10. de Ligt, J. *et al.* Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *New England Journal of Medicine* **367**, 1921-1929 (2012).

11. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *The Lancet* **380**, 1674-1682 (2012).

12. Berg, J.S. *et al.* An informatics approach to analyzing the incidentalome. *Genet Med* **15**, 36-44 (2013).

13. Blekhman, R. *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr Biol* **18**, 883-9 (2008).

14. Hart, T., Brown, K.R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol Syst Biol* **10**, 733 (2014).

15. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).

16. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-6 (2014).

17. Landrum, M.J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-5 (2014).

18. Deciphering Developmental Disorders, S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-8 (2015).

19. Epi, K.C. & Epilepsy Phenome/Genome, P. De novo mutations in epileptic encephalopathies. *Nature* **501**, 217-221 (2013).

20. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-21 (2014).

21. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-9 (1992).

22. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862-4 (1974).

23. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-9 (2010).

24. Green, P. *et al.* Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* **33**, 514-7 (2003).

25. Francioli, L.C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* **47**, 822-6 (2015).

26. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).

27. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262-6 (2015).

28. Kryukov, G.V. *et al.* Genetic Effect of Chemotherapy Exposure in Children of Testicular Cancer Survivors. *Clin Cancer Res* (2015).

29. Loveday, C. *et al.* Mutations in the PP2A regulatory subunit B family genes PPP2R5B, PPP2R5C and PPP2R5D cause human overgrowth. *Hum Mol Genet* **24**, 4775-9 (2015).

30. Hinshaw, S.M., Makrantoni, V., Kerr, A., Marston, A.L. & Harrison, S.C. Structural evidence for Scc4-dependent localization of cohesin loading. *Elife* **4**, e06057 (2015).

31. Coutelier, M. *et al.* A Recurrent Mutation in CACNA1G Alters Cav3.1 T-Type Calcium-Channel Conduction and Causes Autosomal-Dominant Cerebellar Ataxia. *Am J Hum Genet* **97**, 726-37 (2015).

32. Kumar, R. *et al.* THOC2 Mutations Implicate mRNA-Export Pathway in X-Linked Intellectual Disability. *Am J Hum Genet* **97**, 302-10 (2015).

33. O'Rawe, J.A. *et al.* TAF1 Variants Are Associated with Dysmorphic Features, Intellectual Disability, and Neurological Manifestations. *Am J Hum Genet* **97**, 922-32 (2015).

Appendix

Explanation of the appendix

Given that work presented in Chapters 2 and 3 have already been published, I have included the final versions of the main articles in this appendix. Their respective supplements can be found online.

Neale et al *Nature* 2012

Neale BM, Kou Y*, Liu L*, Ma'ayan A*, **Samocha KE***, Sabo A*, Lin CF*, Stevens C, Wang LS, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Schafer C, Liu H, Zhao T, Cai G, Lihm J, Dannenfelser R, Jabado O, Peralta Z, Nagaswamy U, Muzny D, Reid JG, Newsham I, Wu Y, Lewis L, Han Y, Voight BF, Lim E, Rossin E, Kirby A, Flannick J, Fromer M, Shakir K, Fennell T, Garimella K, Banks E, Poplin R, Gabriel S, Depristo M, Wimbish JR, Boone BE, Levy SE, Betancur C, Sunyaev S, Boerwinkle E, Buxbaum JD, Cook EH, Devlin B, Gibbs RA, Roeder K, Schellenberg GD, Sutcliffe JS, Daly MJ. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*. 2012 Apr 4;485(7397):242-5. doi: 10.1038/nature11011.

* indicates co-second authors

LETTER

doi:10.1038/nature11011

Patterns and rates of exonic *de novo* mutations in autism spectrum disorders

Benjamin M. Neale[1,2], Yan Kou[3,4], Li Liu[5], Avi Ma'ayan[3], Kaitlin E. Samocha[1,2], Aniko Sabo[6], Chiao-Feng Lin[7], Christine Stevens[2], Li-San Wang[7], Vladimir Makarov[4,8], Paz Polak[2,9], Seungtai Yoon[4,8], Jared Maguire[2], Emily L. Crawford[10], Nicholas G. Campbell[10], Evan T. Geller[7], Otto Valladares[7], Chad Schafer[5], Han Liu[11], Tuo Zhao[11], Guiqing Cai[4,8], Jayon Lihm[4,8], Ruth Dannenfelser[3], Omar Jabado[12], Zuleyma Peralta[12], Uma Nagaswamy[6], Donna Muzny[6], Jeffrey G. Reid[6], Irene Newsham[6], Yuanqing Wu[6], Lora Lewis[6], Yi Han[6], Benjamin F. Voight[2,13], Elaine Lim[1,2], Elizabeth Rossin[1,2], Andrew Kirby[1,2], Jason Flannick[2], Menachem Fromer[1,2], Khalid Shakir[2], Tim Fennell[2], Kiran Garimella[2], Eric Banks[2], Ryan Poplin[2], Stacey Gabriel[2], Mark DePristo[2], Jack R. Wimbish[14], Braden E. Boone[14], Shawn E. Levy[14], Catalina Betancur[15], Shamil Sunyaev[2,9], Eric Boerwinkle[6,16], Joseph D. Buxbaum[4,8,12,17], Edwin H. Cook Jr[18], Bernie Devlin[19], Richard A. Gibbs[6], Kathryn Roeder[5], Gerard D. Schellenberg[7], James S. Sutcliffe[10] & Mark J. Daly[1,2]

Autism spectrum disorders (ASD) are believed to have genetic and environmental origins, yet in only a modest fraction of individuals can specific causes be identified[1,2]. To identify further genetic risk factors, here we assess the role of *de novo* mutations in ASD by sequencing the exomes of ASD cases and their parents ($n = 175$ trios). Fewer than half of the cases (46.3%) carry a missense or nonsense *de novo* variant, and the overall rate of mutation is only modestly higher than the expected rate. In contrast, the proteins encoded by genes that harboured *de novo* missense or nonsense mutations showed a higher degree of connectivity among themselves and to previous ASD genes[3] as indexed by protein-protein interaction screens. The small increase in the rate of *de novo* events, when taken together with the protein interaction results, are consistent with an important but limited role for *de novo* point mutations in ASD, similar to that documented for *de novo* copy number variants. Genetic models incorporating these data indicate that most of the observed *de novo* events are unconnected to ASD; those that do confer risk are distributed across many genes and are incompletely penetrant (that is, not necessarily sufficient for disease). Our results support polygenic models in which spontaneous coding mutations in any of a large number of genes increases risk by 5- to 20-fold. Despite the challenge posed by such models, results from *de novo* events and a large parallel case–control study provide strong evidence in favour of *CHD8* and *KATNAL2* as genuine autism risk factors.

In spite of the substantial heritability, few genetic risk factors for ASD have been identified[1,2]. Copy number variants (CNVs), in particular *de novo* and large events spanning multiple genes, have been identified as conferring risk[4,5]. Although these CNVs provide important leads to underlying biology, they rarely implicate single genes, are rarely fully penetrant, and many confer risk to a broad range of conditions including intellectual disability, epilepsy and schizophrenia[6]. There are also documented instances of rare single nucleotide variants (SNVs) that are highly penetrant for ASD[3].

Large-scale genetic studies make clear that the origins of ASD risk are multifarious, and recent estimates based on CNV data put the number of independent risk loci in the hundreds[5]. Yet knowledge regarding specific risk-determining genes and the overall genetic architecture for ASD remains incomplete. Although new sequencing technologies provide a catalogue of most variation in the genome, the profound locus heterogeneity of ASD makes it challenging to distinguish variants that confer risk from the background noise of inconsequential SNVs. *De novo* variation, being less frequent and potentially more deleterious, could offer insights into risk-determining genes. Accordingly, we sought to evaluate carefully the observed rate and consequence of *de novo* point mutations in the exomes of ASD subjects.

We performed exome sequencing of 175 ASD probands and their parents across five centres with multiple protocols and validation techniques (Supplementary Information). We used a sensitive and specific analytical pipeline based on current best practices[7–9] to analyse all data and observed no heterogeneity of mutation rate across centres.

In the entire sample, we observed 161 coding region point mutations (101 missense, 50 silent and 10 nonsense), with an additional two conserved splice site (CSS) SNVs and six frameshift insertions/deletions (indels) validated and included in pathway analyses (Supplementary Table 1).

To determine whether the rate of coding region point mutations was elevated, we estimated the mutation rate in light of coverage and base context using two parallel approaches (Supplementary Information). On the basis of both models, the exome target should have a significantly increased (~30%) mutation rate compared to the genome. Conservatively, by assuming the low end of the estimated mutation rate from recent whole-genome data (1.2×10^{-8})[10], we estimate a mutation rate of 1.5×10^{-8} for the exome sequence captured here. The observed point mutation rate of 0.92 per exome is slightly but not significantly elevated versus expectation (Table 1) and is insensitive to adjustment for lower coverage regions (Supplementary Information). Indeed our rate is similar to that of ref. 11.

Per-family events were distributed exquisitely according to the Poisson distribution (Table 1), suggesting limited variation in the underlying rate of *de novo* mutation in ASD families. The relative rates

[1]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [3]Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, New York 10029, USA. [4]Seaver Autism Center for Research and Treatment, Mount Sinai School of Medicine, New York, New York 10029, USA. [5]Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15232, USA. [6]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. [7]Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [8]Department of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA. [9]Division of Genetics, Department of Medicine Brigham & Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. [10]Vanderbilt Brain Institute, Departments of Molecular Physiology & Biophysics and Psychiatry, Vanderbilt University, Nashville, Tennessee 37232, USA. [11]Biostatistics Department and Computer Science Department, Johns Hopkins University, Baltimore, Maryland 21205, USA. [12]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York 10029, USA. [13]Department of Pharmacology, University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania 19104, USA. [14]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA. [15]INSERM U952 and CNRS UMR 7224 and UPMC Univ Paris 06, 75005 Paris, France. [16]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas 77030, USA. [17]Friedman Brain Institute, Mount Sinai School of Medicine, New York, New York 10029, USA. [18]Department of Psychiatry, University of Illinois at Chicago, Chicago, Illinois 60608, USA. [19]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213, USA.

Table 1 | Distribution of events per family

Events per family	All ASD trios		Random mut. exp.‡
	Exon DN SNVs*	Exp.†	
0	71	69.7	73.2
1	62	64.2	63.8
2	28	29.5	27.8
3	10	9.1	8.1
4	2	2.1	1.8
5	1	0.4	0.3
Mean		0.920	0.871

* Exon DN SNVs include all single nucleotide variants in coding sequence but excludes indels and intronic variants.
† The expected distribution of number of trios with a given event count as determined by the Poisson.
‡ Random mut. exp. is the expectation for 175 trios based on the sequence-context mutation rate model M1 (Supplementary Information) based on the count of the number of trios that have at least 10× coverage.

of 'functional' (missense, nonsense, CSS and read-through) versus silent changes did not deviate from expectation (Table 2). We did, however, observe ten nonsense mutations (6.2%), which exceeded expectation (3.3%) (one-tailed $P = 0.04$; Supplementary Information).

We examined missense mutations using PolyPhen-2 scores[12] to measure severity, as some missense variants can severely affect function[13]. These scores showed no deviation from random expectation. The observed PolyPhen-2 scores clearly deviate from standing variation in the parents (Table 2), but such variation, even the rarest category, has survived selective pressure and so is inappropriate for comparison to *de novo* events.

We observed three genes with two *de novo* mutations: *BRCA2* (two missense), *FAT1* (two missense) and *KCNMA1* (one missense, one silent). A gene with two or more non-synonymous *de novo* hits across a panel of trios might indicate strong candidacy. However, simulations (Supplementary Information) show that two such hits are inadequate to define a gene as a conclusive risk factor given the number of observed events in the study.

From analyses of secondary phenotypes (Supplementary Tables 2 and 3), the most striking result is that paternal and maternal age, themselves highly correlated ($r^2 = 0.679$, *P*-value <0.0001), each strongly predicts the number of *de novo* events per offspring (paternal age, $P = 0.0013$; maternal age, $P = 0.000365$), consistent with aggregating mutations in germ cells in the paternal line[14]. Consistent with a liability threshold model, there is an increased rate of *de novo* mutation in female versus male cases (1.214 for females versus 0.914 for males); however, the difference is not significant, owing to limited sample size. Considering phenotypic correlates, we observed no rate difference between subjects with strict autism versus those with a broader ASD classification, between positive and negative family history, or any significant effect of *de novo* mutation on verbal, non-verbal or full-scale IQ (Supplementary Table 3).

Given that hundreds of loci are apparently involved in autism[5] and *de novo* mutations therein affect ASD risk, we modelled different numbers of risk genes and penetrances (Supplementary Information) and show that a model of hundreds of genes with high penetrance mutations is excluded by our data; however, more modest contributions of *de novo* variants are not. For example, up to 20% of cases carrying a *de novo* event conferring a 10- or 20-fold increased risk is consistent with these data (Supplementary Table 4). Thus, our data are consistent with either chance mutation or a modest role for *de novo* mutations on risk. Importantly, a single deleterious event is unlikely to fully explain disease in a patient.

Table 2 | Rates of mutation annotation given variant type

Type of *de novo* mutation	*De novo* (%)*	Random *de novo* (%)	Singletons (%)†	Doubletons (%)†	≥3 (%)†
Missense	62.7	66.1	59.5	55.4	48.8
Nonsense	6.2	3.3	1.2	0.8	0.4
Synonymous	31.1	30.6	39.3	43.8	50.8
PolyPhen-2 missense classification					
Benign	35.0	35.9	46.6	51.3	63.4
Possibly damaging	21.0	18.9	18.8	17.7	15.1
Probably damaging	44.0	45.2	34.7	31.0	21.4

* All indels and failing variants were removed.
† Singletons, doubletons and ≥3 (copies) are only those variants called in 192 parents.

We therefore posed two questions of the group of genes harbouring *de novo* functional mutations: do the protein products of these genes interact with each other more than expected, and are they unusually enriched in, or connected to, previous curated lists of ASD-implicated genes? Using an *in silico* approach (DAPPLE)[15], the protein–protein connectivity defined by InWeb[16] in the set of 113 genes harbouring functional *de novo* mutations was evaluated. These analyses (Fig. 1) showed significantly greater connectivity among the *de novo* identified proteins than would be expected by chance ($P < 0.001$) (Supplementary Information).

Querying previously defined, manually curated lists of genes[3] associated with high risk for ASD with or without intellectual disability (Supplementary Table 5), and high-risk intellectual disability genes (Supplementary Table 6), we asked whether there was significant enrichment for *de novo* mutations in these genes. Five genes with functional *de novo* events were previously associated with ASD and/or intellectual disability (*STXBP1*, *MEF2C*, *KIRREL3*, *RELN* and *TUBA1A*); for four of these genes (all but *RELN*) the previous evidence indicated autosomal dominant inheritance.

