RESEARCH ARTICLE SUMMARY

RNA SPLICING

The human splicing code reveals new insights into the genetic determinants of disease

Hui Y. Xiong,* Babak Alipanahi,* Leo J. Lee,* Hannes Bretschneider, Daniele Merico, Ryan K. C. Yuen, Yimin Hua, Serge Gueroussov, Hamed S. Najafabadi, Timothy R. Hughes, Quaid Morris, Yoseph Barash, Adrian R. Krainer, Nebojsa Jojic, Stephen W. Scherer, Benjamin J. Blencowe, Brendan J. Frey†

INTRODUCTION: Advancing whole-genome precision medicine requires understanding how gene expression is altered by genetic variants, especially those that are far outside of protein-coding regions. We developed a computational technique that scores how strongly genetic variants affect RNA splicing, a critical step in gene expression whose disruption contributes to many diseases, including cancers and neurological disorders. A genome-wide analysis reveals tens of thousands of variants that alter splicing and are enriched with a wide range of known diseases. Our results provide insight into the genetic basis of spinal muscular atrophy, hereditary nonpolyposis colorectal cancer, and autism spectrum disorder.

RATIONALE: We used "deep learning" computer algorithms to derive a computational model that takes as input DNA sequences and applies general rules to predict splicing in human tissues. Given a test variant, which may be up to 300 nucleotides into an intron, our model can be used to compute a score for how much the variant alters splicing. The model is not biased by existing disease annotations or population data and was derived in such a way that it can be used to study diverse diseases and disorders and to determine the consequences of common, rare, and even spontaneous variants.

RESULTS: Our technique is able to accurately classify disease-causing variants and provides insights into the role of aberrant splicing in disease. We scored more than 650,000 DNA variants and found that disease-causing variants have higher scores than common variants and even those associated with disease in genome-wide association studies (GWAS). Our model predicts substantial and unexpected aberrant splicing due to variants within introns and exons, including those far from the splice site. For example, among intronic variants that are more than 30 nucleotides away from any splice site, known disease variants alter splicing nine times as often as common variants; among missense exonic disease variants, those that least affect protein function are more than five times as likely as other variants to alter splicing.

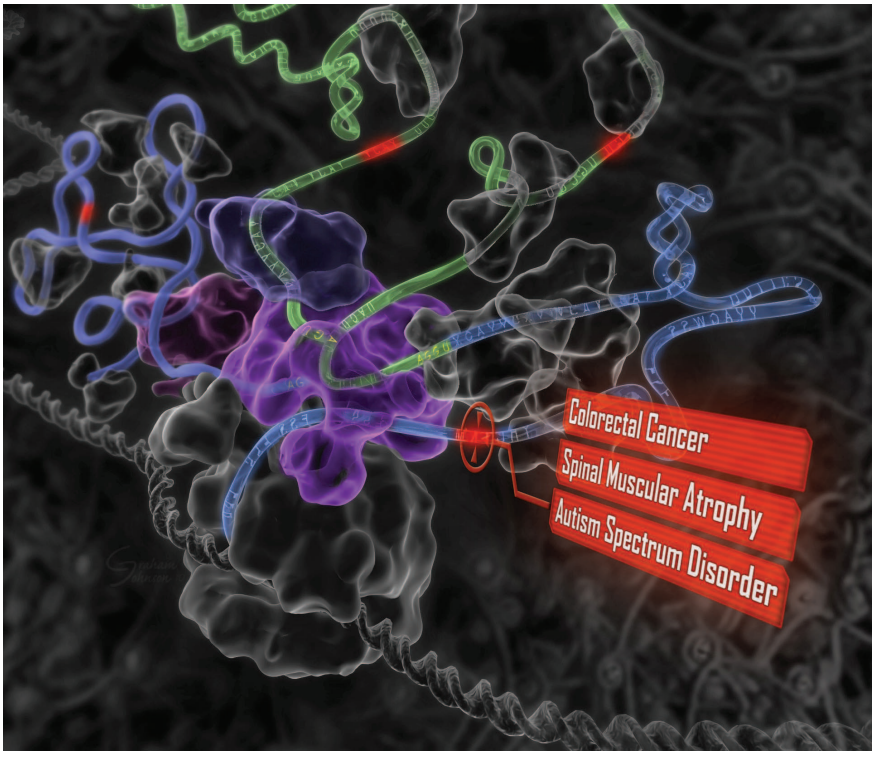
ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.1254806

Autism has been associated with disrupted splicing in brain regions, so we used our method to score variants detected using whole-genome sequencing data from individuals with and without autism. Genes with high-scoring variants include many that have previously been linked with autism, as well as new genes with known neurodevelopmental phenotypes. Most of the high-scoring variants are intronic and cannot be detected by exome analysis techniques.

When we scored clinical variants in spinal muscular atrophy and colorectal cancer genes, up to 94% of variants found to alter splicing using minigene reporters were correctly classified.

CONCLUSION: In the context of precision medicine, causal support for variants independent of existing whole-genome variant studies is greatly needed. Our computational model was trained to predict splicing from DNA sequence alone, without using disease annotations or population data. Consequently, its predictions are independent of and complementary to population data, GWAS, expression-based quantitative trait loci (QTL), and functional annotations of the genome. As such, our technique greatly expands the opportunities for understanding the genetic determinants of disease. ■



"Deep learning" reveals the genetic origins of disease. A computational system mimics the biology of RNA splicing by correlating DNA elements with splicing levels in healthy human tissues. The system can scan DNA and identify damaging genetic variants, including those deep within introns. This procedure has led to insights into the genetics of autism, cancers, and spinal muscular atrophy.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. E-mail: frey@psi.toronto.edu
Cite this article as H. Y. Xiong *et al.*, *Science* **347**, 1254806 (2015). DOI: 10.1126/science.1254806

RESEARCH ARTICLE

RNA SPLICING

The human splicing code reveals new insights into the genetic determinants of disease

Hui Y. Xiong,[1,2,3]* Babak Alipanahi,[1,2,3]* Leo J. Lee,[1,2,3]* Hannes Bretschneider,[1,3,4] Daniele Merico,[5,6,7] Ryan K. C. Yuen,[5,6,7] Yimin Hua,[8] Serge Gueroussov,[2,7] Hamed S. Najafabadi,[1,2,3] Timothy R. Hughes,[2,3,7] Quaid Morris,[1,2,3,7] Yoseph Barash,[1,2,9] Adrian R. Krainer,[8] Nebojsa Jojic,[10] Stephen W. Scherer,[3,5,6,7] Benjamin J. Blencowe,[2,5,7] Brendan J. Frey[1,2,3,4,5,7,10]†

To facilitate precision medicine and whole-genome annotation, we developed a machine-learning technique that scores how strongly genetic variants affect RNA splicing, whose alteration contributes to many diseases. Analysis of more than 650,000 intronic and exonic variants revealed widespread patterns of mutation-driven aberrant splicing. Intronic disease mutations that are more than 30 nucleotides from any splice site alter splicing nine times as often as common variants, and missense exonic disease mutations that have the least impact on protein function are five times as likely as others to alter splicing. We detected tens of thousands of disease-causing mutations, including those involved in cancers and spinal muscular atrophy. Examination of intronic and exonic variants found using whole-genome sequencing of individuals with autism revealed misspliced genes with neurodevelopmental phenotypes. Our approach provides evidence for causal variants and should enable new discoveries in precision medicine.

Regulatory cis elements constitute a substantial portion of the human genome (*1*, *2*) and form the "regulatory code" that directs gene expression, depending on cellular conditions. The development of computational "regulatory models" that can read the code for any gene and predict relative concentrations of transcripts (*3*–*5*) raises the possibility that these models can be used to identify variants that lead to misregulated gene expression and human disease (*6*). Unlike many existing approaches (*7*–*9*), regulatory models do not suffer from the ascertainment biases inherent in databases of disease annotations. Here, we describe a system that uses a regulatory model of splicing to find and score disease mutations (Fig. 1A).

[1]Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario M5S 3G4, Canada. [2]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada. [3]Program on Genetic Networks and Program on Neural Computation & Adaptive Perception, Canadian Institute for Advanced Research, Toronto, Ontario M5G 1Z8, Canada. [4]Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3G4, Canada. [5]McLaughlin Centre, University of Toronto, Toronto, Ontario M5G 0A4, Canada. [6]Centre for Applied Genomics, Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada. [7]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada. [8]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. [9]School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [10]eScience Group, Microsoft Research, Redmond, WA 98052, USA.
*These authors contributed equally to this work. †Corresponding author. E-mail: frey@psi.toronto.edu

A computational model of splicing

Misregulation of splicing contributes substantially to human disease (*10*), so we developed a computational model of splicing regulation that can be applied to any sequence containing a triplet of exons (Fig. 1B). The method extracts DNA sequence features (or cis elements) and, for a given cell type, uses them to predict the percentage of transcripts with the central exon spliced in (Ψ), along with a Bayesian confidence estimate. To train the model, we mined 10,689 exons that displayed evidence of alternative splicing and extracted 1393 sequence features from each exon and its neighboring introns and exons. RNA sequencing (RNA-seq) data from the Illumina Human Body Map 2.0 project (NCBI GSE30611) were used to estimate Ψ for each exon in each of 16 human tissues, and the model was trained to predict Ψ given the tissue type and the sequence features. Unlike existing methods (*3*, *11*, *12*), our computational model was derived using human data, incorporates over 300 new sequence features, and outputs real-valued absolute Ψ values for individual tissues, rather than categorical Ψ values for tissue differences (*13*).

