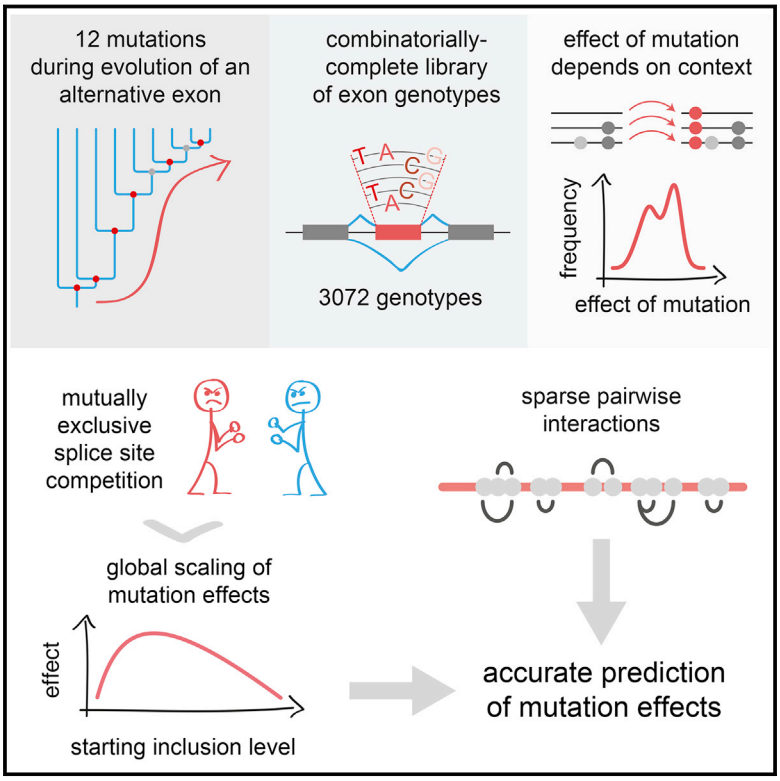
Article

Cell

Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing

Graphical Abstract



Highlights

- Combinatorially complete genotype-to-phenotype map for the evolution of FAS exon 6
- Splicing mutation effects change non-monotonically with exon inclusion
- Splice-site competition introduces a non-linearity into the genotype-phenotype map
- Scaling of splicing perturbation effects is found in exons throughout the genome

Authors

Pablo Baeza-Centurion, Belén Miñana, Jörn M. Schmiedel, Juan Valcárcel, Ben Lehner

Correspondence

juan.valcarcel@crg.eu (J.V.), ben.lehner@crg.eu (B.L.)

In Brief

A quantitative, predictive model for alternative splicing decisions explains the probability of exon inclusion in the context of natural and disease-associated genetic variants

Baeza-Centurion et al., 2019, Cell *176*, 549–563
January 24, 2019
https://doi.org/10.1016/j.cell.2018.12.010

CellPress

Article

Cell

Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing

Pablo Baeza-Centurion,[1,5] Belén Miñana,[2,5] Jörn M. Schmiedel,[1] Juan Valcárcel,[2,3,4,6,*] and Ben Lehner[1,2,4,6,7,*]
[1]Systems Biology Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr Aiguader 88, 08003 Barcelona, Spain
[2]Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr Aiguader 88, 08003 Barcelona, Spain
[3]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain
[4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain
[5]These authors contributed equally
[6]Senior author
[7]Lead Contact
*Correspondence: juan.valcarcel@crg.eu (J.V.), ben.lehner@crg.eu (B.L.)
https://doi.org/10.1016/j.cell.2018.12.010

SUMMARY

Despite a wealth of molecular knowledge, quantitative laws for accurate prediction of biological phenomena remain rare. Alternative pre-mRNA splicing is an important regulated step in gene expression frequently perturbed in human disease. To understand the combined effects of mutations during evolution, we quantified the effects of all possible combinations of exonic mutations accumulated during the emergence of an alternatively spliced human exon. This revealed that mutation effects scale non-monotonically with the inclusion level of an exon, with each mutation having maximum effect at a predictable intermediate inclusion level. This scaling is observed genome-wide for *cis* and *trans* perturbations of splicing, including for natural and disease-associated variants. Mathematical modeling suggests that competition between alternative splice sites is sufficient to cause this non-linearity in the genotype-phenotype map. Combining the global scaling law with specific pairwise interactions between neighboring mutations allows accurate prediction of the effects of complex genotype changes involving >10 mutations.

INTRODUCTION

Accurate quantitative predictions about the behavior of biological systems are still rare. For example, predicting changes in phenotype from changes in genotype is a central challenge in genetics, evolution, agriculture, and personalized medicine (Lehner, 2013). However, predicting the effects of even single mutations in very well-studied genes remains remarkably difficult (Shendure and Akey, 2015).