We then assessed the average distance (D_i, Supplementary Fig. 2) of the *de novo* coding variants in brain-expressed genes (see supplement) to the ASD/intellectual disability list using a protein–protein interaction background network. To enhance power, data from a companion study[11] were used, including the observed silent *de novo* variants and *de novo* variants in unaffected siblings as comparators. The average distance for non-synonymous variants was significantly smaller for the case set than the comparator set (3.66 ± 0.42 versus 3.78 ± 0.59; permutation $P = 0.033$) (Supplementary Fig. 3). Much of this signal comes from 31 synaptic genes identified by three large-scale synaptic proteomic studies ($D_i = 3.47 \pm 0.46$ versus 3.57 ± 0.60; permutation $P = 0.084$) (Fig. 2; see also Supplementary Fig. 4 for the complete data). Taken in total, these independent gene set analyses, along with the modest enrichment of *de novo* variants over background rates in

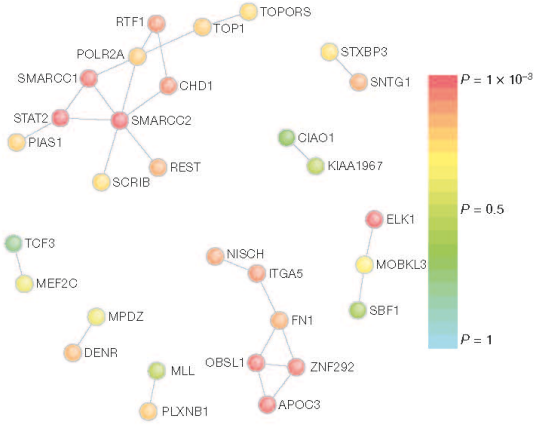


Figure 1 | Protein–protein interaction for genes with an observed functional *de novo* event. Direct protein connections from InWeb, restricting to genes harbouring *de novo* mutations for DAPPLE analysis. Two extensive networks are identified: the first is centred on SMARCC2 with 12 connections across 11 genes; the second is centred on FN1 with 7 connections across 6 genes. The *P* value for each gene having as many connections as those observed is indicated by node colour.

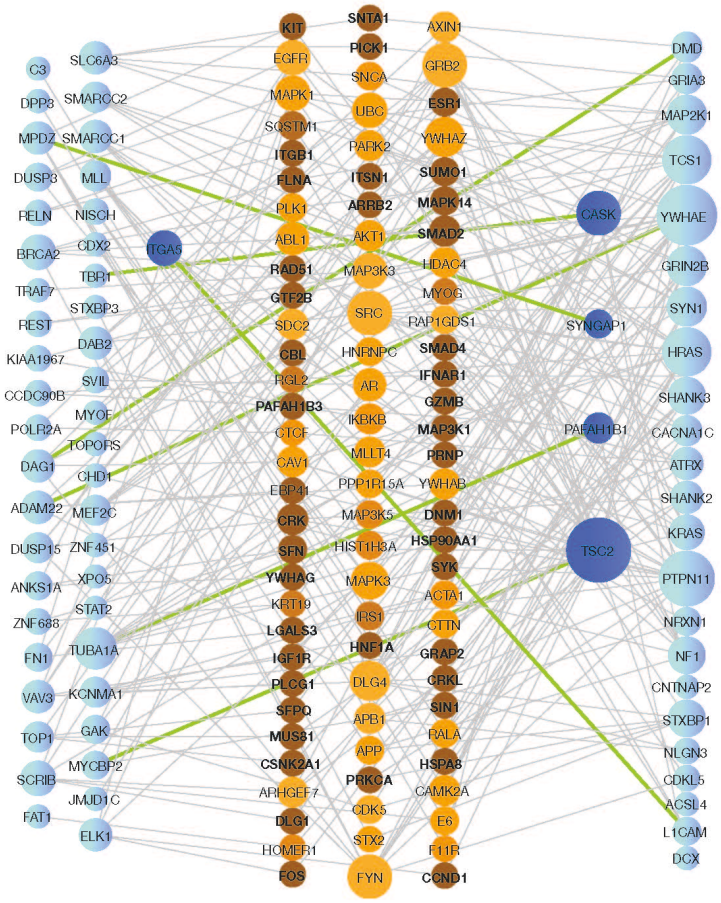


Figure 2 | Direct and indirect protein–protein interaction for genes with a functional *de novo* event and previous ASD genes. PPI network analysis for *de novo* variants and 31 previous synaptic ASD genes (see Supplementary Information). Nodes are sized based on connectivity. Genes harbouring *de novo* variants (left) and previous ASD genes (right) are coloured blue, with dark blue nodes representing genes that belong to one of these lists and are also intermediate proteins. Intermediate proteins (centre) are coloured in shades of orange based on a *P* value computed using a proportion test, where a darker colour represents a lower *P* value. Green edges represent direct connections between genes harbouring *de novo* variants (left) and previous ASD genes. All other edges, connecting to intermediate proteins, are shown in grey.

ASD, indicate that a proportion of the *de novo* events observed in this study probably contribute to autism risk.

Using whole-exome sequencing of autism trios, we demonstrate a rate, functional distribution and predicted impact of *de novo* mutation largely consistent with chance mutational processes governed by sequence context. This lack of significant deviation from random mutational processes indicates a more limited role for the contribution of *de novo* mutations to ASD pathogenesis than has previously been suggested[17], and specifically highlights the fact that observing a single *de novo* mutation, even an apparently 'severe' loss-of-function allele, is insufficient to implicate a gene as a risk factor. Yet the pathway analyses presented here assert that the overall set of genes hit with functional *de novo* mutations is not random and that these genes are biologically related to each other and to previously identified ASD/intellectual disability candidate genes. Modelling the *de novo* mutational process under a range of genetic models reveals that some models are inconsistent with the observed data—for example, 100 rare, fully penetrant Mendelian genes similar to Rett's syndrome—whereas others are not inconsistent, such as spontaneous 'functional' mutation in hundreds of genes that would increase risk by 10- or 20-fold (Supplementary Table 4). Models that fit the data are consistent with the relative risks estimated for most *de novo* CNVs[5] and suggest that *de novo* SNVs, like most CNVs, often combine with other risk factors rather than fully cause disease. Furthermore, these models indicate that *de novo* SNV events will probably explain <5% of the overall variance in autism risk (Supplementary Table 4).

Considering the two companion papers[11,18], 18 genes with two functional *de novo* mutations are observed in the complete data. Using simulations, 11.91 genes on average harbour functional mutations by chance (Supplementary Table 7). Thus, a set of 18 genes with two or more hits is not quite significant ($P = 0.063$). Matching loss-of-function variants, however, at *SCN2A*, *KATNAL2* and *CHD8* (Supplementary Table 7) are unlikely to occur by chance because of the expected very low rate of *de novo* nonsense, splice and frameshift variants. We evaluated these strong candidates further using exome sequencing on 935 cases and 870 controls, and at both *KATNAL2* and

CHD8 three additional loss-of-function mutations were observed in cases with none in controls. No additional loss-of-function mutations were seen at *SCN2A* in the case–control data, but a new splice site *de novo* event has been validated in an additional autism case while this paper was in press, strengthening the evidence for this gene as relevant to autism. Using data from more than 5,000 individuals in the NHLBI Exome Variant Server (http://evs.gs.washington.edu/EVS/) as additional controls, three loss-of-function mutations were seen in *KATNAL2* but none in *CHD8*, making the additional observation of three *CHD8* loss-of-function mutations in our cases significant evidence ($P < 0.01$) of this being a genuine autism susceptibility gene. Not all genes with double hits are nearly so promising (Supplementary Information and Supplementary Tables 8 and 9), supporting the estimate above that most of such observations are simply chance events. Overall, these data underscore the challenge of establishing individual genes as conclusive risk factors for ASD, a challenge that will require larger sample sizes and deeper analytical integration with inherited variation.

METHODS SUMMARY

We ascertained probands using the Autism Diagnostic Interview-Revised (ADI-R), the Autism Diagnostic Observation Schedule-Generic (ADOS) and the DSM-IV diagnosis of a pervasive developmental disorder. All probands met criteria for autism on the ADI-R and either autism or ASD on the ADOS, except for the three subjects that were not assessed with the ADOS. All subjects provided informed consent and the research was approved by institutional human subjects boards.

For 175 trios, we performed exome capture and sequencing using either the Agilent 38Mb SureSelect v2 ($n = 118$), the NimbleGen Seq Cap EZ SR v2 ($n = 51$), or NimbleGen VCRome 2.1 (Baylor $n = 6$). After capture, another round of LM-PCR was performed to increase the quantity of DNA available for sequencing. All libraries were sequenced using an IlluminaHiSeq2000.

All sequence data were processed with Picard (http://picard.sourceforge.net/), which recalibrates quality scores and local realignment at known indels[8] and BWA[7] for mapping reads to hg19. SNPs were called using GATK[8,9] for all trios jointly. Putative *de novo* mutations were identified restricting to sites passing standard filters and both parents were homozygous for the reference sequence and the offspring was heterozygous, and each genotype call was made confidently (see Supplementary Information).

All putative *de novo* events were validated by sequencing the carrier and both parents using Sanger sequencing methods (71 trios) or by using Sequenom MALDI-TOF (104 trios). All events were annotated using RefSeq hg19.

We modelled a Poisson process consistent with the mutation model and observed data. We varied the fraction of genes that influence risk, the probability of a functional variant, and the penetrance of said events.

We performed association tests using SKAT[19], a generalization of C-alpha[20]. Our primary analyses treat case–control data generated at Baylor and Broad sequencing centres separately (23 genes × 2 sites), but we also performed mega- and meta-analyses (23 genes × 2 methods).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 September 2011; accepted 6 March 2012.
Published online 4 April 2012.

1. Lichtenstein, P., Carlstrom, E., Rastam, M., Gillberg, C. & Anckarsater, H. The genetics of autism spectrum disorders and related neuropsychiatric disorders in childhood. *Am. J. Psychiatry* **167,** 1357–1363 (2010).
2. Hallmayer, J. *et al.* Genetic heritability and shared environmental factors among twin pairs with autism. *Arch. Gen. Psychiatry* **68,** 1095–1102 (2011).
3. Betancur, C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380,** 42–77 (2011).
4. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466,** 368–372 (2010).
5. Sanders, S. J. *et al.* Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70,** 863–885 (2011).
6. Sebat, J., Levy, D. L. & McCarthy, S. E. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet.* **25,** 528–535 (2009).
7. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
8. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43,** 491–498 (2011).
9. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
10. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature Genet.* **43,** 712–714 (2011).
11. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* http://dx.doi.org/10.1038/nature10945 (this issue).
12. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).
13. Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80,** 727–739 (2007).
14. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nature Rev. Genet.* **1,** 40–47 (2000).
15. Rossin, E. J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7,** e1001273 (2011).
16. Lage, K. *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl Acad. Sci. USA* **105,** 20870–20875 (2008).
17. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nature Genet.* **43,** 585–589 (2011).
18. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* http://dx.doi.org/10.1038/nature10989 (this issue).
19. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89,** 82–93 (2011).
20. Neale, B. M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* **7,** e1001322 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was directly supported by NIH grants R01MH089208 (M.J.D.), R01 MH089025 (J.D.B.), R01 MH089004 (G.D.S.), R01MH089175 (R.A.G.) and R01 MH089482 (J.S.S.), and supported in part by NIH grants P50 HD055751 (E.H.C.), R01 MH057881 (B.D.) and R01 MH061009 (J.S.S.). Y.K., G.C. and S.Y. are Seaver Fellows, supported by the Seaver Foundation. We thank T. Lehner, A. Felsenfeld and P. Bender for their support and contribution to the project. We thank S. Sanders and M. State for discussions on the interpretation of *de novo* events. We thank D. Reich for comments on the abstract and message of the manuscript. We thank E. Lander and D. Altshuler for comments on the manuscript. We acknowledge the assistance of M. Potter, A. McGrew and G. Crockett without whom these studies would not be possible, and Center for Human Genetics Research resources: Computational Genomics Core, Genetic Studies Ascertainment Core and DNA Resources core, supported in part by NIH NCRR grant UL1 RR024975, and the Vanderbilt Kennedy Center for Research on Human Development (P30 HD015052). This work was supported in part by R01MH084676 (S.S.). We acknowledge the clinicians and organizations that contributed to samples used in this study and the particular support of the Mount Sinai School of Medicine, University of Illinois-Chicago, Vanderbilt University, the Autism Genetics Resource Exchange and the institutions of the Boston Autism Consortium. We acknowledge A. Estes and G. Dawson for patient collection/characterization. We acknowledge partial support from U54 HG003273 (R.A.G.) and U54 HG003067 (E. Lander). J.D.B., B.D., M.J.D., R.A.G., A.S., G.D.S. and J.S.S. are lead investigators in the Autism Sequencing Consortium (ASC). The ASC is comprised of groups sharing massively parallel sequencing data in autism. Finally, we are grateful to the many families, without whose participation this project would not have been possible.

Author Contributions Laboratory work: A.S., C.St., G.C., O.J., Z.P., J.D.B., D.M., I.N., Y.W., L.L., Y.H., S.G., E.L.C., N.G.C. and E.T.G. Data processing: B.M.N., K.E.S., E.L., A.K., J.F., M.F., K.S., T.F., K.G., E.Ba., R.P., M.DeP., S.G., S.Y., V.M., J.L., J.D.B., A.S., C.St., U.N., J.G.R., J.R.W., B.E.B., S.E.L., C.F.L., L.S.W. and O.V. Statistical analysis: B.M.N., L.L., K.E.S., C.Sh., B.F.V., J.M., E.R., S.S., P.P., Y.K., A.M., R.D., C.-F.L., L.-S.W., H.L., T.Z., E.Bo., R.A.G., J.D.B., C.B., E.H.C., J.S.S., G.D.S., B.D., K.R. and M.J.D. Principal Investigators/study design: E.Bo., R.A.G., E.H.C., J.D.B., K.R., B.D., G.D.S., J.S.S. and M.J.D. Y.K., L.L., A.M., K.E.S., A.S. and C.-F.L. contributed equally to this work. E.Bo., J.D.B., E.H.C., B.D., R.A.G., K.R., G.D.S., J.S.S. and M.J.D. are lead investigators of the ARRA Autism Sequencing Collaboration.

Author Information Data included in this manuscript have been deposited at dbGaP under accession number phs000298.v1.p1 and is available for download at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000298.v1.p1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.J.D. (mjdaly@atgu.mgh.harvard.edu), J.D.B. (joseph.buxbaum@mssm.edu) or K.R. (kathryn.roeder@gmail.com).

METHODS

Phenotype assessment. Affected probands were assessed by research-reliable research personnel using Autism Diagnostic Interview-Revised (ADI-R), and the Autism Diagnostic Observation Schedule-Generic (ADOS) and DSM-IV diagnosis of a pervasive developmental disorder was made by a clinician. All probands met criteria for autism on the ADI-R and either autism or ASD on the ADOS, except for the three subjects from AGRE that were not assessed with the ADOS. In all, 85% of probands were classified with autism on both the ADI-R and ADOS. All subjects provided informed consent and the research was approved by institutional human subjects boards.

Exome sequencing, variant identification and *de novo* detection. Exome capture and sequencing was performed at each site using similar methods. Exons were captured using the Agilent 38 Mb SureSelect v2 (University of Pennsylvania and Broad Institute $n = 118$), the NimbleGen Seq Cap EZ SR v2 (Mt Sinai School of Medicine, Vanderbilt University $n = 51$), or NimbleGen VCRome 2.1 (Baylor $n = 6$). After capture, another round of LM-PCR was performed to increase the quantity of DNA available for sequencing. All libraries were sequenced using an IlluminaHiSeq2000.

Sequence processing and variant calling was performed using a similar computational workflow at all sites. Data were processed with Picard (http://picard.sourceforge.net/), which uses base quality-score recalibration and local realignment at known indels[8] and BWA[7] for mapping reads to hg19. SNPs were called using GATK[8,9] for all trios jointly. The variable sites that we have considered in analysis are restricted to those that pass GATK standard filters. From this set of variants, we identified putative *de novo* mutations as sites where both parents were homozygous for the reference sequence and the offspring was heterozygous and each genotype call was made confidently (see Supplementary Information).