We observed good agreement (R^2 = 0.65) between code-predicted Ψ and RNA-seq–assessed Ψ for exons that were held out during training (Fig. 1C). On the task of classifying high ($\Psi \geq 67\%$) versus low ($\Psi \leq 33\%$) inclusion, the area under the receiver-operator characteristic curve (AUC) is 95.5%. For quality control, we only examined exon-tissue combinations (n = 56,784) for which the standard deviation of the RNA-seq–assessed Ψ was less than 10%, and cross-validation was used to ensure that test cases were not used during training (*13*) (table S3). The prediction accuracy was even higher (R^2 = 0.94, AUC = 99.1%) for the 50% of predictions with highest confidence (n = 28,392). The model is robust and accurate for categories of data that were not included during training, including genes with low expression, genes from excluded chromosomes, tissue differences in splicing levels, tissues from independent sources, and splicing levels quantified by reverse transcription polymerase chain reaction (RT-PCR) (*13*).

We next investigated whether our computational model accounts for the effects of known RNA-binding proteins (RBPs), which are key splicing regulators. We compared how well the calculated RBP binding affinity from Ray *et al.* (*14*) correlated with the observed variation in splicing and found 2080 strong correlations ($P < 0.01$, multiple hypothesis-corrected permutation test). Then we correlated the RBP binding affinities with the residual splicing activity not captured by the code, which was obtained by subtracting the code predictions from the observed values. The number of strong correlations dropped to 60, which suggests that our computational model mostly encompasses the collective effects of known RBPs (Fig. 2) (*13*).

Our model also accounts for the effects of disruptions in trans-acting factors. We examined knockdown data for Muscleblind-like (MBNL) RBPs in HeLa cells (*15*). There were 664 exons that exhibited a significant change in RNA-seq–assessed Ψ upon MBNL knockdown, as well as 26,457 exons whose levels did not change significantly upon knockdown. When we scored exons according to how much the model predicted that Ψ would change when the MBNL features were removed in silico, we found that MBNL-regulated exons frequently had higher scores [$P = 6.2 \times 10^{-57}$, Kolmogorov-Smirnov (KS) test, 31.4%]. The computational model predicted the effects of MBNL knockdown more accurately than direct examination of MBNL binding sites [10.9% improvement in the AUC; $P = 1.4 \times 10^{-14}$, bootstrap test (*13*)].

In contrast to correlation-based linear methods, where sequence features act independently, our computational model incorporates crucial context-dependent effects. When we derived tissue-specific linear models by searching over the most predictive set of sequence features, they always accounted for significantly less data variance (R^2 < 0.49) than our context-dependent model (R^2 = 0.65). We found that in our model, the same feature can influence Ψ differently in different cis contexts established by other sequence features and in different trans contexts specified by cell type (*13*) (figs. S14 and S15). For instance, 40 of the 100 most strongly predictive sequence features frequently switched

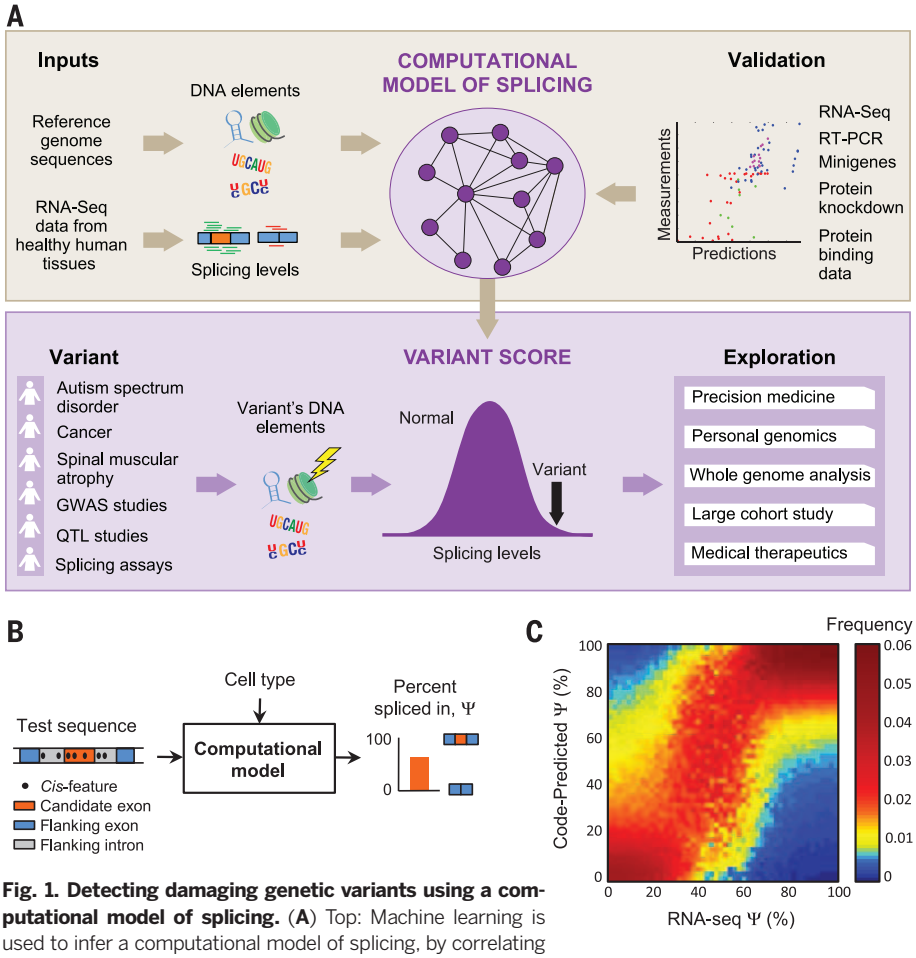


Fig. 1. Detecting damaging genetic variants using a computational model of splicing. (**A**) Top: Machine learning is used to infer a computational model of splicing, by correlating DNA elements with splicing levels in healthy human tissues. Bottom: Genetic variants arising from a wide array of diseases and technologies can be detected and filtered using the computational model, enabling explorations into the genetics of disease. (**B**) For a given cell type, the computational model extracts the regulatory code from a test DNA sequence and predicts the percentage of transcripts with the exon spliced in, Ψ. (**C**) Predictions are made for 10,689 test exons profiled in 16 tissues; exons and tissues are binned according to their RNA-seq–assessed values of Ψ, and for each bin (column) the distribution of code-predicted Ψ is plotted (n = 56,104).

the direction of their effect in at least one tissue, depending on cis context.

We wondered whether our computational model could accurately predict differences in splicing levels between individuals using only their genotype. We analyzed genotype and RNA-seq data for lymphoblastoid cell lines from four individuals (*16*) and used our model to predict Ψ in white blood cells, for pairs of individuals that have differing SNPs (*13*). When we examined 99 exons that exhibited a significant difference in RNA-seq–assessed Ψ between pairs of individuals and whose predicted difference in Ψ was above a noise threshold, we found that our technique correctly predicted the direction of change in 73% of cases ($P = 3.5 \times 10^{-6}$, binomial test).

Genome-wide analysis of splicing misregulation and disease

To assess the implications of genetic variation for splicing regulation, we mapped 658,420 single-nucleotide variations (SNVs) to exonic and intronic sequences containing the regulatory code for ~120,000 exons in ~16,000 genes (*13*). Of these SNVs, 543,525 are single-nucleotide polymorphisms (SNPs), which are common (minor allele frequency or MAF > 1%) (*17*), whereas 114,895 have been linked to diseases and are mostly rare (MAF < 1%) (*18*). To score the effect of every SNV on splicing regulation, we applied the regulatory model to the sequence with and without the SNV and computed the difference in predicted splicing level, ΔΨ, for each tissue (Fig. 3A). We studied the effects of SNVs using the largest value of ΔΨ across tissues, as well as a "regulatory score" that aggregates the magnitude of ΔΨ across tissues (*13*).

The code provides an unprecedented view of the impact of SNVs on splicing regulation (Fig. 3B). It reveals 20,813 unique SNVs that disrupt splicing ($|\Delta\Psi| \geq 5\%$; table S4), frequently in a way that depends on cis context (*13*) (fig. S21). Diverse methods of validation support the functional impact of these disruptions. Intronic SNVs that are close to splice sites frequently cause misregulation, but 465 intronic SNVs that are more than 30 nucleotides (nt) from any splice site also induce substantive changes. Within exons, we found that significant deviations are induced by 9525 nonsense SNVs and 1273 missense SNVs but also by 579 synonymous SNVs—a result supported by recent data showing that synonymous mutations frequently contribute to human cancer (*19*).

To explore the causal implications of high-scoring SNVs in the context of disease, we examined whether disease SNVs are predicted to disrupt splicing ($|\Delta\Psi| \geq 5\%$) more frequently than common SNPs, of which a large portion are thought to be under neutral selection (*20*). We plotted the locations and ΔΨ for 81,608 disease SNVs located up to 100 nt into exons or up to 300 nt into their adjacent introns (Fig. 3C).

Our technique reveals widespread processes whereby disease SNVs cause misregulation of splicing. Databases of disease annotations were not used to train our model, so it is not susceptible to overfitting already discovered disease SNVs or inherent ascertainment biases (*7*–*9*).

We found that intronic disease SNVs that are more than 30 nt from any splice site are 9.0 times as likely to disrupt splicing regulation relative to common SNPs in the same region ($P = 5.1 \times 10^{-68}$, two-sample t test, n = 1639 and n = 24,535). Within exons, synonymous disease SNVs are on average 9.3 times as likely as synonymous SNPs to disrupt splicing regulation ($P = 8.0 \times 10^{-116}$, two-sample t test, n = 2652 and n = 4510).

Missense SNVs have previously been examined mainly in the context of how they alter protein function (*7*). Our method enables the exploration of their effects on splicing regulation. We found that missense disease SNVs are not more likely to disrupt splicing than missense SNPs (P = 0.22, two-sample t test, n = 58,918 and n = 2981), which contradicts previously published evidence that they do ($P \approx 0.05$) (*9*). However, when we examined 789 and 1757 missense disease SNVs that minimally and maximally alter protein function as indicated by Condel (*21*) analysis, we found that SNVs that minimally alter protein function are on average 5.6 times as likely to disrupt splicing regulation ($P = 4.5 \times 10^{-14}$, two-sample t test), elucidating a "disease by misregulation" mechanism (*13*).

We found that within introns, the regulatory scores of 457 SNPs that were implicated in genome-wide association studies (GWAS) and that map to regulatory regions (*22*) are quite similar to non-GWAS SNPs (P = 0.27, KS test, n = 262,804), whereas the scores of disease SNVs are significantly higher ($P < 1 \times 10^{-320}$, KS test, 71.2%, n = 280,638). Fewer than 5% of GWAS SNPs are estimated to cause misregulation in a fashion similar to disease SNVs (*13*), indicating that our method can detect disease SNVs that are not detectable by GWAS (Fig. 4A). In further support of the functional specificity of our approach, we found that the regulatory scores of disease SNVs with

strong experimental evidence are substantially higher than those with weak or indirect evidence (Fig. 4B).

Next, we used the computational model to analyze three human diseases with different characteristics: spinal muscular atrophy (autosomal-recessive single gene), nonpolyposis colorectal cancer (oligogenic), and autism spectrum disorder (multigenic).

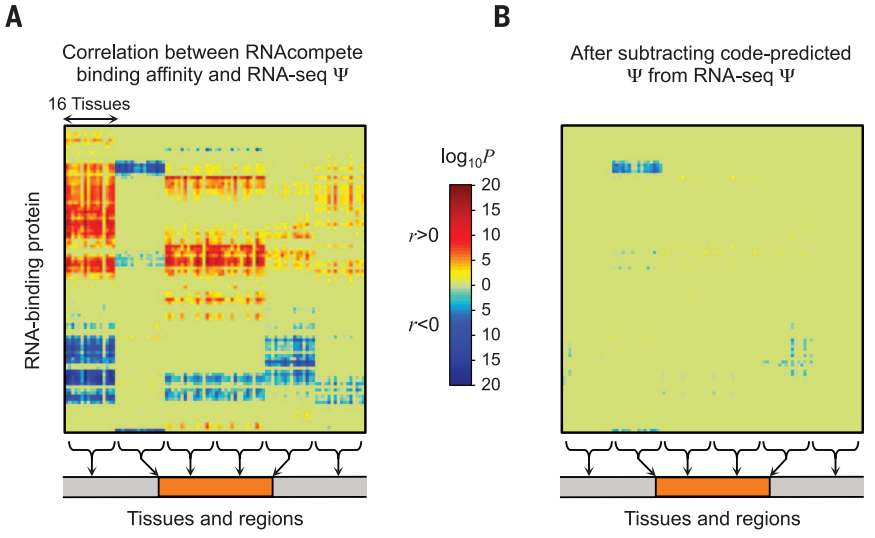


Fig. 2. Accounting for RNA-binding proteins (RBPs). (**A**) Correlations between RNA-seq Ψ and the affinities of RBPs assayed in 98 in vitro experiments (*14*). (**B**) When code-predicted Ψ values are subtracted from RNA-seq–assessed values of Ψ, their correlations with the binding affinities mostly vanish.

Spinal muscular atrophy (SMA)

To explore misregulation of *SMN1/2*, which is associated with SMA, a leading cause of infant mortality (*23*), we used the computational model to simulate the effects of more than 700 known and novel mutations around exon 7 in *SMN1/2*. We first examined the regulatory consequences of four nucleotides that differ between *SMN1* and *SMN2*, labeled C6T, G-44A, A100G, and A215G in Fig. 5A, where "-44" indicates 44 nt upstream of the 3′ splice site. These substitutions are known to lead to decreased inclusion of exon 7 in *SMN2* and loss of function.

Our method predicts that exon 7 skipping is predominantly caused by C6T and to a much lesser degree by G-44A, whereas A100G and A215G are predicted not to have a significant impact on splicing. The prediction for C6T is consistent with previously published mutagenesis data (*23*). Mutagenesis data indicate that A100G enhances skipping by 36% to 63% (*24*) in the *SMN2* context. Using a Z-score threshold of 1, our computational model also predicts a small but significant skipping effect of A100G

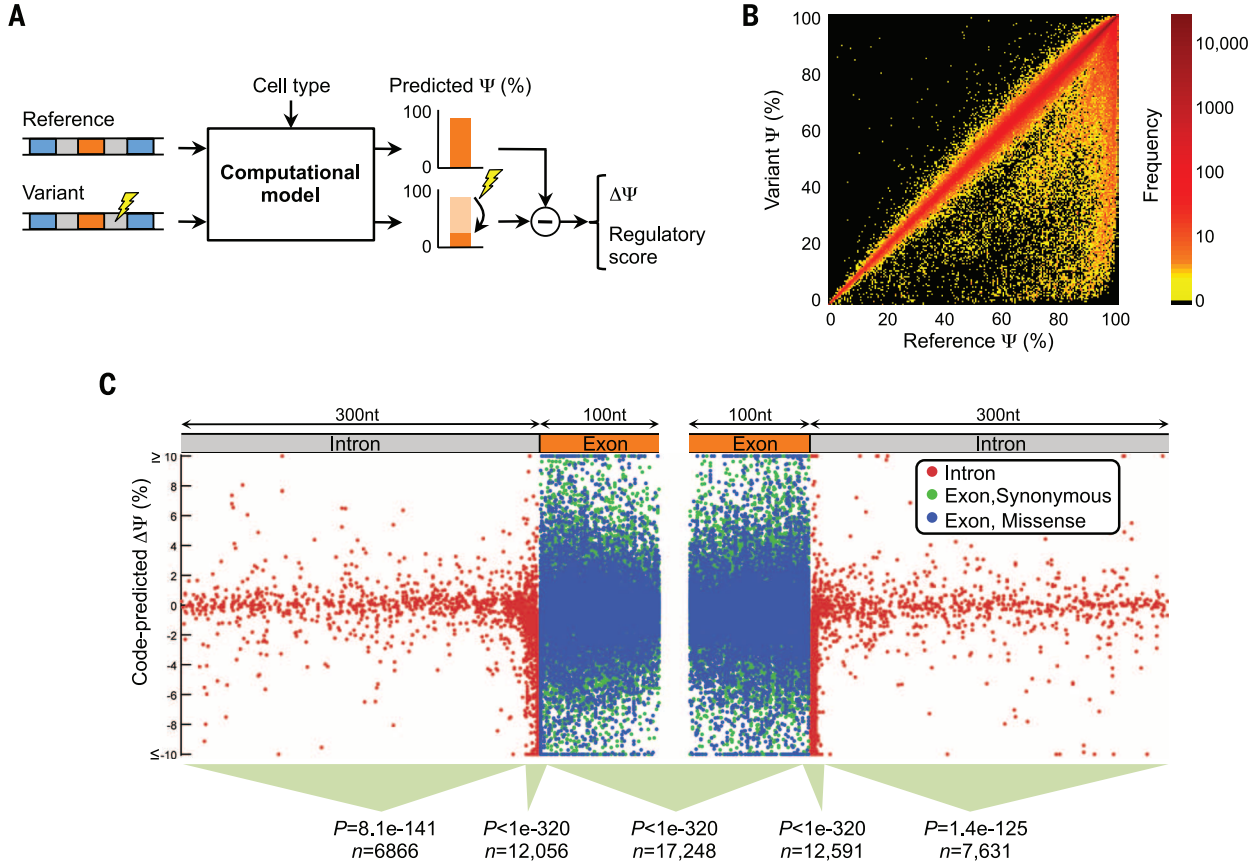


Fig. 3. Genome-wide analysis of genetic variations. (**A**) To assess the effect of a single-nucleotide variation (SNV), the computational model is applied to the reference sequence and the variant. Then, the maximum difference ΔΨ across tissues is computed, along with a "regulatory score" that also accounts for prediction confidence (*13*). (**B**) The effect on Ψ of 658,420 intronic and exonic SNVs. (**C**) Locations and predicted ΔΨ of 81,608 disease-annotated intronic SNVs and synonymous or missense exonic SNVs. In different sequence regions, the scores of disease SNVs tend to be larger than those of SNPs (Ansari-Bradley tests for equal dispersion; *n* includes both types).

in the *SMN2* context. We used minigene reporters to test our predictions and found that in all cases they are supported by the experimental data, including the negligible effect of A100G mutation in the *SMN1* context (Fig. 5B, red). Further, our prediction for G-44A is consistent with antisense oligonucleotide experiments indicating that it overlaps with a splicing suppressor (*25*).