Validation of *de novo* events. Putative *de novo* events were validated by sequencing the carrier and both parents using Sanger sequencing methods (University of Pennsylvania, Mt Sinai School of Medicine, Vanderbilt University, Baylor Medical College) or by Sequenom MALDI-TOF genotyping of trios (Broad).

Gene annotation. All identified mutations were then annotated using RefSeq hg19. The functional impact of variants was assessed for all isoforms of each gene, with the most severe annotation taking priority. Splice site variants were identified as occurring within two base pairs of any intron/exon boundary.

Expectation of *de novo* mutation calculation. To calculate the expected *de novo* rate, we assessed the mutability of all possible trinucleotide contexts in the intergenic region of the human genome for variation in two fashions: fixed genomic differences compared to chimpanzee and baboon[12] and variation identified from the 1,000 Genomes project. The overall mutation rate for the exome was then determined by summing the probability of mutation for all bases in the exome that were captured successfully. We also determined the probability of each class functional mutation by summing the annotated variants.

Pathway analyses. We applied DAPPLE[15], which uses the InWeb database[16], to determine whether there is excess protein–protein interaction across the genes hit by a functional *de novo* event. We also assessed whether these genes were more closely connected to a list of ASD genes[3].

Modelling *de novo* events. We modelled a Poisson process consistent with the expected distribution defined by the mutation model and with the observed data. We varied the fraction of genes that influence risk, the probability a variant in a gene would be functional, and the penetrance of functional *de novo* events. We also simulated a random set of *de novo* events to estimate the probability of hitting a gene multiple times.

Association analysis. We performed association tests using SKAT[19], a generalization of C-alpha[20]. Our primary analyses treat case–control data generated at Baylor and Broad sequencing centres separately (23 genes × 2 sites), but we also performed mega- and meta-analyses (23 genes × 2 methods).

Samocha et al *Nature Genetics* 2014

Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnström K, Mallick S, Kirby A, Wall DP, MacArthur DG, Gabriel SB, dePristo M, Purcell SM, Palotie A, Boerwinkle E, Buxbuam JD, Cook EH, Gibbs RA, Schellenberg GD, Sutcliffe JS, Devlin B, Roeder K, Neale BM, Daly MJ. Leveraging a model of de novo variation to evaluate exome sequencing data. *Nat Genet*. 2014 Sep;46(9):944-50. doi: 10.1038/ng.3050. Epub 2014 Aug 3.

A framework for the interpretation of *de novo* mutation in human disease

Kaitlin E Samocha[1–4], Elise B Robinson[1–3], Stephan J Sanders[5,6], Christine Stevens[2,3], Aniko Sabo[7], Lauren M McGrath[8], Jack A Kosmicki[1,9,10], Karola Rehnström[11,12], Swapan Mallick[13], Andrew Kirby[1,2], Dennis P Wall[9,10], Daniel G MacArthur[1,2], Stacey B Gabriel[2], Mark DePristo[14], Shaun M Purcell[1,2,8,15–17], Aarno Palotie[8,11,12], Eric Boerwinkle[7,18], Joseph D Buxbaum[15–17,19–21], Edwin H Cook Jr[22], Richard A Gibbs[7], Gerard D Schellenberg[23], James S Sutcliffe[24], Bernie Devlin[25], Kathryn Roeder[26,27], Benjamin M Neale[1–3] & Mark J Daly[1–3]

Spontaneously arising (*de novo*) mutations have an important role in medical genetics. For diseases with extensive locus heterogeneity, such as autism spectrum disorders (ASDs), the signal from *de novo* mutations is distributed across many genes, making it difficult to distinguish disease-relevant mutations from background variation. Here we provide a statistical framework for the analysis of excesses in *de novo* mutation per gene and gene set by calibrating a model of *de novo* mutation. We applied this framework to *de novo* mutations collected from 1,078 ASD family trios, and, whereas we affirmed a significant role for loss-of-function mutations, we found no excess of *de novo* loss-of-function mutations in cases with IQ above 100, suggesting that the role of *de novo* mutations in ASDs might reside in fundamental neurodevelopmental processes. We also used our model to identify ~1,000 genes that are significantly lacking in functional coding variation in non-ASD samples and are enriched for *de novo* loss-of-function mutations identified in ASD cases.

Exome sequencing has enabled the identification of *de novo* (newly arising) mutations and has already been effectively used to identify causal variants in rare mendelian diseases. In the case of Kabuki syndrome, the observation of a *de novo* mutation in *MLL2* (*KMT2D*) in 9 of the 10 cases analyzed strongly implicated loss of *MLL2* function as causal[1]. The conclusion that *MLL2* is important in Kabuki syndrome etiology based on the *de novo* mutation findings relies upon the unlikely accumulation of independent and infrequently occurring events in the vast majority of these unrelated cases. By contrast, *de novo* mutations have a smaller role in the pathogenesis of heritable complex traits, such as ASDs, and associated *de novo* mutations are spread across multiple genes. These differences in the etiologic architecture of complex traits make the task of identifying 'causal' genes considerably more challenging. For example, recent exome sequencing studies demonstrated a significant excess of *de novo* loss-of-function mutations in ASD cases but lacked the ability to directly implicate more than a very small number of genes[2–6].

The main complicating factor for interpreting the number of observed *de novo* mutations for a particular gene is the background rate of *de novo* mutation, which can vary greatly between genes. As more individuals are sequenced, multiple *de novo* mutations will inevitably be observed in the same gene by chance. However, if *de novo* mutation has a role in a given disease, we would expect to find

[1]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA. [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [3]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [4]Program in Genetics and Genomics, Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA. [5]Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut, USA. [6]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut, USA. [7]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. [8]Psychiatric and Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA. [9]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. [10]Department of Pathology, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA. [11]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [12]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK. [13]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [14]Synapdx, Lexington, Massachusetts, USA. [15]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [16]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [17]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [18]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas, USA. [19]Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [20]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [21]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [22]Department of Psychiatry, University of Illinois at Chicago, Chicago, Illinois, USA. [23]Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [24]Center for Molecular Neuroscience, Vanderbilt University, Nashville, Tennessee, USA. [25]Department of Psychiatry, University of Pittsburgh Medical School, Pittsburgh, Pennsylvania, USA. [26]Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. [27]Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. Correspondence should be addressed to M.J.D. (mjdaly@atgu.mgh.harvard.edu).

Received 10 December 2013; accepted 9 July 2014; published online 3 August 2014; doi:10.1038/ng.3050

that genes associated with the disease would contain more *de novo* mutations than expected by chance.

Here we develop a statistical model of *de novo* mutation to evaluate the findings from exome sequencing data. With this model, we establish a statistical framework to evaluate the rates of *de novo* mutation, not only on a per-gene basis (in a frequentist manner analogous to that used in common genome-wide association analysis) but also globally and by gene set. We further use this model to predict the expected amount of rare standing variation per gene and to detect those genes that are significantly and specifically deficient in functional variation, likely reflecting processes of selective constraint. Consequently, as selection has reduced standing functional variation in these genes, it is reasonable to hypothesize that mutations in these genes are more likely to be deleterious.

We used the mutational model along with our list of highly constrained genes to evaluate the relationship between *de novo* mutation and ASDs. Most of the families in these analyses were also included in a set of previous studies of *de novo* mutation, which reported an overall excess of *de novo* loss-of-function mutations in ASD cases, as well as multiple *de novo* mutations in specific genes[2–5]. We build on those studies to examine the aggregate rates of *de novo* mutation, the excess of multiply mutated genes and the overlap of *de novo* mutations with gene sets, which highlights the complex relationship between intellectual functioning and the genetic architecture of ASDs.

RESULTS

Basis of the mutational model

Accurate estimation of the expected rate of *de novo* mutation in a gene requires a precise estimate of each gene's mutability. Although gene length is an obvious factor in a gene's mutability, local sequence context is also a well-known source of differences in mutation rate[7]. Accordingly, we extended a previous model of *de novo* mutation based on sequence context and developed gene-specific probabilities for different types of mutation: synonymous, missense, nonsense, essential splice site and frameshift (Online Methods, **Supplementary Fig. 1** and **Supplementary Table 1**)[3]. Underscoring the importance of the sequence context factors in the model, this genome-wide rate yields an expected mutation rate of 1.67×10^{-8} mutations per base per generation for the exome alone. Using counts of rare (minor allele frequency < 0.001) synonymous variants identified in the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP), we found that our per-gene probabilities of mutation were significantly more correlated ($r = 0.940$) with these counts than with gene length alone ($P < 1 \times 10^{-16}$; Online Methods).

Having established accurate per-gene probabilities of mutation, we could then investigate the rates and distribution of *de novo* mutations found in sequencing studies. Specifically, we wished to systematically assess (i) whether cases had genome-wide excesses of certain functional categories of *de novo* mutation; (ii) whether individual genes could be associated via *de novo* mutation with genome-wide statistical significance; (iii) whether specific sets of genes collectively showed significant enrichment of *de novo* mutations; and (iv) whether there were genome-wide excesses of genes with multiple *de novo* mutations. Below we demonstrate the usefulness of the statistical framework in addressing all of these questions with respect to recently generated family exome sequencing data for autism and intellectual disability.

Identifying genes under selective constraint

There has been a long-standing interest in identifying genes in the human genome that are sensitive to mutational changes, as these genes would be the most likely to contribute to disease. Recent work made use of ESP data to create a metric evaluating the proportion of common functional variation in each gene, thereby identifying genes that appeared to be intolerant of mutation[8]. Along these lines, we correlated our calculated per-gene probabilities of mutation with the observed counts of rare missense variants in the ESP data set. In contrast to the high consistency between predicted synonymous mutation rates and observed synonymous counts (expected if the category is under no specific selection), we observed a significant number of genes with a severe deficit in missense variants compared to the expectation generated from predicted mutation rates ($P < 1 \times 10^{-16}$). Such a deficit is consistent with strong evolutionary constraint: when damaging mutations arise, they are quickly removed from the population by purifying selection. To avoid erroneously identified constrained genes, we removed 134 genes with either significantly elevated or decreased synonymous and nonsynonymous rates (both $P < 0.001$; Online Methods).

Comparing both the synonymous and missense mutation predictions of our model to the ESP data set, we identified a list of excessively constrained genes (missense Z score > 3.09; corresponding to $P < 0.001$) that represented roughly 5% of all genes (**Supplementary Table 2**). A high proportion of the most significantly constrained genes (missense constraint $P < 1 \times 10^{-6}$) were associated with autosomal or

Table 1 Evaluation of the rates of *de novo* mutation in ASD cases and unaffected siblings

Genome-wide excesses of mutational events

	Unaffected siblings (n = 343 families)			ASD cases (n = 1,078 families)		
Mutation type	Observed events per exome	Expected events per exome	P value	Observed events per exome	Expected events per exome	P value
Synonymous	0.21	0.27	0.0218[a]	0.25	0.27	0.1065[a]
Missense	0.61	0.62	0.8189[a]	0.64	0.62	0.5721[b]
Loss of function	0.09	0.09	0.4508[b]	0.13	0.09	**2.05×10^{-7}**[a]

Genome-wide excesses of multiply mutated genes

	Unaffected siblings (n = 343 families)			ASD cases (n = 1,078 families)		
Mutation type	Observed genes with ≥2 DNMs	Average expected genes with ≥2 DNMs	P value	Observed genes with ≥2 DNMs	Average expected genes with ≥2 DNMs	P value
Synonymous	0	0.49	1.0	4	3.8	0.5186
Missense	5	2.5	0.1049	33	21	**0.0070**
Loss of function	0	0.039	1.0	6	0.5	**<0.001**
Loss of function + missense	6	3.0	0.0779	48	27	**<0.001**

The top half of the table shows the observed and expected rates of mutation by type per exome for unaffected siblings[2] and ASD cases, including some unpublished US and Finnish trios[2–6]. The bottom half of the table shows the number of genes with multiple *de novo* mutations in unaffected siblings and ASD cases across studies. The average number of expected genes with multiple *de novo* mutations was determined by simulation. DNMs, *de novo* mutations. Significant P values are shown in bold.
[a]Two-tailed. [b]One-tailed.

Table 2 Individually significant genes identified from the analysis of *de novo* mutations in ASD cases

Gene	Mutations	Number of observed loss-of-function mutations	Number of expected loss-of-function mutations	*P* value
DYRK1A	Nonsense, splice site, frameshift	3	0.0072	**6.15×10^{-8}**
SCN2A	Nonsense, nonsense, frameshift	3	0.018	**9.20×10^{-7}**
CHD8	Nonsense, splice site, frameshift	3	0.022	1.76×10^{-6}
KATNAL2	Splice site, splice site	2	0.0049	1.19×10^{-5}
POGZ	Frameshift, frameshift	2	0.013	8.93×10^{-5}
ARID1B	Frameshift, frameshift	2	0.018	1.57×10^{-4}

Shown are genes with multiple *de novo* loss-of-function mutations across 1,078 ASD cases. Loss-of-function mutations include nonsense, frameshift and splice site–disrupting mutations. Number of expected loss-of-function mutations refers to the expected number of *de novo* loss-of-function mutations based on the probability of mutation for the gene as determined by our model. The genome-wide significance threshold is 1×10^{-6}. Significant *P* values are shown in bold.

X-linked dominant, largely sporadic mendelian disease entries in the Online Mendelian Inheritance in Man database (OMIM; $n = 27/86$). By contrast, a set of genes for which the missense constraint was very close to the expectation ($n = 111$; $-0.01 < Z < 0.01$) had only 2 *de novo* or dominant disease inheritance entries in OMIM, a number significantly different from that for the highly constrained set ($P < 1 \times 10^{-8}$). For the 86 most highly constrained genes, no autosomal recessive mendelian disorders have been documented. However, 11 of the 111 genes with average levels of constraint have been identified as causal in autosomal recessive mendelian disorders. The significant excess of recessive disease-causing genes in the middle part of the distribution in comparison to the constrained set ($P < 0.003$) underscores the idea that recessive inheritance models do not induce strong constraint.

Mutation rates for ASDs and intellectual disability

We applied the model to two primary data sets: published results from ASD sequencing studies[2–6] with a collection of additional unpublished ASD family trios and published results from individuals with severe intellectual disability[9,10]. Comparisons of the predicted number of mutations per exome and the observed data from the 1,078 ASD cases as well as the 343 sequenced unaffected siblings[2–6] are shown in **Table 1**. The model's predictions matched the observed data for the unaffected siblings well, but the cases showed a significant excess of *de novo* loss-of-function mutations ($P = 2.05 \times 10^{-7}$), consistent with the findings of the individual sequencing studies. Using our model to simulate null *de novo* mutation sets, we found that there were significantly more genes with two or more *de novo* loss-of-function mutations than would be expected by chance ($P < 0.001$; six observed when less than one was expected; **Supplementary Table 3**). Notably, although we did not observe a global excess of *de novo* missense mutations, we did observe an excess of genes with 2 or more functional (loss-of-function or missense) *de novo* mutations (48 such genes were observed when the average number expected was 27; $P < 0.001$) and genes with 2 or more *de novo* missense mutations alone (33 such genes were observed when the average number expected was 21; $P = 0.007$ for missense variants; **Table 1**). No such excess of genes containing multiple *de novo* mutations was seen in the unaffected siblings (**Table 1**). Of note, our framework also supports the assessment of many other weightings and combinations of alleles—such as missense variants only (optimal for pure gain-of-function disease models), predicted damaging missense variants only and exact probability estimates for specific combinations of loss-of-function and missense variants—beyond those shown above.