To explore mutations that may result in gain of *SMN2* function, we simulated the regulatory effects of all 420 possible point mutations in 140 nt of intronic sequence upstream of exon 7 (Fig. 5B). Minigene reporter data for the top three predictions confirm that none of them exhibit decreased inclusion and two of them cause increased inclusion (Fig. 5, B and C, green). Together, the predictions for *SMN1* and *SMN2* mutations (Fig. 5C) have a Spearman correlation of 0.82 with the experimental data ($P = 0.017$, $n = 7$, one-sided permutation test).

We generated a literature-curated compendium of mutagenesis data for 85 variations located in three exonic regulatory regions previously tested using in vivo selection, plus an intronic region. When our model is used to predict ΔΨ for these cases (Fig. 5D), the direction of regulation is correct in 85% of cases and the Spearman correlation is 0.74 ($P = 5.7 \times 10^{-16}$, one-sided permutation test). We additionally used our method to simulate ΔΨ for 101 mutants selected in vivo to increase Ψ, with point mutations in the first 6 nt in exon 7 and also in the entire exon (*23*). Increases in Ψ are correctly predicted in 98.7% of the 78 high-confidence cases (table S6).

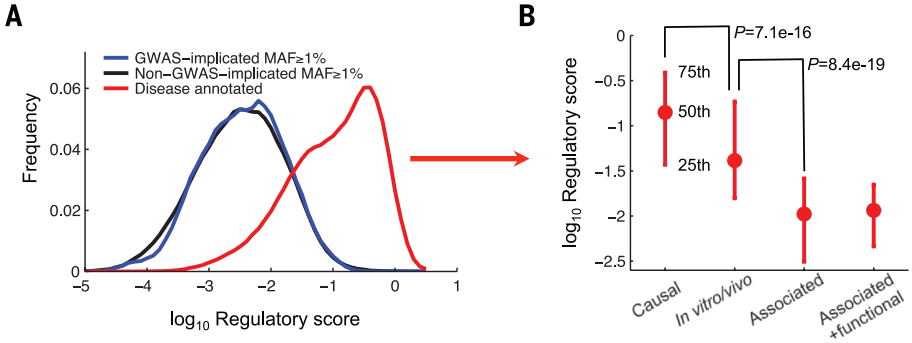


Fig. 4. Regulatory scores of GWAS SNPs. (**A**) Distributions of regulatory scores for GWAS-implicated SNPs ($n = 457$), non–GWAS-implicated SNPs ($n = 262{,}347$), and disease SNVs ($n = 18{,}291$) in introns. (**B**) Regulatory scores of disease-annotated intronic SNVs that are causal ($n = 17{,}631$), supported by in vitro and in vivo data ($n = 224$), only associated ($n = 324$), or associated but have additional functional evidence ($n = 112$). *P* values (*t* test) are indicated.

Nonpolyposis colorectal cancer

Lynch syndrome, or hereditary nonpolyposis colorectal cancer, accounts for ~3% of colorectal cancer cases (*26*), and nearly 90% of reported variations occur in the DNA mismatch repair genes *MLH1* and *MSH2* (*27*). Numerous studies have shown that misregulation of splicing accounts for a major portion of cases (*28*) but also that existing computational predictions for variations that do not directly disrupt splice sites

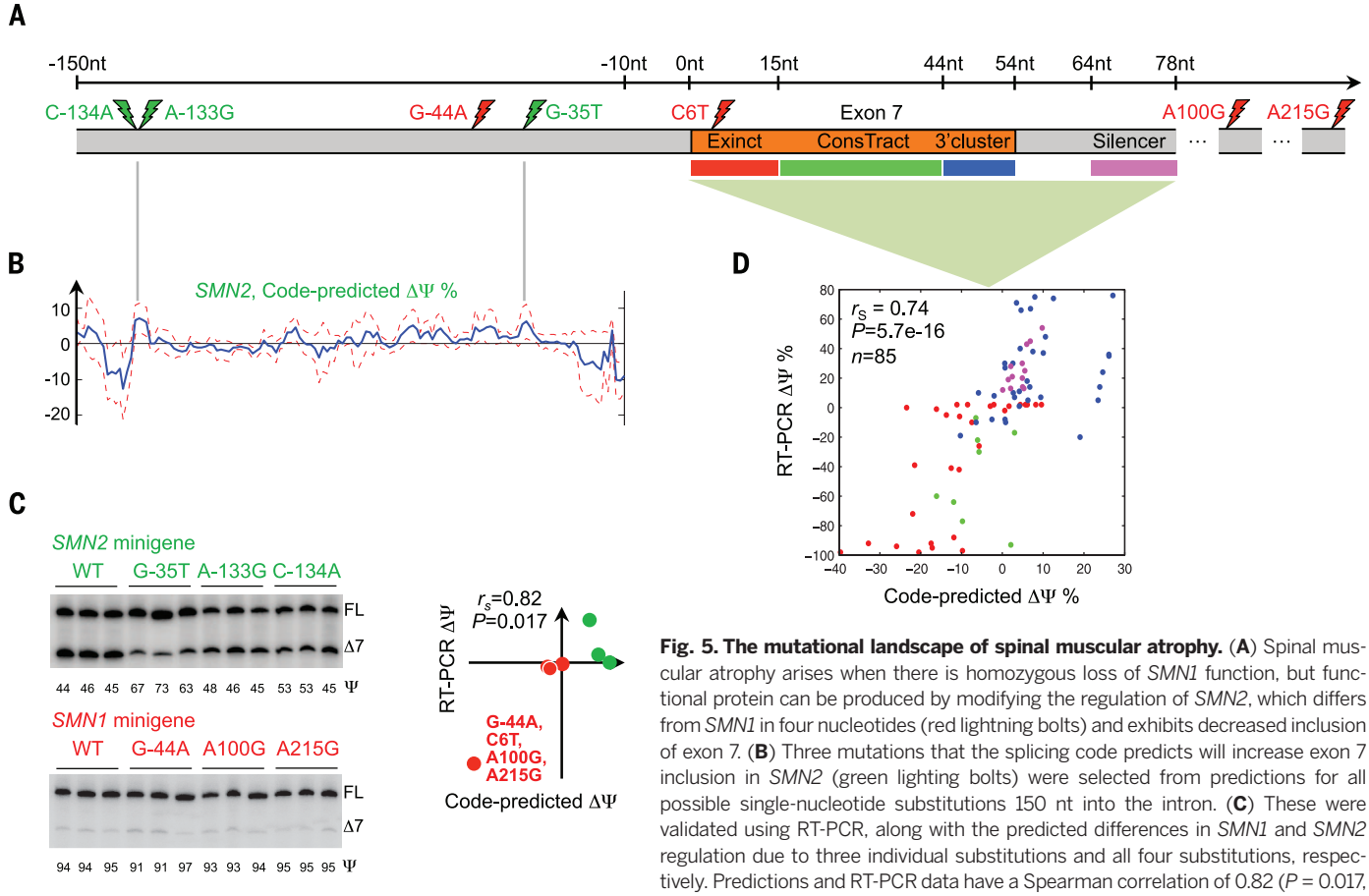


Fig. 5. The mutational landscape of spinal muscular atrophy. (**A**) Spinal muscular atrophy arises when there is homozygous loss of *SMN1* function, but functional protein can be produced by modifying the regulation of *SMN2*, which differs from *SMN1* in four nucleotides (red lightning bolts) and exhibits decreased inclusion of exon 7. (**B**) Three mutations that the splicing code predicts will increase exon 7 inclusion in *SMN2* (green lighting bolts) were selected from predictions for all possible single-nucleotide substitutions 150 nt into the intron. (**C**) These were validated using RT-PCR, along with the predicted differences in *SMN1* and *SMN2* regulation due to three individual substitutions and all four substitutions, respectively. Predictions and RT-PCR data have a Spearman correlation of 0.82 ($P = 0.017$, one-sided permutation test). (**D**) Predicted ΔΨ values for 85 individual mutations located in four regions are plotted against RT-PCR–assessed values; the Spearman correlation is 0.74 ($P = 5.7 \times 10^{-16}$, one-sided permutation test).

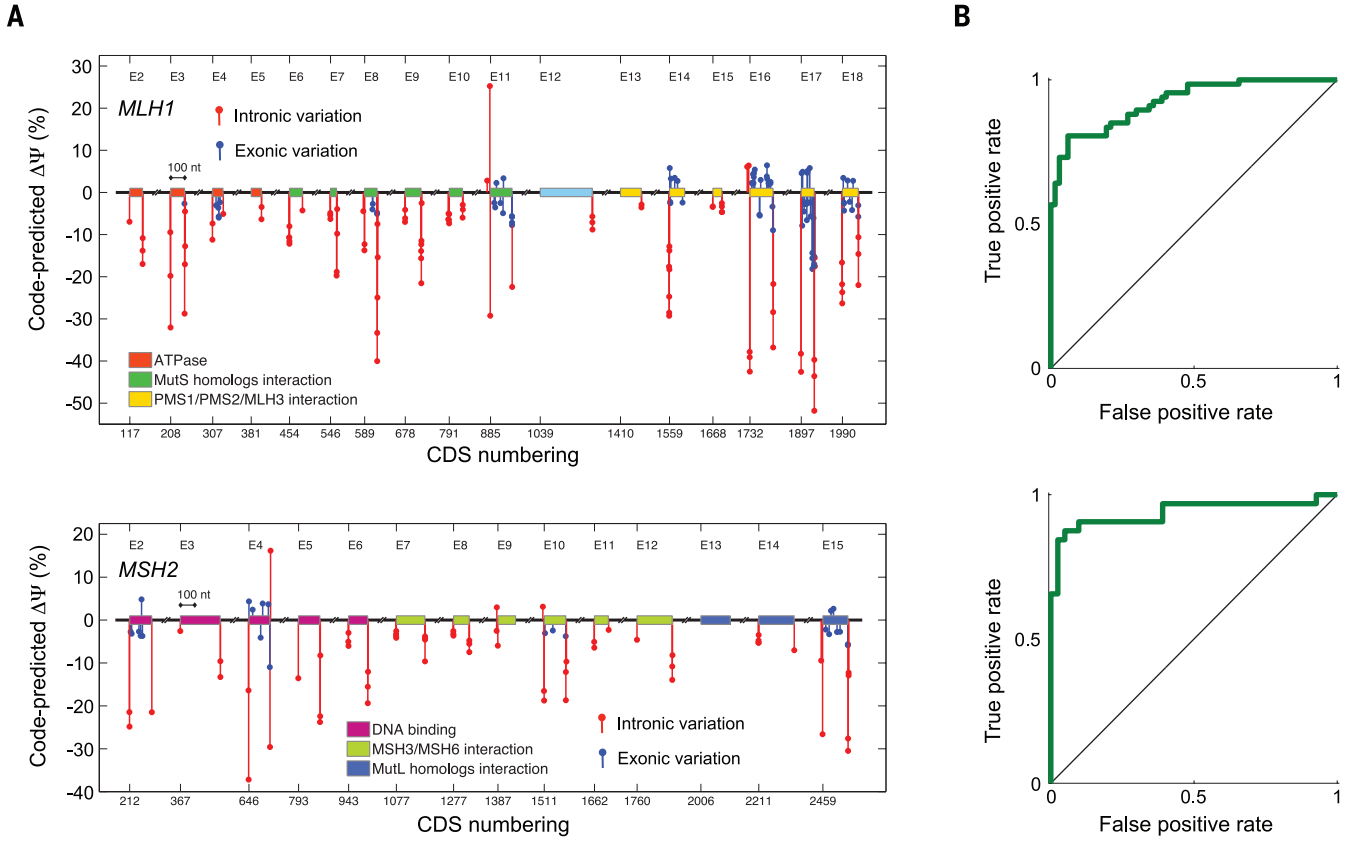


Fig. 6. The mutational landscape of nonpolyposis colorectal cancer. (**A**) Predicted ΔΨ for mutations in *MLH1* and *MSH2* arising in patients with nonpolyposis colorectal cancer, or Lynch syndrome. Coding sequence (CDS) numbering is based on GenBank NM_000249.3 and NM_000251.2 and starts at A of the ATG translation initiation codon. (**B**) Validation using 134 *MLH1* variations tested by RT-PCR (AUC = 92.4%, $P = 2.8 \times 10^{-24}$, one-sided permutation test) and 73 *MSH2* variations (AUC = 93.8%, $P = 8.7 \times 10^{-15}$, one-sided permutation test).

are not correlated with experimental data (*28*, *29*). It has been suggested that this is because existing tools do not take interactions between regulatory features into account (*29*).

We evaluated 977 SNVs, 156 of which are nonsense, in *MLH1* and *MSH2* (*27*) using our computational model and found that high levels of misregulation are predicted (Fig. 6A and tables S7 and S8) (*13*): 32.3% of SNVs exhibited a ΔΨ that was larger than that of 95% of common SNVs ($P = 4.2 \times 10^{-135}$, one-sided binomial test). To avoid bias, we excluded *MLH1*, *MSH2*, and their variants during model training. Additionally, the majority of predictions are concordant with published RT-PCR data (tables S9 and S10). When predicted ΔΨ was used to classify increased skipping versus no change for SNVs where RT-PCR data were available, AUCs of 92.4% and 93.8% (Fig. 6B) were achieved for 134 *MLH1* and 73 *MSH2* variants [$P = 2.8 \times 10^{-24}$ and $P = 8.7 \times 10^{-15}$, one-sided permutation tests (*13*)].

To further test the specificity of our method, we mapped 80 common SNPs to *MLH1* and *MSH2* and compared their regulatory scores to those of the SNVs found in patients. Common SNPs had significantly lower scores ($P = 8.1 \times 10^{-11}$, KS test, 40.0%, $n = 1058$), indicating that our method successfully detects causal variants (*13*).

Our method sheds light on unresolved hypotheses for the mechanisms of specific mutations. Three missense substitutions in the second nucleotide of codon 659 in exon 17 of *MLH1* are observed in Lynch syndrome patients: c.1976G>T, c.1976G>C, and c.1976G>A. Evidence indicates that c.1976G>A likely does not change protein function, which suggests that the mechanism is splicing misregulation (*30*–*32*). Indeed, RT-PCR data indicate that c.1976G>T and c.1976G>C induce increased exon skipping (*30*). However, previous computational analyses either fail to predict misregulation (*31*) or, because the mutations increase the strength of an exonic splicing enhancer, erroneously predict increased exon inclusion (*13*, *33*). We applied our computational model and found that it confidently and correctly predicts increased skipping in all three cases (table S10) and also correctly predicts that c.1976G>C has a stronger effect than c.1976G>T. We can thus hypothesize that c.1976G>A induces aberrant splicing and renders the translated protein dysfunctional.

Autism spectrum disorder (ASD)

ASD is a neurodevelopmental condition characterized by language deficiency, restricted and repetitive interests, and challenges in social skills. It is highly heritable, but its substantial clinical and genetic heterogeneity has complicated the identification of all etiologic genetic variants (*34*). Through the study of rare genetic variants, ~100 genes have now been implicated in ASD (*35*), and these are estimated to account for ~20% of the etiologic cause in different cohorts examined (*36*, *37*). More recent studies using whole-genome sequencing revealed higher yields of contributing mutations, but these studies have focused only on exonic regions (*38*). Common genetic variants may also have an effect in ASD, but few studies replicate the same loci (*39*). Splicing misregulation as a cause of ASD is evidenced by examples of genes involved in ASD, such as neurexins and neuroligins, that are extensively alternatively spliced (*40*), as well as by recent transcriptomic analyses showing consistent deviations in alternative splicing patterns in the cortical regions of ASD cases (*41*).

To identify genes with SNVs that potentially cause splicing misregulation in ASD cases, we used our regulatory model to analyze the genomes of five idiopathic ASD cases, which do not have ASD-associated cytogenetic markers such as chromosome 15q duplication (*13*). We sequenced these genomes using brain samples from the Autism Tissue Program (*42*) and selected the genomes of 12 controls consisting of three subgroups of four controls each. As a control, we clustered the ASD and control genomes using genome-wide genetic similarity and verified that they cluster by ethnic group but not by disease condition or other covariates; this