Some of the genes that had 2 or more *de novo* loss-of-function mutations across the 1,078 subjects with ASD are listed in **Table 2**. The results for all genes can be found in **Supplementary Table 4**. A conservative significance threshold of $P = 1 \times 10^{-6}$ was used, correcting for 18,271 genes and 2 tests. Considering this set of 1,078 trios as a single experiment, 2 genes (*DYRK1A* and *SCN2A*) exceeded this conservative genome-wide significance threshold for more *de novo* loss-of-function mutations than predicted. *SCN2A* also had significantly more functional *de novo* mutations than expected. *CHD8*, with three *de novo* loss-of-function mutations and one missense mutation, was very close to the significance threshold in these studies ($P = 1.76 \times 10^{-6}$ for loss-of-function mutations; $P = 3.20 \times 10^{-5}$ for functional mutations). However, a recent targeted sequencing study found 7 additional *de novo* loss-of-function mutations in *CHD8* in ASD cases[11], bringing the total number of *de novo* loss-of-function mutations in *CHD8* to 10, a number that was highly significant ($P = 8.38 \times 10^{-20}$ when accounting for the total number of trios ($n = 2{,}750$) examined in the combination of the targeted and exome-wide studies). These results offer the encouraging point that, as with genome-wide association studies (GWAS), larger collaborative exome sequencing efforts for trios will define unambiguous risk factors. It is important to note, however, that not all genes with a large number of *de novo* mutations had significant *P* values. For example, *TTN* had four missense *de novo* mutations in ASD cases but had a *P* value that was not even nominally significant ($P = 0.18$), owing to the enormous size of the gene. Even having two *de novo* loss-of-function mutations was on occasion not enough to provide compelling significance (*POGZ*; two frameshift mutations; $P = 8.93 \times 10^{-5}$). In comparison, none of the genes found to contain multiple *de novo* mutations in the unaffected siblings crossed the significance threshold (**Supplementary Table 5**).

These analyses were also applied to the results from the sequencing studies of moderate to severe (IQ < 60) intellectual disability[9,10].

Table 3 Evaluation of the rates of *de novo* mutation in cases with intellectual disability

Genome-wide excesses of mutational events

Mutation type	Intellectual disability cases: Observed events per exome	Intellectual disability cases: Expected events per exome	*P* value
Synonymous	0.19	0.27	0.0267[a]
Missense	0.70	0.62	0.2380[a]
Loss of function	0.24	0.09	**6.49×10^{-7}**[b]

Genome-wide excesses of multiply mutated genes

Mutation type	Intellectual disability cases: Observed genes with ≥2 DNMs	Intellectual disability cases: Average expected genes with ≥2 DNMs	*P* value
Synonymous	1	0.092	0.0879
Missense	3	0.47	**0.0090**
Loss of function	2	0.011	**<0.001**
Loss of function + missense	6	0.60	**<0.001**

The top half of the table shows the observed and expected rates of mutation by type per exome for cases of intellectual disability ($n = 151$ families)[9,10]. The bottom half of the table shows the number of genes with multiple *de novo* mutations in intellectual disability cases across studies. The average number of expected genes with multiple *de novo* mutations was determined by simulation. DNMs, *de novo* mutations. Significant *P* values are shown in bold. [a]Two-tailed. [b]One-tailed.

Table 4 Individually significant genes identified from the analysis of *de novo* mutations in individuals with intellectual disability

Gene	Mutations	Number of loss-of-function mutations	Number of missense mutations	Number of DNMs expected	*P* value	Test
SYNGAP1	Splice site, frameshift, frameshift	3	0	0.0017	**8.15×10^{-10}**	Loss of function
SCN2A	Missense, nonsense, frameshift, frameshift	3	1	0.0025	**2.56×10^{-9}**	Loss of function
SCN2A	Missense, nonsense, frameshift, frameshift	3	1	0.019	**5.01×10^{-9}**	Loss of function + missense
STXBP1	Missense, missense, splice site	1	2	0.0071	**5.87×10^{-8}**	Loss of function + missense
TCF4	Missense, missense	0	2	0.0069	2.39×10^{-5}	Loss of function + missense
GRIN2A	Missense, missense	0	2	0.016	1.34×10^{-4}	Loss of function + missense
TRIO	Missense, missense	0	2	0.033	5.60×10^{-4}	Loss of function + missense

Shown are genes with multiple functional *de novo* mutations across 151 cases of intellectual disability[9,10]. Loss-of-function mutations include nonsense, frameshift and splice site–disrupting mutations. The genome-wide significance threshold is 1×10^{-6}. The number of mutations is either compared to the expected number for loss-of-function mutations only or for both loss-of-function and missense mutations, as indicated by the number of DNMs expected and test columns. Significant *P* values are shown in bold.

Intellectual disability, like ASD, showed a significant excess of *de novo* loss-of-function mutations ($P = 6.49 \times 10^{-7}$; **Table 3**). Even with a much smaller sample size ($n = 151$), there were genes with significantly more loss-of-function and functional *de novo* mutations than predicted by the model (**Table 4**). The data for intellectual disability also showed significantly more genes with multiple missense, loss-of-function and functional *de novo* mutations than predicted ($P = 0.009$ for missense mutations; $P < 0.001$ for loss-of-function and functional mutations).

In our ASD sample, we then investigated the rate of *de novo* events as a function of IQ; roughly 80% of this sample had an IQ assessment attempted. We found that the rate of *de novo* loss-of-function mutation in ASD cases with a measured IQ above average was no different than the expectation (IQ ≥ 100; $n = 229$; 0.08 *de novo* loss-of-function mutations per exome in comparison to the expectation of 0.09; $P = 0.59$). By contrast, the rate in the rest of the sample was substantially higher than the expectation ($n = 572$; rate of 0.17 *de novo* loss-of-function mutations per exome; $P = 1.17 \times 10^{-10}$). Furthermore, when directly compared (rather than being compared to our expectation), these two groups were significantly different from each other ($P < 0.001$), confirming a difference in genetic architecture among ASDs as a function of IQ (**Supplementary Table 6**). These conclusions were unchanged in separate analyses of nonverbal and verbal IQ as well as full-scale IQ (**Supplementary Table 6**).

Gene set enrichment

Given the significant global excess of *de novo* loss-of-function mutations in ASD cases, we wanted to evaluate whether the set of genes harboring *de novo* loss-of-function mutations had significant overlap with several sets of genes proposed to be relevant to autism or describing biochemical pathways. We used the probabilities of mutation to determine the fraction of loss-of-function mutations expected to fall into the given gene set. We then used the binomial distribution to evaluate the number of observed loss-of-function mutations overlapping with the set in comparison to the established expectation. When we applied this analysis to a set of 112 genes reported to be disrupted in individuals with ASDs or autistic features, we observed no enrichment of *de novo* loss-of-function mutations (**Fig. 1**, Betancur)[12]. By contrast, we applied this analysis to a recent study of 842 genes found to interact with the fragile X mental retardation protein (FMRP) *in vivo* and found a highly significant overlap (2.3-fold enrichment; $P < 0.0001$; **Fig. 1**)[2,13]. This enrichment with the targets of FMRP held even when we removed the *de novo* mutations identified in the study by Iossifov *et al.*[2], which initially reported an enrichment of *de novo* mutations in ASD cases in FMRP-associated genes (2.5-fold enrichment; $P < 0.0001$).

We then evaluated the group of individuals from the ASD studies who had a *de novo* loss-of-function event in one of the targets of FMRP. On average, these cases were enriched for having a measured IQ of <100 (Fisher's exact test $P = 4.01 \times 10^{-4}$; **Supplementary Table 7**) as well as a significantly reduced male/female ratio ($P = 0.02$; **Supplementary Table 8**) as compared to the remaining sequenced cases (**Supplementary Note**). These individuals represented about 3% of the total sample, when, at most, a 1% overlap would be expected. The estimated odds ratio (OR) of *de novo* loss-of-function events in the set of FMRP target genes was around 6, very similar to the ORs estimated for large copy number variants (CNVs) that disrupt multiple genes[14]. In addition, the OR for the published cases of moderate to severe intellectual disability noted above (IQ < 60; not ascertained for ASDs) having a *de novo* loss-of-function event in the set of FMRP targets was roughly 10.

The same analysis was applied to the list of *de novo* loss-of-function events from the unaffected siblings of ASD cases and additional control individuals ($n = 647$)[2,4,5,15]. There was a significant enrichment when evaluating overlap with the set of autism-related genes ($P = 0.0095$; **Fig. 1**). However, no significance was observed for overlap with the *in vivo* targets of FMRP. The list of *de novo* loss-of-function mutations from the individuals with intellectual disability, on the other hand, was significant for both sets ($P < 1 \times 10^{-4}$ for both sets; **Supplementary Fig. 2**). Even the *de novo* missense mutations found in the intellectual disability cases showed significant overlap with both

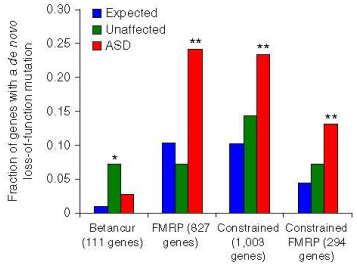


Figure 1 The expected and observed fraction of genes with a *de novo* loss-of-function mutation in ASD cases and unaffected controls for four gene sets of interest. ASD cases ($n = 1,078$) and unaffected controls ($n = 647$) were sequenced across various studies (refs. 2–6,10,15). "Betancur" refers to a set of genes reported to be disrupted in individuals with ASDs or autistic features; of the 112 on the list[12], we could evaluate 111. "FMRP" refers to the genes whose mRNAs are bound and regulated by the fragile X mental retardation protein, as identified by Darnell *et al.*[13]. The "constrained" category is a set of 1,003 genes that we defined as significantly lacking rare missense variation, indicating intolerance to mutation. The targets of FMRP that are also considered constrained by our metric make up the "constrained FMRP" category. $^*P < 0.01$, $^{**}P < 1 \times 10^{-4}$, binomial test.

sets under study ($P = 0.02$ for autism-related genes and $P < 0.0001$ for the targets of FMRP; **Supplementary Fig. 2**).

Evaluating constrained genes

We further applied the enrichment analysis to our set of constrained genes and found that they contained more *de novo* loss-of-function mutations than expected by chance (2.3-fold enrichment; $P < 0.0001$; **Fig. 1**). We observed a greater fold enrichment when focusing on the subset of constrained genes that were also identified in the FMRP study (3.0-fold enrichment; $P < 0.0001$; **Fig. 1**)[13]. We note that the FMRP targets showed significant overlap with the constrained set of genes (OR = 1.29; $P < 0.0001$), which is consistent with the report that the targets of FMRP are under greater purifying selection than expected[2]. All enrichments were demonstrated to be independent of gene length (**Supplementary Note**).

The genes that contained a *de novo* missense or loss-of-function mutation in the intellectual disability cases also showed a significant enrichment for both the constrained gene set and the set of constrained targets of FMRP ($P < 0.0001$ for all lists). In comparison, no enrichment was found with either set for the list of genes that had a *de novo* loss-of-function mutation in unaffected siblings and control individuals.

In addition to treating constraint as a dichotomous trait, we also evaluated the missense *Z* score for each of the genes with a *de novo* loss-of-function mutation. We found that the distribution of missense *Z* scores for genes with a *de novo* loss-of-function mutation in unaffected individuals was no different than the overall distribution of scores (Wilcoxon $P = 0.8325$; **Fig. 2**). By contrast, both the genes with a *de novo* loss-of-function mutation in ASD and intellectual disability cases had values significantly shifted toward high constraint (Wilcoxon $P < 1 \times 10^{-6}$ for both). Furthermore, we compared the distribution of *Z* scores among each of the three groups. Both the ASD and intellectual disability distributions were significantly different from the distribution of missense *Z* scores for unaffected individuals ($P = 0.0148$ and 0.0012, respectively). The intellectual disability missense *Z* scores were also significantly higher than the corresponding ASD values ($P = 0.0319$).

When evaluating the ASD cases split by IQ group, we found no enrichment of genes with *de novo* loss-of-function mutations with either constrained genes or targets of FMRP in the group with IQ of ≥100 ($P > 0.5$ for both sets of genes), but we found very strong enrichment in the set with IQ of <100 ($P < 0.0001$ for both sets of genes). These results underscore the idea that phenotypically distinct subsets of ASD cases may have significantly different contributions from *de novo* mutation.

Comparison of constraint metric with existing methods

Identifying constrained genes by comparing observed nonsynonymous sites to the expectation is conceptually similar to the traditional approach of detecting selective pressure by comparing observed nonsynonymous sites to observed synonymous sites (for example, d_N/d_S) that has been used extensively. Our approach should in principle achieve greater statistical power to detect constrained genes; comparison of an observation to an expectation is statistically more powerful than contrasting that observation with a generally smaller second observation (the number of observed synonymous variants). To investigate this claim, we identified genes that had significant evidence for selective constraint using the d_N/d_S metric (their ratio of synonymous to nonsynonymous sites deviated from the genome-wide average at $P < 0.001$; **Supplementary Note**). There were only 377 of these genes, over half of which overlapped with the constrained gene list defined by our method ($n = 1{,}003$; overlap of 237 genes). The genes identified as significantly constrained by only our metric (the top 10 of which included *RYR2*, *MLL* (*KMT2A*), *MLL2* and *SYNGAP1*) were still significantly enriched for known causes of autosomal and X-linked dominant forms of mendelian disease ($P = 5 \times 10^{-4}$). We therefore conclude that the model-based approach to identifying constrained genes adds substantial power to traditional approaches. The importance of this increased power to detect constraint is further articulated in the ASD and intellectual disability analyses below.

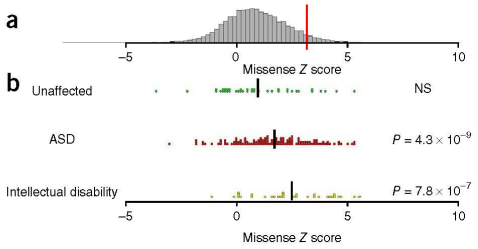


Figure 2 Distributions of missense *Z* scores and *Z* scores for genes containing *de novo* loss-of-function mutations identified in unaffected individuals, ASD cases and intellectual disability cases. (**a**) Distribution of missense *Z* scores. The red bar indicates a *Z* score of 3.09, or the threshold for inclusion in the set of 1,003 constrained genes. (**b**) Missense *Z* scores for genes containing *de novo* loss-of-function mutations in unaffected individuals, ASD cases and intellectual disability cases[2–6,9,10,15]. Black bars indicate the mean *Z* score of each group: 0.94, 1.68 and 2.46 for unaffected individuals, ASD cases and intellectual disability cases, respectively. Although the missense *Z* scores of the *de novo* loss-of-function mutations found in unaffected siblings matched the overall distribution (Wilcoxon $P = 0.8325$; NS, not significant), *de novo* loss-of-function mutations found in both ASD and intellectual disability cases were significantly shifted toward more extreme constraint values ($P < 1 \times 10^{-6}$ for both). All *P* values for deviation from the overall distribution are listed on the right side of the figure. In addition, the distributions of missense *Z* scores for each of the three *de novo* lists were all individually significant at $P < 0.05$.