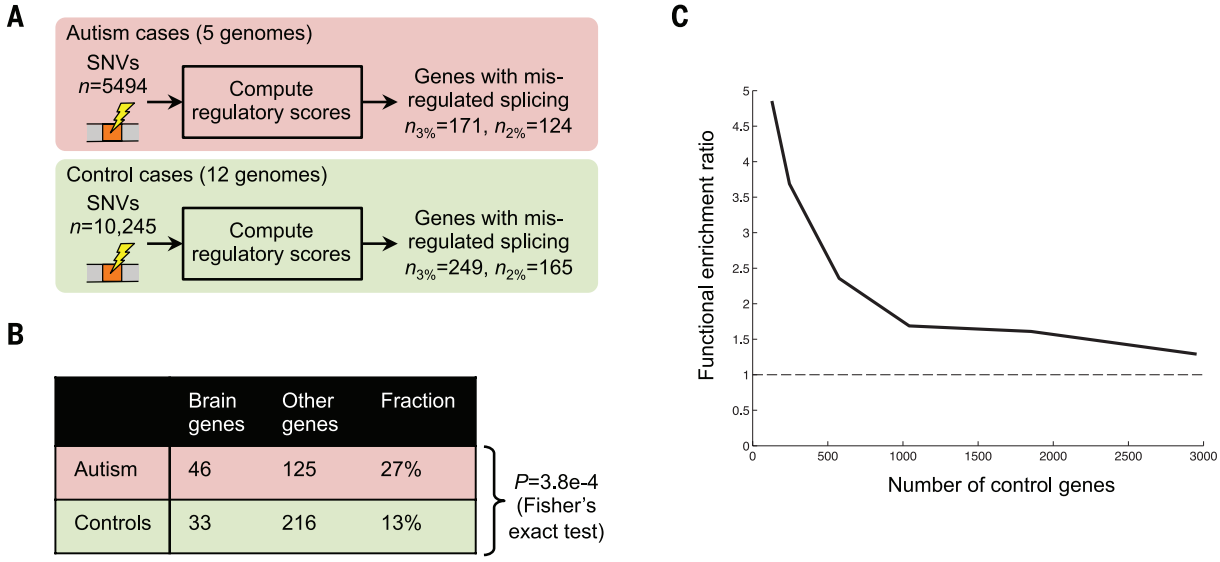


Fig. 7. Splicing misregulation in individuals with autism. (**A**) Genes containing at least one SNV that the computational model predicts will cause decreased exon inclusion were identified in five autism spectrum disorder (ASD) cases and 12 controls by thresholding $\Delta\Psi$ using either the 2nd or 3rd percentile of $\Delta\Psi$ for SNPs. (**B**) Genes that our method predicts are misregulated in ASD cases more frequently have high expression in brain tissues than in control cases. (**C**) The effect of varying the threshold on $\Delta\Psi$, and thus the number of case and control genes, on the odds ratio for the enrichment of central nervous system development genes (GO:0007417); in all cases, $P < 0.05$.

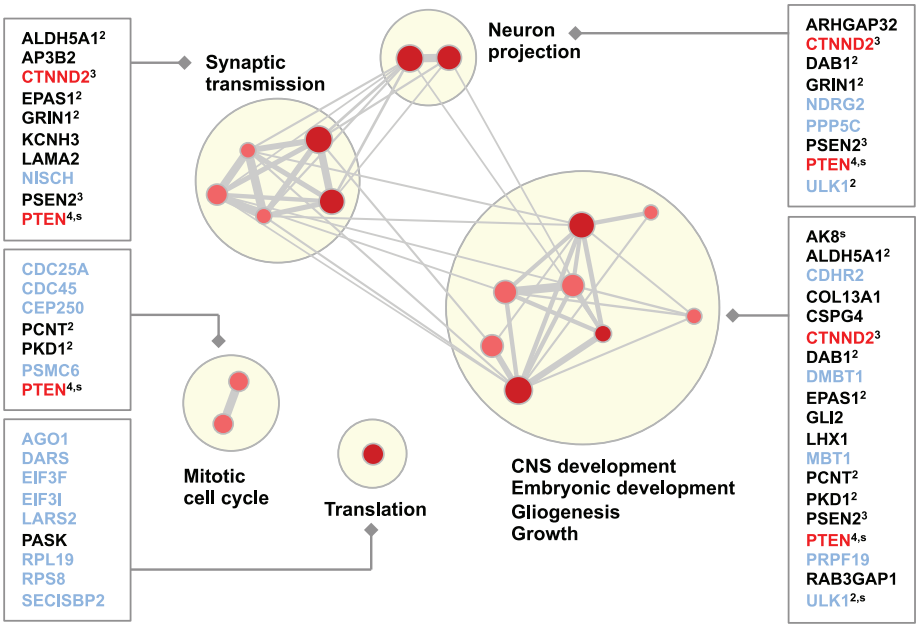


Fig. 8. Misregulated genes and functional categories enriched in individuals with autism. Gene Ontology and pathway categories that are enriched ($P \leq 0.01$, Fisher's exact test) in misregulated genes from ASD cases relative to controls were identified ($n = 18$), along with the corresponding set of genes from ASD cases. Each gene set is shown as a red or pink dot, depending on whether the 2nd- or 3rd-percentile threshold was used for detection (Fig. 7A), and size is proportional to the number of genes in the set. Edge thickness indicates the fraction of genes shared between two sets. Groups of functionally related gene sets are highlighted by blond discs. The names of novel genes that are not already implicated in ASD and have neural-related phenotypes are shown in black, the names of genes already implicated in ASD are in red, and other gene names are in pale blue. If a gene is in multiple categories, the number of categories is shown in superscript; genes in which a stop codon is introduced by the SNV are labeled "s."

result indicates that the ASD and control SNVs are not grossly biased by nondisease effects (*13*) (fig. S27).

The genomes of cases and controls were scanned for SNVs (*13*) and, to focus our analysis on rare variants, we retained only high-quality homozygous and heterozygous reference SNVs (in which one allele matches the reference allele) that did not correspond to common SNPs. This resulted in a median of ~42,000 SNVs per subject.

We examined genes with high expression in brain tissues, which are more frequently implicated in ASD, and did not find an enrichment of SNVs in ASD cases versus controls [$P = 0.24$, Fisher's exact test (*13*)]. Aiming to separate causal SNVs from noncausal ones, we identified SNVs that our technique predicts will cause splicing misregulation (Fig. 7A). All variants were mapped onto the splicing code within canonical Ensembl transcripts, resulting in 15,739 SNVs, whose code-predicted $\Delta\Psi$s were then computed (table S13). We identified genes with misregulated splicing in cases and also in controls by applying a threshold to $\Delta\Psi$ equal to the 2nd and also the 3rd percentile of $\Delta\Psi$ for common SNPs (Fig. 7B) (*13*), and genes misregulated in both cases and controls were removed from further analysis.

Among genes that our technique predicts are misregulated in ASD cases ($n = 171$), 27% have high expression in brain, whereas for controls ($n = 249$), only 13% have high expression in brain ($P = 3.8 \times 10^{-4}$, Fisher's exact test). When we examined genes with low or no expression in brain tissues, we did not observe significant differences (*13*). Further, when we made the threshold used to identify misregulated genes more stringent, we found that enrichment of ASD-related functions was amplified (Fig. 7C). These results open the door to discovering new genetic determinants of ASD and also suggest that more generally, our splicing model can be used to sift through variants to support precision medicine and whole-genome variant studies.

We tested Gene Ontology annotation and pathway-based gene sets for enrichment in misregulated genes; to account for biases such as gene length, we tested the gene enrichment in ASD genomes relative to control genomes. Interestingly, we found categories related to synaptic transmission and to neuron projection and growth (Fig. 8). Gene permutation analysis shows that enrichment in neurodevelopmental gene sets is significant (empirical false discovery rate < 4%). In addition, repeating the analysis for a subset of control genomes versus another subset of control genomes did not produce any significant results, and top-ranking gene sets were not neurodevelopmental.

We found 39 genes with predicted splicing alterations that are associated with at least one enriched function, and we additionally prioritized 19 of these genes as more compelling ASD disease candidates because they are known to have neurological, neurobehavioral, or neurodevelopmental phenotypes in human [Human Phenotype Ontology (HPO) and Online Mendelian Inheritance in Man (OMIM)] or mouse [Mouse Genomics Informatics/Mammalian Phenotype Ontology (MGI/MPO)] (table S16). The analysis reveals interesting candidates, and only *CTNND2* and *PTEN* have been previously implicated or suggested to play a role in ASD (*35*, *43*). Our study suggests new candidate ASD genes, including *ALDH5A1*, *GLI2*, *GRIN1*, *KCNH3*, *LAMA2*, and *NISCH*, in addition to other possibilities. Our results are robust to choices made in the analysis (*13*) and can be combined with other approaches [e.g., (*44*)] to develop diagnostic techniques.

Discussion

Our results from profiling the genome-wide effects of more than 650,000 SNVs shed light on how genetic variation affects splicing. Further, our in-depth results from the analysis of thousands of variations in diverse disorders, including spinal muscular atrophy, nonpolyposis colorectal cancer, and autism, exemplify the wide range of applicability of our technique and provide insights into the genetic determinants of these diseases.

In the context of precision medicine, the importance of providing causal evidence for putative variants with the goal of avoiding the effects of confounding factors, such as population stratification, has recently been underscored (*45*, *46*). The ability of our computational technique [SPANR (splicing-based analysis of variants); see (*13*) and http://tools.genes.toronto.edu] to provide regulatory evidence for a variant's disruptiveness is supported by accurate predictions for test sequences that were not used during training, discrimination of disease variants even though the model was not trained using disease labels, and strong correlation between code-predicted changes in splicing induced by mutations and experimental data using minigene reporters.

Our approach contrasts with techniques that use functional annotations of the genome (*2*, *8*, *47*), tools that are trained using existing disease annotations and thus suffer from overfitting to known mutations or severe selection bias (*7*–*9*, *48*, *49*); GWAS (*50*, *51*); and expression-based quantitative trait loci (QTL) (*16*, *52*). To compare our method with using functional genome annotations, we removed missense exonic SNVs that may affect phenotype without changing splicing regulation, yielding 26,403 SNVs that map to canonical Ensembl transcripts. At a false positive rate of 0.1%, we found that scoring SNVs by their overlap with functional annotations detects 1.4% of disease variants, whereas our method is 25 times as sensitive and detects 35.9% of disease variants (*13*).