Several groups have previously published approaches and specific gene sets from these that are also aimed at identifying genes under excessive purifying selection or generally intolerant of functional mutation. Bustamante *et al.*[16] expanded on the McDonald-Kreitman framework[17], contrasting fixed differences in the primate lineage to polymorphic differences in humans to identify a set of genes under weak negative selection, while more recently Petrovski *et al.*[8] used the excess of rare versus common missense variation within humans to flag genes intolerant of functional variation. We found a reasonable correlation between our metric of constraint and the residual variation intolerance score (RVIS) of Petrovski *et al.* (**Supplementary Fig. 3**)[8]. A comparison of these approaches as applied to the prioritization of known haploinsufficient genes, as well as to the *de novo* loss-of-function mutations in autism described here, is provided in the **Supplementary Note** and demonstrates that the two human-only approaches (constraint and RVIS) perform better on these tasks of identifying medical genetics lesions of severe effect in modern humans (**Supplementary Table 9**). Intriguingly, both of these other approaches use independent information from each other and from our approach (which uses the absence of rare functional variation in comparison to the expectation within humans), raising the possibility that composite scores employing all three sources of information could add further value in highlighting which genes are most sensitive to heterozygous mutation.

DISCUSSION

We have developed a framework for evaluating excesses of *de novo* mutations identified through exome sequencing. Even though this framework can be leveraged to evaluate excesses of mutation across a study and in gene sets, the key focus is on evaluating the significance for individual genes. Given the small number of observed *de novo* events per gene, simple case-control comparisons cannot achieve any meaningful level of significance. For example, observing 3 *de novo* loss-of-function mutations in a small gene in 1,000 case trios is perhaps quite compelling, especially if no such mutations were identified in 1,000 control trios. However, a simple three-to-zero case-control comparison in this situation would yield no compelling statistical evidence (one-tailed $P = 0.125$). Incidence of such extremely rare events, however, can be evaluated if the expected rate of such events is known. Sequencing large numbers of control trios to gather empirical rate estimates on a per-gene basis that are accurate is infeasible and inefficient. The calibrated model and statistical approach described here can achieve a close approximation of this ideal. Our method, therefore, offers the ability to evaluate the rate of rare variation in individual genes in situations where burden tests would fail.

Other groups have developed similar statistical frameworks[11,18]; notably, the Epi4K Consortium[18] used the same base model we began with[3] to interpret event rates. Our model, however, has two primary strengths. First, our model of *de novo* mutation incorporates additional factors beyond sequence context that affect mutation rate. Both the depth of coverage (how many sequence reads were present on average) for each base and the regional divergence around the gene between humans and macaques independently and significantly improve the predictive value of our model (**Supplementary Note**). Second, given the high correlation between the number of rare synonymous variants in ESP and the probability of a synonymous mutation determined by our full model, we have a metric to evaluate the extent to which genes in the human genome show evidence of selective constraint. The list of 1,003 genes that we define as constrained contains an enrichment of genes known to cause severe human disease—an observation analogous to that recently made in using empirical comparison of common and rare rates of functional variation to evaluate intolerance to mutation[8]. In fact, site count deficits and shifts in site frequency each contribute independent information to the definition of constraint and can in principle be combined in a composite test.

The results of our metric were compared to both the scores created by Petrovski *et al.*[8] and the loci identified as being under negative selection by Bustamante *et al.*[16]. Overall, our metric and the RVIS metric defined by Petrovski *et al.* worked similarly well, reinforcing the benefits that could come from combining the two approaches. It is unsurprising that these methods outperform the evolutionary ones on the specific matter of genes intolerant to heterozygous mutation. Evolutionary methods examining differences between polymorphism and fixed differences, which are more sensitive to weaker negative selection, require that mutations be tolerated well enough to become polymorphic in the first place. By contrast, approaches measuring the complete absence of variation will pick up the most strongly intolerant genes.

Ideally, we can conceptualize defining two metrics of genic constraint, one based on missense variants and the other based on loss-of-function variants. With only 6,503 individuals in ESP, we are underpowered to determine significant deviations for most genes with respect to loss-of-function variants. As sample size increases, our ability to calculate constraint improves. For example, if the sample size were to increase by an order of magnitude, we would be able to evaluate approximately 66% of genes using loss-of-function variants. We therefore view the constrained gene list as a work in progress, to be updated when larger exome sequencing data sets become available.

Applying our statistical framework to *de novo* mutations from 1,078 ASD cases shows that, although there is no global excess in *de novo* missense mutations, there are significantly more genes that contain multiple *de novo* missense mutations than expected. We also see significant overlap between the list of genes with a *de novo* loss-of-function mutation in ASD cases and the set of constrained genes that we defined. In addition, there is significant overlap between the genes with a *de novo* loss-of-function mutation and the targets of FMRP, as reported in Iossifov *et al.*[2]. All of the significant signals in ASD—the global excess of *de novo* loss-of-function mutations, the excess of genes with multiple functional *de novo* mutations, the overlap between the genes with *de novo* loss-of-function mutation and both constrained genes and the targets of FMRP—are not found in the subset of ASD cases with IQ of ≥100. The lack of signal in this subset indicates that genetic architecture among ASDs varies as a function of IQ. Overall, the probabilities of mutation defined by our full model and list of constrained genes can be used to critically evaluate the observed *de novo* mutations from sequencing studies and to aid in the identification of variants and genes that have a critical role in disease.

URLs. Online Mendelian Inheritance in Man (OMIM), http://omim.org/; Exome Variant Server, http://evs.gs.washington.edu/EVS/; site to query constraint information and *de novo* mutations from published studies, http://atgu.mgh.harvard.edu/webtools/gene-lookup/; Picard, http://picard.sourceforge.net/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. New data included in this manuscript have been deposited in the database of Genotypes and Phenotypes (dbGaP), merged with our published data under accession phs000298.v1.p1.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

All data from published studies are available in the respective publications. All newly generated data and computational tools used in this paper will be available online as downloadable material. We have also constructed a website to query genes that provides information on constraint and the *de novo* mutations found in the specified gene across published studies of *de novo* mutation. We would like to thank E. Daly and M. Chess for their contributions to data analysis and the construction of the website, respectively. We acknowledge the following resources and families who contributed to them: the National Institute of Mental Health (NIMH) repository (U24MH068457); the Autism Genetic Resource Exchange (AGRE) Consortium, a program of Autism Speaks (1U24MH081810 to C.M. Lajonchere); The Autism Simplex Collection (TASC) (grant from Autism Speaks); the Simons Foundation Autism Research Initiative (SFARI) Simplex Collection (grant from the Simons Foundation); and The Autism Consortium (grant from the Autism Consortium). This work was directly supported by US National Institutes of Health (NIH) grants R01MH089208 (M.J.D.), R01MH089025 (J.D.B.), R01MH089004 (G.D.S.), R01MH089175 (R.A.G.) and R01MH089482 (J.S.S.) and was supported in part by US NIH grants P50HD055751 (E.H.C.), R01MH057881 (B.D.) and R01MH061009 (J.S.S.). We acknowledge partial support from grants U54HG003273 (R.A.G.) and U54HG003067 (E. Lander). We thank T. Lehner (NIMH), A. Felsenfeld (National Human Genome Research Institute) and P. Bender (NIMH) for their support and contribution to the project. E.B., J.D.B., B.D., M.J.D., R.A.G., K. Roeder, A.S., G.D.S. and J.S.S. are lead investigators in the ARRA Autism Sequencing Collaboration (AASC). We would also like to thank the NHLBI GO Exome Sequencing Project (ESP) and its ongoing studies that produced and provided exome variant calls on the web: the Lung GO Sequencing Project (HL-102923), the

Women's Health Initiative (WHI) Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010).

AUTHOR CONTRIBUTIONS

K.E.S., B.M.N. and M.J.D. conceived and designed the mutational model and constraint methods. K.E.S. and E.B.R. executed the analyses. K.E.S., E.B.R., L.M.M., J.A.K., S.M., A.K., D.P.W., D.G.M., S.M.P., J.D.B., B.D. and K. Roeder contributed to analysis concepts and methods. K.E.S., S.J.S., C.S., A.S., K. Rehnström, S.B.G., M.D., A.P., E.B., J.D.B., E.H.C., R.A.G., G.D.S., J.S.S., B.D., K. Roeder, B.M.N. and M.J.D. contributed autism sequencing, evaluation and manuscript comments. K.E.S., E.B.R., B.M.N. and M.J.D. performed the primary writing.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Ng, S.B. *et al.* Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790–793 (2010).
2. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
3. Neale, B.M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
4. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).
5. Sanders, S.J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
6. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nat. Genet.* **43**, 585–589 (2011).
7. Antonarakis, S.E. CpG dinucleotides and human disorders. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Chichester, UK, 2006).
8. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
9. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
10. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
11. O'Roak, B.J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–1622 (2012).
12. Betancur, C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380**, 42–77 (2011).
13. Darnell, J.C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
14. Sanders, S.J. *et al.* Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885 (2011).
15. Xu, B. *et al. De novo* gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet.* **44**, 1365–1369 (2012).
16. Bustamante, C.D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).
17. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
18. Epi4K Consortium & Epilepsy Phenome/Genome Project. *De novo* mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).

ONLINE METHODS

***De novo* mutation information.** Published *de novo* mutations were collected for both ASD[2–6] and severe intellectual disability[9,10]. Updated *de novo* calls were provided from two of the ASD studies[3,5]. Details about sample collection, sequencing and variant processing can be found in the separate studies.

Additional sequencing. Exome sequencing of the additional families ($n = 129$) was performed at the Broad Institute. Exons were captured using Agilent 38Mb SureSelect v2. After capture, a round of ligation-mediated PCR was performed to increase the quantity of DNA available for sequencing. All libraries were sequenced using an Illumina HiSeq 2000 instrument. Data were processed with Picard, which uses base quality score recalibration and local realignment at known indels[19] and Burrows-Wheeler Aligner (BWA)[20] to map reads to hg19. SNPs were called using the Genome Analysis Toolkit (GATK) for all trios jointly[19,21]. The variable sites that we have considered in analysis were restricted to those that passed GATK standard filters. From this set of variants, we identified putative *de novo* mutations and validated them as previously described[3]. Autism Consortium samples ($n = 78$ trios) were collected in Boston under institutional review board (IRB) approval from Harvard Medical School, Massachusetts General Hospital, Children's Hospital Boston, Tufts–New England Medical Center and Boston University Medical Center with ADI and ADOS assessment. Finnish autism samples ($n = 51$ trios) were collected under IRB approval at the University of Helsinki with ADI and ADOS assessment and consented for autism research only. In both studies, all participants gave written informed consent, although, as autism is classified as a childhood disorder, many subjects are children, with informed consent provided by their parents or guardians.

Mutational model. We wanted to create an accurate model of *de novo* mutation for each gene. To do so, we extended a previous sequence context–based model of *de novo* mutation to derive gene-specific probabilities of mutation for each of the following mutation types: synonymous, missense, nonsense, essential splice site and frameshift[3]. In brief, local sequence context was used to determine the probability of each base in the coding region mutating to each other possible base and then to determine the coding impact of each possible mutation. These probabilities of mutation were summed across genes to create a per-gene probability of mutation for the aforementioned mutation types (see the **Supplementary Note** for more details). Here we applied the method to exons and immediately flanking essential splice sites, but note that the framework is applicable to non-genic sequences. While fitting the expected rates of mutation to observed data, we added a term for local primate divergence across 1 Mb (to capture additional unmeasured sources of regional mutational variability) and another for the average depth of sequence of each nucleotide (to capture inefficiency of variant discovery at lower sequencing depths); both terms significantly improved the fit of the model to observed data (details in the **Supplementary Note**). We also investigated a regional replication timing term[22] but found no evidence for it significantly improving the model (**Supplementary Note**).

To evaluate the predictive value of the model of *de novo* coding mutations, we extracted synonymous variants that were seen 10 times or fewer in the 6,503 individuals in ESP and compared the number of these rare variants in each gene to (i) the length of the gene and (ii) the probability of a synonymous mutation for that gene as determined by our model. Although gene length alone showed high correlation ($r = 0.880$), our full model showed significantly greater correlation ($r = 0.940$; $P < 1 \times 10^{-16}$). Of note, the stochastic variability of counts from ESP is such that, if the model were perfect, the correlation to any instance of these data would be 0.975, indicating that little additional gene-to-gene variability remains to be explained. The relative rates of different types of coding mutation were quite similar to those in previous work based on primate substitutions[23]. With this calibrated model of relative mutability, we determined the absolute expected mutation rate per gene by applying a genome-wide mutation rate of 1.2×10^{-8} mutations per base pair per generation (**Supplementary Note**)[24,25].

Removing potential false positive constrained genes. To identify genes that appeared to be significantly constrained, we used our probabilities of mutation to predict the expected amount of synonymous and nonsynonymous variation in ESP data. Those genes that had the expected amount of synonymous variation but were significantly ($P < 0.001$) deficient for missense variation were labeled as constrained. To ensure that genes were not nominated as being constrained erroneously, we excluded from all analyses 134 genes in which the observed synonymous and nonsynonymous rates were both significantly elevated or significantly decreased (both $P < 0.001$). Upon inspection, this list contained a number of genes that contained an internal duplication (for example, *FLG*), a nearby pseudogene (for example, *AHNAK2*) and a number of cases where recent duplications and/or annotation errors have led to the same sequence being assigned to two genes (for example, *SLX1A* and *SLX1B*). These are all scenarios where standard exome processing pipelines systematically undercall variation (reads are unmapped owing to uncertainty on which gene to assign them to) or overcall false variants owing to read misplacement. This further suggests that a byproduct of this analysis framework is the identification of a residual set of challenging genes for current exome sequencing pipelines.

19. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
20. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
21. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
22. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91**, 1033–1040 (2012).
23. Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
24. Campbell, C.D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44**, 1277–1281 (2012).
25. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712–714 (2011).

De Rubeis et al *Nature* 2014

De Rubeis S, He X, Goldberg AP, Poultney CS, **Samocha K**, Ercument Cicek A, Kou Y, Liu L, Fromer M, Walker S, Singh T, Klei L, Kosmicki J, Fu SC, Aleksic B, Biscaldi M, Bolton PF, Brownfeld JM, Cai J, Campbell NG, Carracedo A, Chahrour MH, Chiocchetti AG, Coon H, Crawford EL, Crooks L, Curran SR, Dawson G, Duketis E, Fernandez BA, Gallagher L, Geller E, Guter SJ, Sean Hill R, Ionita-Laza I, Jimenez Gonzalez P, Kilpinen H, Klauck SM, Kolevzon A, Lee I, Lei J, Lehtimäki T, Lin CF, Ma'ayan A, Marshall CR, McInnes AL, Neale B, Owen MJ, Ozaki N, Parellada M, Parr JR, Purcell S, Puura K, Rajagopalan D, Rehnström K, Reichenberg A, Sabo A, Sachse M, Sanders SJ, Schafer C, Schulte-Rüther M, Skuse D, Stevens C, Szatmari P, Tammimies K, Valladares O, Voran A, Wang LS, Weiss LA, Jeremy Willsey A, Yu TW, Yuen RK; The DDD Study; Homozygosity Mapping Collaborative for Autism; UK10K Consortium; The Autism Sequencing Consortium, Cook EH, Freitag CM, Gill M, Hultman CM, Lehner T, Palotie A, Schellenberg GD, Sklar P, State MW, Sutcliffe JS, Walsh CA, Scherer SW, Zwick ME, Barrett JC, Cutler DJ, Roeder K, Devlin B, Daly MJ, Buxbaum JD. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*. 2014 Nov 13;515(7526):209-15. doi: 10.1038/nature13772. Epub 2014 Oct 29.

ARTICLE

doi:10.1038/nature13772

Synaptic, transcriptional and chromatin genes disrupted in autism

A list of authors and their affiliations appears at the end of the paper

The genetic architecture of autism spectrum disorder involves the interplay of common and rare variants and their impact on hundreds of genes. Using exome sequencing, here we show that analysis of rare coding variation in 3,871 autism cases and 9,937 ancestry-matched or parental controls implicates 22 autosomal genes at a false discovery rate (FDR) < 0.05, plus a set of 107 autosomal genes strongly enriched for those likely to affect risk (FDR < 0.30). These 107 genes, which show unusual evolutionary constraint against mutations, incur *de novo* loss-of-function mutations in over 5% of autistic subjects. Many of the genes implicated encode proteins for synaptic formation, transcriptional regulation and chromatin-remodelling pathways. These include voltage-gated ion channels regulating the propagation of action potentials, pacemaking and excitability-transcription coupling, as well as histone-modifying enzymes and chromatin remodellers—most prominently those that mediate post-translational lysine methylation/demethylation modifications of histones.

Features of subjects with autism spectrum disorder (ASD) include compromised social communication and interaction. Because the bulk of risk arises from *de novo* and inherited genetic variation[1–10], characterizing which genes are involved informs ASD neurobiology and reveals part of what makes us social beings.

Whole-exome sequencing (WES) studies have proved fruitful in uncovering risk-conferring variation, especially by enumerating *de novo* variation, which is sufficiently rare that recurrent mutations in a gene provide strong evidence for a causal link to ASD. *De novo* loss-of-function (LoF) single-nucleotide variants (SNVs) or insertion/deletion (indel) variants[11–15] are found in 6.7% more ASD subjects than in matched controls and implicate nine genes from the first 1,000 ASD subjects analysed[11–16]. Moreover, because there are hundreds of genes involved in ASD risk, ongoing WES studies should identify additional ASD genes as an almost linear function of increasing sample size[11].

Here we conduct the largest ASD WES study so far, analysing 16 sample sets comprising 15,480 DNA samples (Supplementary Table 1 and Extended Data Fig. 1). Unlike earlier WES studies, we do not rely solely on counting *de novo* LoF variants, rather we use novel statistical methods to assess association for autosomal genes by integrating *de novo*, inherited and case-control LoF counts, as well as *de novo* missense variants predicted to be damaging. For many samples original data from sequencing performed on Illumina HiSeq 2000 systems were used to call SNVs and indels in a single large batch using GATK (v2.6)[17]. *De novo* mutations were called using enhancements of earlier methods[14] (Supplementary Information), with calls validating at extremely high rates.

After evaluation of data quality, high-quality alternative alleles with a frequency of <0.1% were identified, restricted to LoF (frameshifts, stop gains, donor/acceptor splice site mutations) or probably damaging missense (Mis3) variants (defined by PolyPhen-2 (ref. 18)). Variants were classified by type (*de novo*, case, control, transmitted, non-transmitted) and severity (LoF, Mis3), and counts tallied for each gene.

Some 13.8% of the 2,270 ASD trios (two parents and one affected child) carried a *de novo* LoF mutation—significantly in excess of both the expected value[19] (8.6%, $P < 10^{-14}$) and what was observed in 510 control trios (7.1%, $P = 1.6 \times 10^{-5}$) collected here and previously published[15]. Eighteen genes (Table 1) exhibited two or more *de novo* LoF mutations. These genes are all known or strong candidate ASD genes, but given the number of trios sequenced and gene mutability[14,19], we would expect to observe this in approximately two such genes by chance. While we expect only two *de novo* Mis3 events in these 18 genes, we observe 16 ($P = 9.2 \times 10^{-11}$, Poisson test). Because most of our data exist in cases and controls and because we observed an additional excess of transmitted LoF events in the 18 genes, it is evident that the optimal analytical framework must involve an integration of *de novo* mutation with variants observed in cases and controls and transmitted or untransmitted from carrier parents. Investigating beyond *de novo* LoFs is also critical given that many ASD risk genes and loci have mutations that are not completely penetrant.

Transmission and *de novo* association

We adopted TADA (transmission and *de novo* association), a weighted, statistical model integrating *de novo*, transmitted and case-control variation[20]. TADA uses a Bayesian gene-based likelihood model including per-gene mutation rates, allele frequencies, and relative risks of particular classes of sequence changes. We modelled both LoF and Mis3 sequence variants. Because no aggregate association signal was detected for inherited Mis3 variants, they were not included in the analysis. For each gene, variants of each class were assigned the same effect on relative risk. Using a prior probability distribution of relative risk across genes for each class of variants, the model effectively weighted different classes of variants in this order: *de novo* LoF > *de novo* Mis3 > transmitted LoF, and allowed for a distribution of relative risks across genes for each class. The strength of association was assimilated across classes to produce a gene-level Bayes factor with a corresponding FDR *q* value. This framework increases the power compared to the use of *de novo* LoF variants alone (Extended Data Fig. 2).

TADA identified 33 autosomal genes with an FDR < 0.1 (Table 1) and 107 with an FDR < 0.3 (Supplementary Tables 2 and 3 and Extended Data Fig. 3). Of the 33 genes, 15 (45.5%) are known ASD risk genes[9]; 11 have been reported previously with mutations in ASD patients but were not classed as true risk genes owing to insufficient evidence (*SUV420H1* (refs 11, 15), *ADNP*[12], *BCL11A*[15], *CACNA2D3* (refs 15, 21), *CTTNBP2* (ref. 15), *GABRB3* (ref. 21), *CDC42BPB*[13], *APH1A*[14], *NR3C2* (ref. 15), *SETD5* (refs 14, 22) and *TRIO*[11]) and 7 are completely novel (*ASH1L*, *MLL3* (also known as *KMT2C*), *ETFB*, *NAA15*, *MYO9B*, *MIB1* and *VIL1*). *ADNP* mutations have recently been identified in 10 patients with ASD and other shared clinical features[23]. Two of the newly discovered genes,

Table 1 | ASD risk genes

dnLoF count	FDR ≤ 0.01	0.01 < FDR ≤ 0.05	0.05 < FDR ≤ 0.1
≥2	*ADNP, ANK2, ARID1B, CHD8, CUL3, DYRK1A, GRIN2B, KATNAL2, POGZ, SCN2A, SUV420H1, SYNGAP1, TBR1*	*ASXL3, BCL11A, CACNA2D3, MLL3*	*ASH1L*
1		*CTTNBP2, GABRB3, PTEN, RELN*	*APH1A, CD42BPB, ETFB, NAA15, MYO9B, MYT1L, NR3C2, SETD5, TRIO*
0		*MIB1*	*VIL1*

TADA analysis of LoF and damaging missense variants found to be *de novo* in ASD subjects, inherited by ASD subjects, or present in ASD subjects (versus control subjects). dnLoF, de *novo* LoF events.

ASH1L and *MLL3*, converge on chromatin remodelling. *MYO9B* plays a key role in dendritic arborization[24]. *MIB1* encodes an E3 ubiquitin ligase critical for neurogenesis[25] and is regulated by miR-137 (ref. 26), a microRNA that regulates neuronal maturation and is implicated in schizophrenia risk[27].

When the WES data from genes with an FDR < 0.3 were evaluated for the presence of deletion copy number variants (CNVs) (such CNVs are functionally equivalent to LoF mutations), 34 CNVs meeting quality and frequency constraints (Supplementary Information) were detected in 5,781 samples (Extended Data Fig. 1). Of the 33 genes with an FDR < 0.1, 3 contained deletion CNVs mapping to 3 ASD subjects and one parent. Of the 74 genes meeting the criterion 0.1 ≤ FDR < 0.3, about one-third could be false positives. Deletion CNVs were found in 14 of these genes and the data supported risk status for 10 of them (Extended Data Table 1 and Extended Data Fig. 4). Two of these ten, *NRXN1* and *SHANK3*, were previously implicated in ASD[2,3,10]. The risk from deletion CNVs, as measured by the odds ratio, is comparable to that from LoF SNVs in cases versus controls or transmission of LoF variants from parents to offspring.

Estimated odds ratios of top genes

Inherent in our conception of the biology of ASD is the notion that there is variation between genes in their impact on risk; for a given class of variants (for example, LoF) some genes have a large impact, others smaller, and still others have no effect at all. In addition, misannotation of variants, among other confounds, can yield false variant calls in subjects (Supplementary Information). These confounds can often be overcome by examining the data in a manner orthogonal to gene discovery. For example, females have greatly reduced rates of ASD relative to males (a 'female protective effect'). Consequently, and regardless of whether this is diagnostic bias or biological protection, females have a higher liability threshold, requiring a larger genetic burden before being diagnosed[22,28,29]. A corollary is that if a variant has the same effect on autism liability in males as it does in females, that variant will be present at a higher frequency in female ASD cases compared to males. Importantly, the magnitude of the difference is proportional to risk as measured by the odds ratio; hence, the effect on risk for a class of variants can be estimated from the difference in frequency between males and females.

Genes with an FDR < 0.1 show profound female enrichment for *de novo* events ($P = 0.005$ for LoF, $P = 0.004$ for Mis3), consistent with *de novo* events having large impacts on liability (odds ratio ≥ 20; Extended Data Fig. 5). However, genes with an FDR between 0.1 and 0.3 show substantially less enrichment for female events, consistent with a modest impact for LoF variants (odds ratio range 2–4, whether transmitted or *de novo*) and little to no effect from Mis3 variants. The

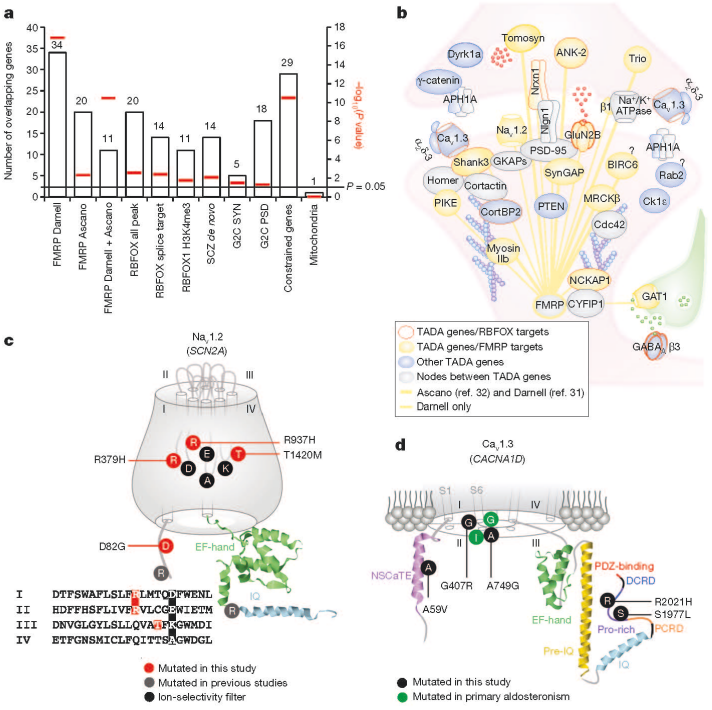


Figure 1 | ASD genes in synaptic networks. **a,** Enrichment of 107 TADA genes in: FMRP targets from two independent data sets[31,32] and their overlap; RBFOX targets; RBFOX targets with predicted alterations in splicing; RBFOX1 and H3K4me3 overlapping targets; genes with *de novo* mutations in schizophrenia (SCZ); human orthologues of Genes2Cognition (G2C) mouse synaptosome (SYN) or PSD genes; constrained genes; and genes encoding mitochondrial proteins (as a control). Red bars indicate empirical *P* values (Supplementary Information). **b,** Synaptic proteins encoded by TADA genes. **c,** *De novo* Mis3 variants in $Na_v1.2$ (*SCN2A*). The four repeats (I–IV) with P-loops, the EF-hand, and the IQ domain are shown, as are the four amino acids (DEKA) forming the inner ring of the ion-selectivity filter. **d,** Variants in $Ca_v1.3$ (*CACNA1D*). Part of the channel is shown, including helices one and six (S1 and S6) for domains I–IV, the NSCaTE motif, the EF-hand domain, the pre-IQ, IQ, proximal (PCRD) and distal (DCRD) C-terminal regulatory domains, the proline-rich region, and the PDZ domain-binding motif.

results are consistent with inheritance patterns: LoF mutations in FDR < 0.1 genes are rarely inherited from unaffected parents whereas those in the 0.1 ≤ FDR < 0.3 group are far more often inherited than they are *de novo* mutations.

By analysing the distribution of relative risk over inferred ASD genes[20], the number of ASD risk genes can be estimated. The estimate relies on the balance of genes with multiple *de novo* LoF mutations versus those with only one: the larger the number of ASD genes, the greater proportion that will show only one *de novo* LoF. This approach yields an estimate of 1,150 ASD genes (Supplementary Information). While there are many more genes to be discovered, many will have a modest impact on risk compared to the genes in Table 1.

Enrichment analyses

Gene sets with an FDR < 0.3 are strongly enriched for genes under evolutionary constraint[19] ($P = 3.0 \times 10^{-11}$; Fig. 1a and Supplementary Table 4), consistent with the hypothesis that heterozygous LoF mutations in these genes are ASD risk factors. Over 5% of ASD subjects carry *de novo* LoF mutations in our FDR < 0.3 list. We also observed that genes in the FDR < 0.3 list had a significant excess of *de novo* non-synonymous events detected by the largest schizophrenia WES study so far[30] ($P = 0.0085$; Fig. 1a), providing further evidence for overlapping risk loci between these disorders and independent confirmation of the signal in the gene sets presented here.

We found significant enrichment for genes encoding messenger RNAs targeted by two neuronal RNA-binding proteins: FMRP[31] (also known as FMR1), mutated or absent in fragile X syndrome ($P = 1.20 \times 10^{-17}$, 34 targets[31], of which 11 are corroborated by an independent data set[32]), and RBFOX (RBFOX1/2/3) ($P = 0.0024$, 20 targets, of which 12 overlap with FMRP), with RBFOX1 shown to be a splicing factor dysregulated in ASD[33,34] (Fig. 1a). These two pathways expand the complexity of ASD neurobiology to post-transcriptional events, including splicing and translation, both of which sculpt the neural proteome.

We found nominal enrichment for human orthologues of mouse genes encoding synaptic ($P = 0.031$) and post-synaptic density (PSD) proteins[35] ($P = 0.046$; Fig. 1a, b and Supplementary Tables 4–6). Enrichment analyses for InterPro, SMART or Pfam domains (FDR < 0.05 and a minimum of five genes per category) reveal an overrepresentation of DNA- or histone-related domains: eight genes encoding proteins with InterPro zinc-finger FYVE PHD domains (142 such annotated genes in the genome; FDR = 7.6×10^{-4}), and five with Pfam Su(var)3-9, enhancer-of-zeste, trithorax (SET) domains (39 annotated in the genome; FDR = 8.2×10^{-4}).

Integrating complementary data

To implicate additional genes in risk for ASD, we used a model called DAWN (detecting association with networks)[36]. DAWN evokes a hidden Markov random field framework to identify clusters of genes that show strong association signals and highly correlated co-expression in a key tissue and developmental context. Previous research suggests human mid-fetal prefrontal and motor-somatosensory neocortex is a critical nexus for risk[16], thus we evaluated gene co-expression data from that tissue together with TADA scores for genes with an FDR < 0.3. Because this list is enriched for genes under evolutionary constraint, we generalized DAWN to incorporate constraint scores (Supplementary Information). When TADA results, gene co-expression in mid-fetal neocortex and constraint scores are jointly modelled, DAWN identifies 160 genes that plausibly affect risk (Fig. 2), 91 of which are not in the 107 TADA genes with an FDR < 0.3. Moreover, the model parameter describing evolutionary constraint is an important predictor of clusters of putative risk genes ($P = 0.018$).