Relative to state-of-the-art methods that examine perturbations of motifs and genome annotations but do not account for changes in gene regulation (*48*, *49*), our method is nearly 10 times as sensitive in each of several sequence regions (fig. S18). Our technique does not directly detect variants associated with a phenotype of interest. However, when it is combined with phenotype-matched genotype data such as those generated by whole-genome sequencing, it can detect variants relevant to phenotype, as demonstrated by our autism analysis.

In contrast to GWAS (*50*), splicing QTL analysis (*52*), and other methods that use allele frequencies within populations to score variants (*47*), our technique does not directly depend on allele frequencies. As demonstrated above, our method can reliably detect rare and even spontaneous disease variants. To provide evidence that our method is not dependent on allele frequency, we separately analyzed rare variants (0.1% < MAF < 1%), moderately common variants (1% < MAF < 5%), and disease variants [annotated in the Human Gene Mutation Database (HGMD), mostly rare]. We found that the disease variants have regulatory scores significantly different from those of the rare and common variants, but the distribution of regulatory scores is indistinguishable for rare and common variants (*13*). Furthermore, when we examined 15,386 disease variants and 1519 common SNPs within intronic regions with moderate to high conservation across vertebrates (PhastCons score > 0.5), we found that our method more accurately detects disease variants ($P < 1 \times 10^{-320}$, KS test, 60.1%) than scoring them using conservation ($P = 2.2 \times 10^{-166}$, KS test, 38.2%).

Our approach can be combined with population-based methods so as to amplify their specificity and identify causal variants in the context of specific diseases, either by providing more refined scores or by scoring variants in the same linkage disequilibrium block as a GWAS- or QTL-identified noncausal SNP. When we evaluated 453 splicing QTLs that were identified using blood samples and the genotypes of 922 individuals (*52*), we found that a subset of splicing QTLs had high regulatory scores, as computed using our method, relative to those of common SNPs in general ($P = 4.2 \times 10^{-10}$, KS test, 15.4%).

Potential sources of prediction error include unaccounted-for RNA features, inaccuracies in computed features, imperfect modeling of splicing levels, and limitations due to a focus on cassette splicing. Even so, the method described here performs well, as assessed both by validation of splicing prediction using several diverse sources of data and by its ability to detect disease mutations.

We anticipate that it will be important to seek regulatory models that encompass other major steps in gene regulation, including chromatin dynamics, transcription, polyadenylation, mRNA turnover, protein synthesis, and protein stabilization. These processes influence transcript levels in a highly integrated manner within the cell, so modeling them jointly should lead to more accurate predictions. Moreover, evidence suggests that DNA elements previously thought to be pertinent to only one regulatory process may in fact span several steps in the regulatory chain. Examples include nucleosome positioning, epigenetic modifications, and chromatin interactions (*53*).

Materials and methods

Details of all data sets, learning algorithms, statistical analyses, experimental validation, and Web tool implementation are provided in the supplementary materials. In brief, the human splicing code was assembled using 1393 carefully designed sequence features extracted from each of the 10,689 alternatively spliced exons and their corresponding Ψ values profiled in 16 normal tissues from human BodyMap 2.0 (NCBI GSE30611) RNA-seq data. The features of an exon were extracted from its proximal genomic sequences, including exon and intron lengths, splice site signals, counts of splicing factor motifs, trinucleotide frequencies, retrovirus repeats, nucleosome positioning, RNA secondary structures, etc. The computational model was learned using a Bayesian deep learning algorithm, with extreme care exercised to prevent overfitting. Because the model was built using the reference genome only, its performance was first validated using held-out data, including additional RNA-seq (*54*), RT-PCR, RBP binding (*14*), and MBNL knockdown (*15*) data sets. The model was further evaluated using genome-wide SNVs, including common SNPs in dbSNP135 (*17*), point mutations in HGMD (*18*), and rare variants from ANNOVAR (*55*). Finally, the splicing model was applied in three disease studies: SMA, hereditary nonpolyposis colorectal cancer, and ASD. A large amount of literature-curated data from splicing assays was used to validate our predictions for SMA and nonpolyposis colorectal cancer mutations, with additional mutagenesis experiments carried out for SMA. When applying our computational model to ASD, we performed whole-genome sequencing on five ASD and four control subjects (deposited at the European Genome-Phenome Archive, www.ebi.ac.uk/ega, with accession number EGAS00001000928). Our SPANR Web tool (http://tools.genes.toronto.edu) is programmed in Python under the Flask Web framework (http://flask.pocoo.org) and makes use of MongoDB (www.mongodb.org) and the Celery distributed task queue (http://celery.readthedocs.org).