A subnetwork obtained by seeding the 160 DAWN genes within a high-confidence protein–protein interactome[14] confirmed that the putative genes are enriched for neuronal functions. We kept the largest connected component, containing 95 seed DAWN genes, 50 of which were in the FDR < 0.3 gene set. The DAWN gene products form four natural

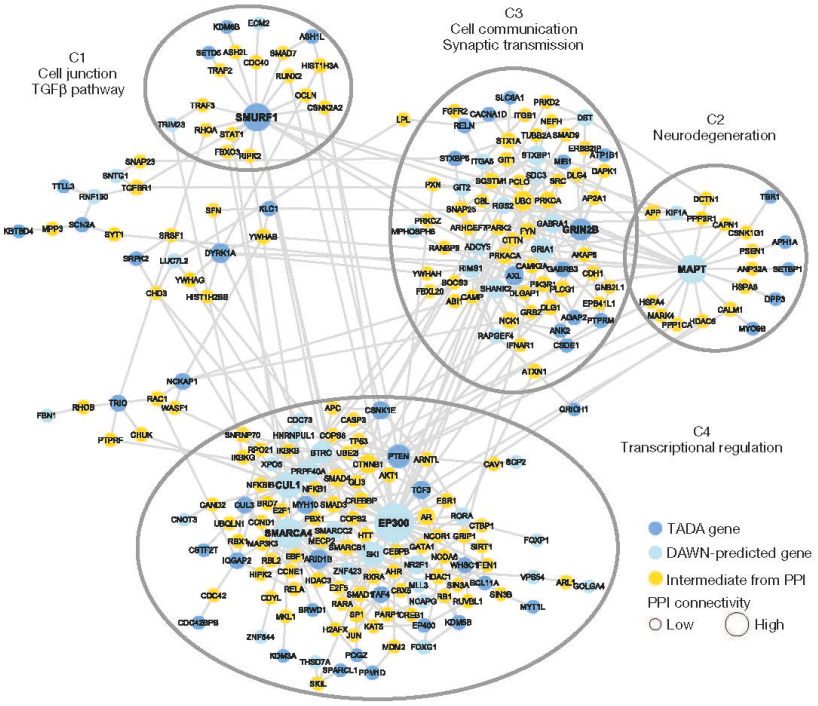


Figure 2 | ASD genes in neuronal networks. Protein–protein interaction network created by seeding TADA and DAWN-predicted genes. Only intermediate genes that are known to interact with at least two TADA and/or DAWN genes are included. Four natural clusters (C1–C4) are demarcated with black ellipses. All nodes are sized on the basis of degree of connectivity.

clusters on the basis of network connectivity (Fig. 2). We visualized the enriched pathways and biological functions for each of these clusters on 'canvases'[37] (Extended Data Fig. 6). Many of the previously known ASD risk genes fall in cluster C3, including genes involved in synaptic transmission and cell–cell communication. Cluster C4 is enriched for genes related to transcriptional and chromatin regulation. Many TADA and DAWN genes in this cluster interact tightly with other transcription factors, histone-modifying enzymes and DNA-binding proteins. Five TADA genes in the cluster C2 are bridged to the rest of the network through *MAPT*, as inferred by DAWN. The enrichment results for cluster C2 indicate that genes implicated in neurodegenerative disorders could also have a role in neurodevelopmental disorders.

Emergent results

Amongst the critical synaptic components found to be mutated in our study are voltage-gated ion channels involved in fundamental processes including the propagation of action potentials (for example, the $Na_v1.2$ channel), neuronal pacemaking and excitability–transcription coupling (for example, the $Ca_v1.3$ channel) (Fig. 1b). We identified four LoF and five Mis3 variants in *SCN2A* ($Na_v1.2$), three Mis3 variants in *CACNA1D* ($Ca_v1.3$) and two LoF variants in *CACNA2D3* ($\alpha_2\delta$-3 subunit). Remarkably, three *de novo* Mis3 variants in *SCN2A* affected residues mutated in homologous genes in patients with other syndromes, including Brugada syndrome (*SCN5A*) or epilepsy disorders (*SCN1A*) (Arg379His and Arg 937His). These arginines, as well as the threonine mutated in Thr1420Met, cluster to the P-loops forming the ion selectivity filter, located in proximity to the inner ring (DEKA motif) (Fig. 1c). Because homologous channels mutated in these arginines do not conduct inward Na^+ currents[38,39], Arg379His and Arg937His mutations might have similar effect.

Two *de novo* *CACNA1D* variants (Gly407Arg and Ala749Gly) emerged at positions proximal to residues mutated in patients with primary aldosteronism and neurological deficits (Fig. 1d). The reported mutations interfere with channel activation and inactivation[40]. Amongst variants found in cases, Ala59Val maps to the NSCaTE domain, also important for Ca^{2+}-dependent inactivation, and Ser1977Leu and Arg2021His co-cluster in the carboxy-terminal proline-rich domain, the site of interaction with SHANK3, a key PSD scaffolding protein. Mutations in RIMS1 and RIMBP2, which can associate with $Ca_v1.3$, were found in our cohort (but with an FDR > 0.3).

Chromatin remodelling involves histone-modifying enzymes (encoded by histone-modifier genes, HMGs) and chromatin remodellers (readers) that recognize specific histone post-translational modifications and orchestrate their effects on chromatin. Our gene set is enriched in HMGs (9 HMGs out of 152 annotated in HIstome[41], Fisher's exact test, $P = 2.2 \times 10^{-7}$). Enrichment in the gene ontology term 'histone-lysine *N*-methytransferase activity' (5 genes out of 41 so annotated; FDR = 2.2×10^{-2}) highlights this as a prominent pathway.

Lysines on histones 3 and 4 can be mono-, di- or tri-methylated, providing a versatile mechanism for either activation or repression of transcription. Of 107 TADA genes, five are SET lysine methyltransferases, four are jumonji lysine demethylases, and two are readers (Fig. 3a). RBFOX1 co-isolates with histone H3 trimethyl Lys 4 (H3K4me3)[42], and our data set is enriched in targets shared by RBFOX1 and H3K4me3 ($P = 0.0166$; Fig. 1a and Supplementary Table 4). Some *de novo* missense variants targeting these genes map to functional domains (Extended Data Fig. 7).

For the H3K4me2 reader *CHD8*, we extended our analyses in search of additional *de novo* variation in the cases of the case-control sample. By sequencing complete parent–child trios for many *CHD8* variants, five variants were found to be *de novo*, two of which affect essential splice sites and cause LoF by exon skipping or activation of cryptic splice sites in lymphoblastoid cells (Fig. 3b).

Given the role of HMGs in transcription, we reasoned that TADA genes might be interconnected through transcription 'routes'. We searched for a connected network (seeded by 9 TADA HMGs) in a transcription factor interaction network (ChEA)[43]. We found that 46 TADA genes

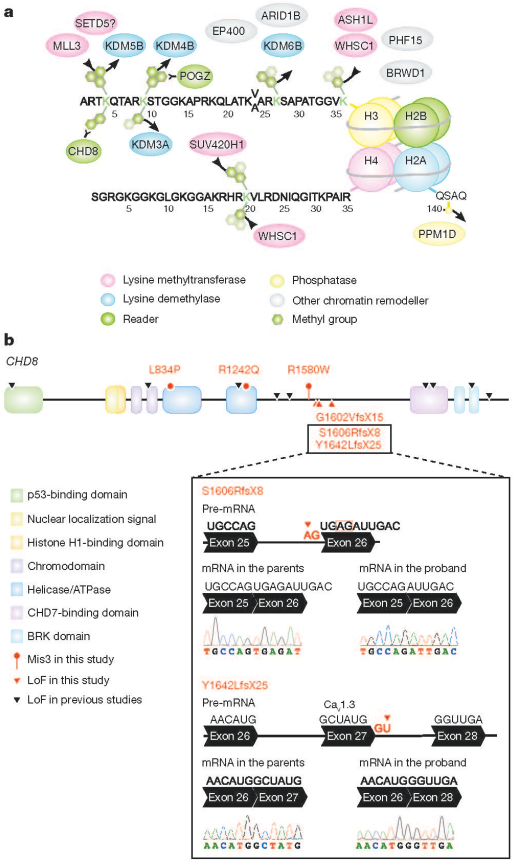


Figure 3 | ASD genes in chromatin remodelling. a, TADA genes cluster to chromatin-remodelling complexes. Amino-terminals of histones H3, H4 and part of H2A are shown. Lysine methyltransferases add methyl groups, whereas lysine demethylases remove them. **b,** *De novo* Mis3 and LoF variants in CHD8. The box shows the outcome of reverse transcription PCR and Sanger sequencing in lymphoblastoid cells for two newly identified *de novo* splice-site variants. The first mutation affects an acceptor splice site (red arrow), causing the activation of a cryptic splice site (red box), a four-nucleotide deletion, frame shift and a premature stop. The second mutation affects a donor splice site (red arrow), causing exon skipping, frame shift and a premature stop.

are directly interconnected in a 55-gene cluster (Extended Data Fig. 8) ($P = 0.002$; 1,000 random draws), for a total of 69 when including all known HMGs (Fig. 4) ($P = 0.001$; 1,000 random draws).

Examining the Human Gene Mutation Database we found that the 107 TADA genes included 21 candidate genes for intellectual disability, 3 for epilepsy, 17 for schizophrenia, 9 for congenital heart disease and 6 for metabolic disorders (Fig. 5).

Conclusions

Complementing earlier reports, ASD subjects show a clear excess of *de novo* LoF mutations above expectation, with a concentration of such events in a handful of genes. While this handful has a large effect on risk, most ASD genes have a much smaller impact. This gradient emerges most notably from the contrast of risk variation in male and female ASD subjects. Unlike some earlier studies, but consistent with expectation, the data also show clear evidence for effect of *de novo* missense SNVs

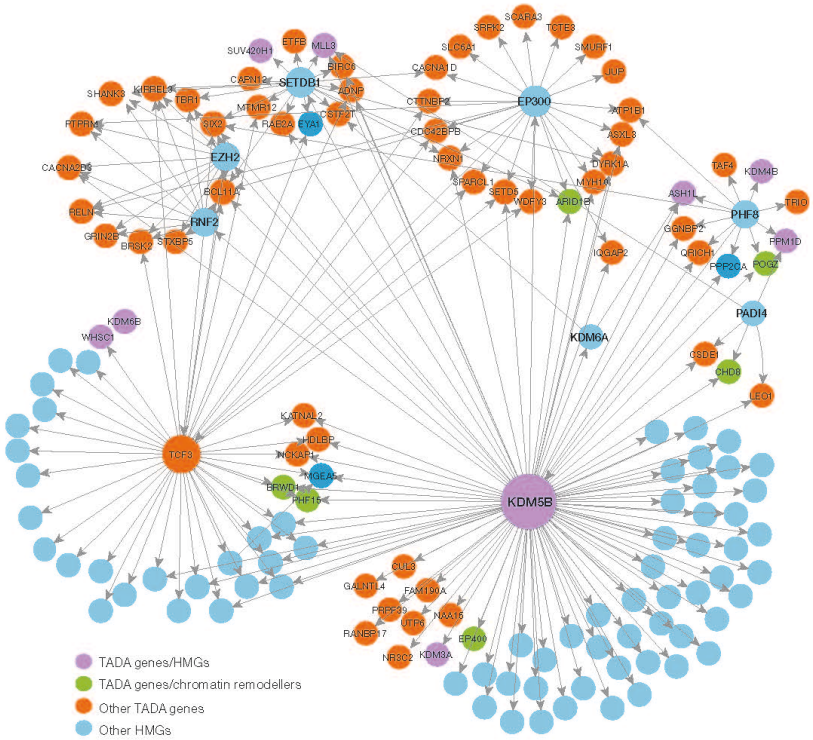


Figure 4 | Transcription regulation network of TADA genes. Edges indicate transcription regulators (source nodes) and their gene targets (target nodes) based on the ChEA network; interactions among only HMGs are ignored.

on risk; for risk generated by LoF variants transmitted from unaffected parents; and for the value of case-control design in gene discovery. By integrating data on *de novo*, inherited and case-control variation, the yield of ASD gene discoveries was doubled over what would be obtained from a count of *de novo* LoF variants alone. ASD genes almost uniformly show strong constraints against variation, a feature we exploit to implicate other genes in risk.

Three critical pathways for typical development are damaged by risk variation: chromatin remodelling, transcription and splicing, and synaptic function. Chromatin remodelling controls events underlying the formation of neural connections, including neurogenesis and neural differentiation[44], and relies on epigenetic marks as post-translational modifications of histones. Here we provide extensive evidence for HMGs and readers in sporadic ASD, implicating specifically lysine methylation and extending the mutational landscape of the emergent ASD gene *CHD8* to missense variants. Splicing is implicated by the enrichment of RBFOX targets in the top ASD candidates. Risk variation also affects multiple classes and components of synaptic networks, from receptors and ion channels to scaffolding proteins. Because a wide set of synaptic genes is disrupted in idiopathic ASD, it seems reasonable to suggest that altered chromatin dynamics and transcription, induced by disruption of relevant genes, leads to impaired synaptic function as well. *De novo* mutations in ASD[11–15], intellectual disability[45] and schizophrenia[30] cluster to synaptic genes, and synaptic defects have been reported in models of these disorders[46]. Integrity of synaptic function is essential for neural physiology, and its perturbation could represent the intersection between diverse neuropsychiatric disorders[47].

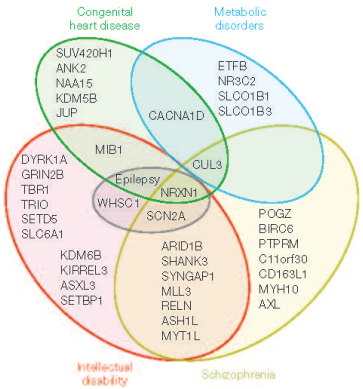


Figure 5 | Involvement in disease of ASD genes. The Venn diagram shows the overlap in disease involvement for the TADA genes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 May; accepted 18 August 2014.
Published online 29 October 2014.

1. Ronald, A. & Hoekstra, R. A. Autism spectrum disorders and autistic traits: a decade of new twin studies. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **156,** 255–274 (2011).
2. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316,** 445–449 (2007).
3. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466,** 368–372 (2010).
4. Klei, L. *et al.* Common genetic variants, acting additively, are a major source of risk for autism. *Mol. Autism* **3,** 9 (2012).

5. Gaugler, T. *et al.* Most inherited risk for autism resides with common variation. *Nature Genet.* **46,** 881–885 (2014).
6. Yu, T. W. *et al.* Using whole-exome sequencing to identify inherited causes of autism. *Neuron* **77,** 259–273 (2013).
7. Lim, E. T. *et al.* Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron* **77,** 235–242 (2013).
8. Poultney, C. S. *et al.* Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. *Am. J. Hum. Genet.* **93,** 607–619 (2013).
9. Betancur, C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380,** 42–77 (2011).
10. Glessner, J. T. *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459,** 569–573 (2009).
11. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237–241 (2012).
12. O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338,** 1619–1622 (2012).
13. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485,** 246–250 (2012).
14. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485,** 242–245 (2012).
15. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74,** 285–299 (2012).
16. Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155,** 997–1007 (2013).
17. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43,** 491–498 (2011).
18. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).
19. Samocha, K. E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nature Genet.* **46,** 944–950 (2014).
20. He, X. *et al.* Integrated model of *de novo* and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.* **9,** e1003671 (2013).
21. Girirajan, S. *et al.* Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am. J. Hum. Genet.* **92,** 221–237 (2013).
22. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **94,** 677–694 (2014).
23. Helsmoortel, C. *et al.* A SWI/SNF-related autism syndrome caused by *de novo* mutations in *ADNP*. *Nature Genet.* **46,** 380–384 (2014).
24. Long, H. *et al.* Myo9b and RICS modulate dendritic morphology of cortical neurons. *Cereb. Cortex* **23,** 71–79 (2013).
25. Yoon, K. J. *et al.* Mind bomb 1-expressing intermediate progenitors generate Notch signaling to maintain radial glial cells. *Neuron* **58,** 519–531 (2008).
26. Smrt, R. D. *et al.* MicroRNA miR-137 regulates neuronal maturation by targeting ubiquitin ligase Mind bomb-1. *Stem Cells* **28,** 1060–1070 (2010).
27. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genet.* **45,** 1150–1159 (2013).
28. Robinson, E. B., Lichtenstein, P., Anckarsater, H., Happe, F. & Ronald, A. Examining and interpreting the female protective effect against autistic behavior. *Proc. Natl Acad. Sci. USA* **110,** 5258–5262 (2013).
29. Jacquemont, S. *et al.* A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. *Am. J. Hum. Genet.* **94,** 415–425 (2014).
30. Fromer, M. *et al. De novo* mutations in schizophrenia implicate synaptic networks. *Nature* **506,** 179–184 (2014).
31. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146,** 247–261 (2011).
32. Ascano, M. Jr. *et al.* FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature* **492,** 382–386 (2012).
33. Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep.* **6,** 1139–1152 (2014).
34. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474,** 380–384 (2011).
35. Collins, M. O. *et al.* Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *J. Neurochem.* **97** (suppl. 1), 16–23 (2006).
36. Liu, L. *et al.* DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. *Mol. Autism* **5,** 22 (2014).
37. Tan, C. M., Chen, E. Y., Dannenfelser, R., Clark, N. R. & Ma'ayan, A. Network2Canvas: network visualization on a canvas with enrichment analysis. *Bioinformatics* **29,** 1872–1878 (2013).
38. Vatta, M. *et al.* Genetic and biophysical basis of sudden unexplained nocturnal death syndrome (SUNDS), a disease allelic to Brugada syndrome. *Hum. Mol. Genet.* **11,** 337–345 (2002).
39. Volkers, L. *et al.* $Na_v1.1$ dysfunction in genetic epilepsy with febrile seizures-plus or Dravet syndrome. *Eur. J. Neurosci.* **34,** 1268–1275 (2011).
40. Scholl, U. I. *et al.* Somatic and germline *CACNA1D* calcium channel mutations in aldosterone-producing adenomas and primary aldosteronism. *Nature Genet.* **45,** 1050–1054 (2013).
41. Khare, S. P. *et al.* HIstome–a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Res.* **40,** D337–D342 (2012).
42. Feng, J. *et al.* Chronic cocaine-regulated epigenomic changes in mouse nucleus accumbens. *Genome Biol.* **15,** R65 (2014).
43. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26,** 2438–2444 (2010).
44. Ronan, J. L., Wu, W. & Crabtree, G. R. From neural development to cognition: unexpected roles for chromatin. *Nature Rev. Genet.* **14,** 347–359 (2013).
45. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380,** 1674–1682 (2012).
46. Penzes, P., Cahill, M. E., Jones, K. A., VanLeeuwen, J. E. & Woolfrey, K. M. Dendritic spine pathology in neuropsychiatric disorders. *Nature Neurosci.* **14,** 285–293 (2011).
47. Zoghbi, H. Y. Postnatal neurodevelopmental disorders: meeting at the synapse? *Science* **302,** 826–830 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by National Institutes of Health (NIH) grants U01MH100233, U01MH100209, U01MH100229 and U01MH100239 to the Autism Sequencing Consortium. Sequencing at Broad Institute was supported by NIH grants R01MH089208 (M.J.D.) and new sequencing by U54 HG003067 (S. Gabriel, E. Lander). Other funding includes NIH R01 MH089482, R37 MH057881 (B.D. and K.R.), R01 MH061009 (J.S.S.), UL1TR000445 (NCAT to VUMC); P50 HD055751 (E.H.C.); MH089482 (J.S.S.), NIH R01 MH083565 and RC2MH089952 (C.A.W.), NIMH MH095034 (P.S), MH077139 (P.F. Sullivan); 5UL1 RR024975 and P30 HD15052. The DDD Study is funded by HICF-1009-003 and WT098051. UK10K is funded by WT091310. We also acknowledge The National Children's Research Foundation, Our Lady's Children's Hospital, Crumlin; The Meath Foundation; AMNCH, Tallaght; The Health Research Board, Ireland and Autism Speaks, U.S.A. C.A.W. is an Investigator of the Howard Hughes Medical Institute. S.D.R., A.P.G., C.S.P., Y.K. and S.-C.F. are Seaver fellows, supported by the Seaver foundation. A.P.G. is also supported by the Charles and Ann Schlaifer Memorial Fund. P.F.B. is supported by a UK National Institute for Health Research (NIHR) Senior Investigator award and the NIHR Biomedical Research Centre in Mental Health at the South London & Maudsley Hospital. A.C. is supported by María José Jove Foundation and the grant FIS PI13/01136 of the Strategic Action from Health Carlos III Institute (FEDER). This work was supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai. We acknowledge the assistance of D. Hall and his team at National Database for Autism Research. We thank Jian Feng for providing a list of targets of both RBFOX1 and H3K4me3. We thank M. Potter for data coordination; K. Moore and J. Reichert for technical assistance; and, S. Lindsay for helping with molecular validation. We acknowledge the clinicians and organizations that contributed to samples used in this study. Finally, we are grateful to the many families whose participation made this study possible.

Author Contributions Study conception and design: J.D.B., D.J.C., M.J.D., S.D.R., B.D., M.F., A.P.G., X.H., T.L., C.S.P., K.Ro., M.W.S. and M.E.Z. Data analysis: J.C.B., P.F.B., J.D.B., J.C., A.E.C, D.J.C., M.J.D., S.D.R., B.D., M.F., S.-C.F., A.P.G., X.H., L.K., J.K., Y.K., L.L., A.M., C.S.P., S.P., K.Ro., K.S., C.S., T.S., C.St., S.W., L.W. and M.E.Z. Contribution of samples, WES data or analytical tools: B.A., J.C.B., M.B., P.F.B., J.D.B., J.C., N.G.C., A.C., M.H.C., A.G.C., A.E.C, H.C., E.L.C., L.C., S.R.C., D.J.C., M.J.D., G.D., S.D.R., B.D., E.D., B.A.F., C.M.F., M.F., L.G., E.G., M.G., A.P.G., S.J.G., X.H., R.H., C.M.H., I.I.-L., P.J.G., H.K., S.M.K., L.K., A.K., J.K., Y.K., I.L., J.L., T.Le., C.L., L.L., A.M., C.R.M., A.L.M., B.N., M.J.O., N.O., A.P., M.P., J.R.P., C.S.P., S.P., K.P., D.R., K.R., A.R., K.Ro., A.S., M.S., K.S., S.J.S., C.S., G.D.S., S.W.S., M.S.-R., T.S., P.S., D.S., M.W.S., C.St., J.S.S., P.Sz., K.T., O.V., A.V., S.W., C.A.W., L.W., L.A.W., J.A.W., T.W.Y., R.K.C.Y., M.E.Z. Writing of the paper: J.C.B., J.D.B., E.H.C., D.J.C., M.J.D., S.D.R., B.D., M.G., A.P.G., X.H., C.S.P., K.Ro., S.W.S., M.E.Z. Leads of ASC committees: J.D.B., E.H.C., M.J.D., B.D., M.G., K.Ro., M.W.S., J.S.S., M.E.Z. Administration of ASC: J.M.B.

Author Information New data included in this manuscript have been deposited at dbGAP merged with our published data under accession number phs000298.v1.p1 and is available for download at (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000298.v1.p1). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.D.B. (joseph.buxbaum@mssm.edu) or M.J.D. (mjdaly@broadinstitute.org).

Silvia De Rubeis[1,2], Xin He[3], Arthur P. Goldberg[1,2,4], Christopher S. Poultney[1,2], Kaitlin Samocha[5], A. Ercument Cicek[3], Yan Kou[1,2], Li Liu[6], Menachem Fromer[2,4,5], Susan Walker[7], Tarjinder Singh[8], Lambertus Klei[9], Jack Kosmicki[5], Shih-Chen Fu[1,2], Branko Aleksic[10], Monica Biscaldi[11], Patrick F. Bolton[12], Jessica M. Brownfeld[1,2], Jinlu Cai[1,2], Nicholas G. Campbell[13,14], Angel Carracedo[15,16], Maria H. Chahrour[17,18], Andreas G. Chiocchetti[19], Hilary Coon[20,21], Emily L. Crawford[13,14], Lucy Crooks[8], Sarah R. Curran[12], Geraldine Dawson[22], Eftichia Duketis[19], Bridget A. Fernandez[23], Louise Gallagher[24], Evan Geller[25], Stephen J. Guter[26], R. Sean Hill[17,18], Iuliana Ionita-Laza[27], Patricia Jimenez Gonzalez[28], Helena Kilpinen[29], Sabine M. Klauck[30], Alexander Kolevzon[1,2,31], Irene Lee[32], Jing Lei[6], Terho Lehtimäki[33], Chiao-Feng Lin[25], Avi Ma'ayan[34], Christian R. Marshall[7], Alison L. McInnes[35], Benjamin Neale[36], Michael J. Owen[37], Norio Ozaki[10], Mara Parellada[38], Jeremy R. Parr[39], Shaun Purcell[2], Kaija Puura[40], Deepthi Rajagopalan[7], Karola Rehnström[8], Abraham Reichenberg[1,2,41], Aniko Sabo[42], Michael Sachse[19], Stephan J. Sanders[43], Chad Schafer[6], Martin Schulte-Rüther[44], David Skuse[32,45], Christine Stevens[36], Peter Szatmari[46], Kristiina Tammimies[7], Otto Valladares[25], Annette Voran[47], Li-San Wang[25], Lauren A. Weiss[43], A. Jeremy Willsey[43], Timothy W. Yu[17,18], Ryan K. C. Yuen[7], The DDD Study*, Homozygosity Mapping Collaborative for Autism*, UK10K Consortium*, The Autism Sequencing Consortium*, Edwin H. Cook[26], Christine M. Freitag[19], Michael Gill[24], Christina M. Hultman[48], Thomas Lehner[49], Aarno Palotie[5,50,51,52], Gerard D. Schellenberg[25], Pamela Sklar[2,4,53], Matthew W. State[43], James S. Sutcliffe[13,14], Christopher A. Walsh[17,18],

Stephen W. Scherer[7,54], Michael E. Zwick[55], Jeffrey C. Barrett[8], David J. Cutler[55], Kathryn Roeder[6,3], Bernie Devlin[9], Mark J. Daly[17,36,56] & Joseph D. Buxbaum[1,2,4,53,57,58]

[1]Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [2]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York 10029, New York, USA. [3]Ray and Stephanie Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. [4]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [5]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [6]Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. [7]Program in Genetics and Genome Biology, The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario M5G 0A4, Canada. [8]The Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, UK. [9]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213, USA. [10]Department of Psychiatry, Graduate School of Medicine, Nagoya University, Nagoya 466-8550, Japan. [11]Department of Child and Adolescent Psychiatry, Psychotherapy, and Psychosomatics, University Medical Center Freiburg; Center for Mental Disorders, 79106 Freiburg, Germany. [12]Department of Child Psychiatry & SGDP Centre, King's College London Institute of Psychiatry, Psychology & Neuroscience, London, SE5 8AF, UK. [13]Vanderbilt Brain Institute, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. [14]Department of Molecular Physiology and Biophysics and Psychiatry, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. [15]Genomic Medicine Group, CIBERER, University of Santiago de Compostela and Galician Foundation of Genomic Medicine (SERGAS), 15706 Santiago de Compostela, Spain. [16]Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah 21589, Kingdom of Saudi Arabia. [17]Harvard Medical School, Boston, Massachusetts 02115, USA. [18]Division of Genetics and Genomics, Boston Children's Hospital, Boston, Massachusetts 02115, USA. [19]Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Goethe University Frankfurt, 60528 Frankfurt, Germany. [20]Department of Internal Medicine, University of Utah, Salt Lake City, Utah 84132, USA. [21]Department of Psychiatry, University of Utah, Salt Lake City, Utah 84108, USA. [22]Duke Institute for Brain Sciences, Duke University, Durham, North Carolina 27708, USA. [23]Disciplines of Genetics and Medicine, Memorial University of Newfoundland, St John's, Newfoundland A1B 3V6, Canada. [24]Department of Psychiatry, School of Medicine, Trinity College Dublin, Dublin 8, Ireland. [25]Geisinger Health System, Danville, Pennsylvania 17822, USA. [26]Institute for Juvenile Research, Department of Psychiatry, University of Illinois at Chicago, Chicago, Illinois 60608, USA. [27]Department of Biostatistics, Columbia University, New York, New York 10032, USA. [28]Hospital Nacional de Niños Dr Saenz Herrera, CCSS, Child Developmental and Behavioral Unit, San José, Costa Rica. [29]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [30]Division of Molecular Genome Analysis, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. [31]Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [32]Institute of Child Health, University College London, London, WC1N 1EH, UK. [33]Department of Clinical Chemistry, Fimlab Laboratories, SF-33100 Tampere, Finland. [34]Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [35]Department of Psychiatry Kaiser Permanente, San Francisco, California 94118, USA. [36]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [37]MRC Centre for Neuropsychiatric Genetics and Genomics, and the Neuroscience and Mental Health Research Institute, Cardiff University, Cardiff, CF24 4HQ, UK. [38]Child and Adolescent Psychiatry Department, Hospital General Universitario Gregorio Marañón, IiSGM, CIBERSAM, Universidad Complutense, 28040 Madrid, Spain. [39]Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, NE2 4HH, UK. [40]Department of Child Psychiatry, University of Tampere and Tampere University Hospital, 33521 Tampere, Finland SF-33101. [41]Department of Preventive Medicine, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [42]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA. [43]Department of Psychiatry, University of California at San Francisco, San Francisco, California 94143–0984, USA. [44]Department of Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy, Translational Brain Medicine in Psychiatry and Neurology, University Hospital RWTH Aachen / JARA Brain Translational Medicine, 52056 Aachen, Germany. [45]Department of Child and Adolescent Mental Health, Great Ormond Street Hospital for Children, National Health Service Foundation Trust, London, WC1N 3JH, UK. [46]Department of Psychiatry and Behavioural Neurosciences, Offord Centre for Child Studies, McMaster University, Hamilton, Ontario L8S 4K1, Canada. [47]Department of Child and Adolescent Psychiatry, Saarland University Hospital, D-66424 Homburg, Germany. [48]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-171 77 Stockholm, Sweden. [49]National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland 20892-9663, USA. [50]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [51]Institute for Molecular Medicine Finland, University of Helsinki, FI-00014 Helsinki, Finland. [52]Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [53]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [54]McLaughlin Centre, University of Toronto, Toronto, Ontario M5S 1A1, Canada. [55]Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia 30322, USA. [56]Center for Human Genetic Research, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [57]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [58]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA.

*Lists of participants appear in the Supplementary Information.