REFERENCES AND NOTES

1. K. Lindblad-Toh *et al.*, A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011). doi: 10.1038/nature10530; pmid: 21993624
2. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). doi: 10.1038/nature11247; pmid: 22955616
3. Y. Barash *et al.*, Deciphering the splicing code. *Nature* **465**, 53–59 (2010). doi: 10.1038/nature09000; pmid: 20445623
4. C. Zhang *et al.*, Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science* **329**, 439–443 (2010). doi: 10.1126/science.1191150; pmid: 20558669
5. N. L. Barbosa-Morais *et al.*, The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012). doi: 10.1126/science.1230612; pmid: 23258890
6. E. Segal, J. Widom, From DNA sequence to transcriptional behaviour: A quantitative approach. *Nat. Rev. Genet.* **10**, 443–456 (2009). doi: 10.1038/nrg2591; pmid: 19506578
7. F. Gnad, A. Baucom, K. Mukhyala, G. Manning, Z. Zhang, Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* **14** (suppl. 3), S7 (2013). pmid: 23819521
8. M. Kircher *et al.*, A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014). doi: 10.1038/ng.2892; pmid: 24487276
9. M. Mort *et al.*, MutPred Splice: Machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol.* **15**, R19 (2014). doi: 10.1186/gb-2014-15-1-r19; pmid: 24451234
10. T. Sterne-Weiler, J. R. Sanford, Exon identity crisis: Disease-causing mutations that disrupt the splicing code. *Genome Biol.* **15**, 201 (2014). doi: 10.1186/gb4150; pmid: 24456648
11. H. Y. Xiong, Y. Barash, B. J. Frey, Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics* **27**, 2554–2562 (2011). pmid: 21803804
12. Y. Barash *et al.*, AVISPA: A web tool for the prediction and analysis of alternative splicing. *Genome Biol.* **14**, R114 (2013). doi: 10.1186/gb-2013-14-10-r114; pmid: 24156756
13. See supplementary materials on *Science* Online.
14. D. Ray *et al.*, A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013). doi: 10.1038/nature12311; pmid: 23846655
15. H. Han *et al.*, MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**, 241–245 (2013). doi: 10.1038/nature12270; pmid: 23739326
16. T. Lappalainen *et al.*, Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013). doi: 10.1038/nature12531; pmid: 24037378
17. S. T. Sherry *et al.*, dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001). doi: 10.1093/nar/29.1.308; pmid: 11125122
18. P. D. Stenson *et al.*, The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009). doi: 10.1186/gm13; pmid: 19348700
19. F. Supek, B. Miñana, J. Valcárcel, T. Gabaldón, B. Lehner, Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**, 1324–1335 (2014). doi: 10.1016/j.cell.2014.01.051; pmid: 24630730
20. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
21. A. González-Pérez, N. López-Bigas, Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.* **88**, 440–449 (2011). doi: 10.1016/j.ajhg.2011.03.004; pmid: 21457909
22. L. A. Hindorff *et al.*, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9362–9367 (2009). doi: 10.1073/pnas.0903103106; pmid: 19474294
23. R. N. Singh, Evolving concepts on human SMN pre-mRNA splicing. *RNA Biol.* **4**, 7–10 (2007). doi: 10.4161/rna.4.1.4535; pmid: 17592254
24. T. Kashima, N. Rao, J. L. Manley, An intronic element contributes to splicing repression in spinal muscular atrophy. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 3426–3431 (2007). doi: 10.1073/pnas.0700343104; pmid: 17307868
25. Y. Hua, T. A. Vickers, H. L. Okunola, C. F. Bennett, A. R. Krainer, Antisense masking of an hnRNP A1/A2 intronic splicing silencer corrects SMN2 splicing in transgenic mice. *Am. J. Hum. Genet.* **82**, 834–848 (2008). doi: 10.1016/j.ajhg.2008.01.014; pmid: 18371932
26. R. A. Barnetson *et al.*, Classification of ambiguous mutations in DNA mismatch repair genes identified in a population-based study of colorectal cancer. *Hum. Mutat.* **29**, 367–374 (2008). doi: 10.1002/humu.20635; pmid: 18033691
27. P. Peltomäki, H. Vasen, Mutations associated with HNPCC predisposition — Update of ICG-HNPCC/INSiGHT mutation database. *Dis. Markers* **20**, 269–276 (2004). doi: 10.1155/2004/305058; pmid: 15528792
28. S. Arnold *et al.*, Classifying MLH1 and MSH2 variants using bioinformatic prediction, splicing assays, segregation, and tumor characteristics. *Hum. Mutat.* **30**, 757–770 (2009). doi: 10.1002/humu.20936; pmid: 19267393
29. B. Betz *et al.*, Comparative in silico analyses and experimental validation of novel splice site and missense mutations in the genes MLH1 and MSH2. *J. Cancer Res. Clin. Oncol.* **136**, 123–134 (2010). doi: 10.1007/s00432-009-0643-z; pmid: 19669161
30. M. Nyström-Lahti *et al.*, Missense and nonsense mutations in codon 659 of MLH1 cause aberrant splicing of messenger RNA in HNPCC kindreds. *Genes Chromosomes Cancer* **26**, 372–375 (1999). doi: 10.1002/(SICI)1098-2264(199912)26:4<372::AID-GCC12>3.0.CO;2-V; pmid: 10534773
31. P. Lastella, N. C. Surdo, N. Resta, G. Guanti, A. Stella, In silico and in vivo splicing analysis of MLH1 and MSH2 missense mutations shows exon- and tissue-specific effects. *BMC Genomics* **7**, 243 (2006). doi: 10.1186/1471-2164-7-243; pmid: 16995940
32. J. Kosinski, I. Hinrichsen, J. M. Bujnicki, P. Friedhoff, G. Plotz, Identification of Lynch syndrome mutations in the MLH1-PMS2 interface that disturb dimerization and mismatch repair. *Hum. Mutat.* **31**, 975–982 (2010). doi: 10.1002/humu.21301; pmid: 20533529
33. P. J. Smith *et al.*, An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum. Mol. Genet.* **15**, 2490–2508 (2006). doi: 10.1093/hmg/ddl171; pmid: 16825284
34. J. D. Buxbaum *et al.*, The Autism Sequencing Consortium: Large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron* **76**, 1052–1056 (2012). doi: 10.1016/j.neuron.2012.12.008; pmid: 23259942
35. C. Betancur, Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380**, 42–77 (2011). doi: 10.1016/j.brainres.2010.11.078; pmid: 21129364
36. B. Devlin, S. W. Scherer, Genetic architecture in autism spectrum disorder. *Curr. Opin. Genet. Dev.* **22**, 229–237 (2012). doi: 10.1016/j.gde.2012.03.002; pmid: 22463983
37. I. Iossifov *et al.*, De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012). doi: 10.1016/j.neuron.2012.04.009; pmid: 22542183
38. Y. H. Jiang *et al.*, Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am. J. Hum. Genet.* **93**, 249–263 (2013). doi: 10.1016/j.ajhg.2013.06.012; pmid: 23849776
39. R. Anney *et al.*, A genome-wide scan for common alleles affecting risk for autism. *Hum. Mol. Genet.* **19**, 4072–4082 (2010). doi: 10.1093/hmg/ddq307; pmid: 20663923
40. T. C. Südhof, Neuroligins and neurexins link synaptic function to cognitive disease. *Nature* **455**, 903–911 (2008). doi: 10.1038/nature07456; pmid: 18923512
41. I. Voineagu *et al.*, Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011). doi: 10.1038/nature10110; pmid: 21614001
42. R. F. Wintle *et al.*, A genotype resource for postmortem brain samples from the Autism Tissue Program. *Autism Res.* **4**, 89–97 (2011). doi: 10.1002/aur.173; pmid: 21254448
43. D. Pinto *et al.*, Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010). doi: 10.1038/nature09146; pmid: 20531469
44. M. Uddin *et al.*, Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder. *Nat. Genet.* **46**, 742–747 (2014). doi: 10.1038/ng.2980; pmid: 24859339
45. E. Skafidas *et al.*, Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. *Mol. Psychiatry* **19**, 504–510 (2014). doi: 10.1038/mp.2012.126; pmid: 22965006
46. E. B. Robinson *et al.*, Response to 'Predicting the diagnosis of autism spectrum disorder using gene pathway analysis'. *Mol. Psychiatry* **19**, 860–861 (2014). doi: 10.1038/mp.2013.125; pmid: 24145379
47. E. Khurana *et al.*, Integrative annotation of variants from 1092 humans: Application to cancer genomics. *Science* **342**, 1235587 (2013). doi: 10.1126/science.1235587; pmid: 24092746
48. K. H. Lim, L. Ferraris, M. E. Filloux, B. J. Raphael, W. G. Fairbrother, Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11093–11098 (2011). doi: 10.1073/pnas.1101135108; pmid: 21685335
49. A. Woolfe, J. C. Mullikin, L. Elnitski, Genomic features defining exonic variants that modulate splicing. *Genome Biol.* **11**, R20 (2010). doi: 10.1186/gb-2010-11-2-r20; pmid: 20158892
50. B. E. Stranger *et al.*, Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007). doi: 10.1126/science.1136678; pmid: 17289997
51. J. A. Tennessen *et al.*, Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012). doi: 10.1126/science.1219240; pmid: 22604720
52. A. Battle *et al.*, Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014). doi: 10.1101/gr.155192.113; pmid: 24092820
53. U. Braunschweig, S. Gueroussov, A. M. Plocik, B. R. Graveley, B. J. Blencowe, Dynamic integration of splicing within gene regulatory pathways. *Cell* **152**, 1252–1269 (2013). doi: 10.1016/j.cell.2013.02.034; pmid: 23498935
54. D. Brawand *et al.*, The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011). doi: 10.1038/nature10532; pmid: 22012392
55. K. Wang, M. Li, H. Hakonarson, ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010). doi: 10.1093/nar/gkq603; pmid: 20601685

ACKNOWLEDGMENTS

B.J.F. designed the study and wrote the manuscript, with input from B.J.B., S.W.S., B.A., N.J., Y.B., Q.M., and other co-authors. H.Y.X. trained the models. H.Y.X., B.J.F., N.J., and L.J.L. developed the bootstrap method for quantifying **Ψ**. L.J.L. mined exons, mapped reads and designed additional features, extending the original feature set developed by Y.B., B.J.F., and B.J.B. H.Y.X., B.J.F., H.S.N., Q.M., and T.R.H. analyzed RNAcompete data. B.J.F., H.Y.X., L.J.L., and B.A. performed the genome-wide SNV analysis. L.J.L., Y.H., B.J.F., and A.R.K. tested predictions for *SMN1/2*. B.A. and B.J.F. tested predictions for *MLH1/MSH2*. B.A., D.M., R.K.C.Y., B.J.F., and S.W.S. analyzed ASD genomes. S.G. conducted wild-type RT-PCR assays. H.B. developed the Web tool and feature visualization, with input from B.J.F., H.Y.X., B.A., and L.J.L. We thank the Center of Applied Genomics at the Toronto Hospital for Sick Children for providing HGMD annotations, G. Schroth at Illumina for providing the BodyMap RNA-seq data, and M. Brudno, O. Buske, A. Delong, C. Smith, T. Sterne-Weiler, J. Wilson, and J. Valcarel for comments on the manuscript and the Web tool. Supported by Canadian Institutes for Advanced Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), John C. Polanyi, the University of Toronto McLaughlin Centre, and the Ontario Genomics Institute (OGI) (B.J.F.); CIHR and McLaughlin (B.J.B.); McLaughlin, Genome Canada, OGI, and Autism Speaks (S.W.S.); NIH grant R37-GM42699a (A.R.K.); CIHR (Q.M.); an Autism Research Training Fellowship (B.A.); a CIHR Banting Fellowship (H.S.N.); and an NSERC Alexander Graham Bell Scholarship (S.G.). S.W.S. holds the GlaxoSmithKline-CIHR Chair in Genome Sciences. B.J.B. holds the Banbury Chair of Medical Research at the University of Toronto. B.J.F., S.W.S., and T.R.H. are Fellows of the Canadian Institute for Advanced Research. B.J.F. holds the Canada Research Chair in Biological Computation.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/347/6218/1254806/suppl/DC1
Materials and Methods
Figs. S1 to S32
Tables S1 to S18
References (*56*–*81*)

13 April 2014; accepted 19 November 2014
Published online 18 December 2014;
10.1126/science.1254806

The human splicing code reveals new insights into the genetic determinants of disease

Hui Y. Xiong, Babak Alipanahi, Leo J. Lee, Hannes Bretschneider, Daniele Merico, Ryan K. C. Yuen, Yimin Hua, Serge Gueroussov, Hamed S. Najafabadi, Timothy R. Hughes, Quaid Morris, Yoseph Barash, Adrian R. Krainer, Nebojsa Jojic, Stephen W. Scherer, Benjamin J. Blencowe and Brendan J. Frey

Science **347** (6218), 1254806.
DOI: 10.1126/science.1254806originally published online December 18, 2014

Predicting defects in RNA splicing

Most eukaryotic messenger RNAs (mRNAs) are spliced to remove introns. Splicing generates uninterrupted open reading frames that can be translated into proteins. Splicing is often highly regulated, generating alternative spliced forms that code for variant proteins in different tissues. RNA-binding proteins that bind specific sequences in the mRNA regulate splicing. Xiong *et al.* develop a computational model that predicts splicing regulation for any mRNA sequence (see the Perspective by Guigó and Valcárcel). They use this to analyze more than half a million mRNA splicing sequence variants in the human genome. They are able to identify thousands of known disease-causing mutations, as well as many new disease candidates, including 17 new autism-linked genes.

Science, this issue 10.1126/science.1254806; see also p. 124