One reason for the difficulty of genetic prediction is that the consequence of a mutation often changes depending on the genetic background where it is made. This is true considering the complete variation within a genome, but also when only considering additional variation within an individual gene (Lehner, 2011; Phillips, 2008). Comprehensive mutagenesis of individual proteins (Diss and Lehner, 2018; Fowler et al., 2010; Olson et al., 2014; Sarkisyan et al., 2016) and RNAs (Domingo et al., 2018; Li et al., 2016; Puchta et al., 2016) has revealed abundant pairwise interactions between mutations within genes. Changes in phenotype also occur when more than two mutations are combined that cannot be predicted from the phenotypes of the constituent pairwise combinations (Domingo et al., 2018; Sailer and Harms, 2017; Weinreich et al., 2013). These interactions between mutations are known as genetic interactions or epistasis, with pairwise (2^{nd} order) and higher-order (3^{rd}, 4^{th} etc. order) interactions all important for accurate genetic prediction in the few cases where this has been systematically evaluated (Domingo et al., 2018; Phillips, 2008; Poelwijk et al., 2016; Sailer and Harms, 2017; Weinreich et al., 2013).

Deep mutagenesis combined with selection for function and deep sequencing has also been used to quantify the effects of mutations on gene expression. This has included mutagenesis of gene promoters (Kinney et al., 2010; Patwardhan et al., 2009), transcriptional enhancers (Melnikov et al., 2012; Patwardhan et al., 2012), 5′ and 3′ UTRs (Dvir et al., 2013; Holmqvist et al., 2013; Shalem et al., 2015), and intronic and exonic regions that regulate splicing (Braun et al., 2018; Julien et al., 2016; Ke et al., 2018; Rosenberg et al., 2015).

Alternative splicing is a key regulated step in gene expression frequently perturbed in human disease (Daguenet et al., 2015) with 10% of disease-causing exonic mutations altering splicing (Soemedi et al., 2017), and it has been estimated that up to 1/3 of all disease-associated alleles alter splicing (Havens et al., 2013).

Quantifying the effects of all possible single-nucleotide (nt) changes within a model alternatively spliced exon, exon 6 of the *FAS* gene, we previously reported that over 60% of single mutations alter inclusion of the exon. Moreover, testing double mutants revealed frequent non-linear interactions between pairs of mutations, making it difficult to predict the exact level of splicing when mutations are combined (Julien et al., 2016).

FAS exon 6 is alternatively spliced in humans, with inclusion varying across cell types and conditions, and encodes the transmembrane domain of the FAS/CD95 death receptor. mRNAs that skip exon 6 encode a secreted protein lacking the transmembrane domain that acts as a decoy receptor. The alternative splicing of *FAS* therefore switches the protein from a pro- to an anti-apoptotic molecule (Cascino et al., 1995).

Here, we use *FAS* exon 6 as a model system to investigate how higher-order combinations of mutations interact to cause phenotypic change and the extent to which it is possible to make accurate genetic predictions about changes in genotype involving multiple mutations. We show that *FAS* exon 6 became alternatively spliced during the evolution of primates. Combining the 12 substitutions that separate the sequence of the human exon from the primate ancestor in all 3,072 possible combinations and quantifying the effects on splicing, we reveal a non-monotonic mathematical law for how mutations combine to alter splicing. This non-intuitive scaling may simply be a consequence of mutually exclusive splice-site competition. Scaling is observed in other deep mutagenesis datasets, for natural genetic variants and for *trans* perturbations to the alternative splicing of endogenous mRNAs. Finally, we show that, if this general nonlinearity in the genotype-to-phenotype map is taken into account, a small number of specific proximal pairwise interactions are sufficient to accurately predict the effect of >10 mutations when combined.

RESULTS

Reconstructing the Evolution of *FAS* Exon 6

In humans, *FAS* exon 6 is alternatively spliced such that skipping of the exon switches the FAS protein from a membrane-bound pro-apoptotic isoform to a soluble anti-apoptotic isoform of the protein. Alternative splicing of this exon varies among tissues, with percentage spliced-in (PSI) values ranging across five human tissues from ~70% in lung to ~95% in kidney (Figure 1A; Table S1).

To investigate the evolution of *FAS* exon 6 alternative splicing, we analyzed RNA sequencing data from 5 tissues in different primates. Inclusion of the exon also changes across tissues in chimpanzees and Old World monkeys (Figure 1A; Table S1), but in New World monkeys and lemurs the exon is nearly constitutive in all tissues (Figure 1A; Table S1), as is also true in mice (Figure S1A; Table S2). This suggests that variable skipping of this exon evolved within the primate lineage.

The intronic sequences 5′ and 3′ of *FAS* exon 6 are largely invariant across primates (Figure S1D). However, there have been 12 nt substitutions at 11 positions since the last common ancestor of primates (Figures 1A, 1B, and S1B). Parsimony assigns these substitutions to 6 nodes of the species tree (Figures 1A and S1B). We constructed these inferred evolutionary intermediates and quantified their PSI in a minigene construct containing *FAS* exons 5–7 and the intervening introns. While the human exon was included at 60%, exons with the sequence inferred for the ancestors of primates (at the root of the phylogenetic tree in Figure 1A), haplorrhines (node 2) and simians (node 3) were nearly constitutive, with PSIs of 96%, 97%, and 96%, respectively (Figures 1A and S1C). Exons with the sequences of more recent intermediates had intermediate levels of inclusion: 57% and 79% for the common ancestors of catarrhines and great apes. The increase in inclusion as more ancestral substitutions are added is also consistent with the increase in average inclusion levels in the RNA sequencing data from humans to chimpanzees to New World monkeys (Figure 1A; Table S1), suggesting that the decrease in exon 6 inclusion was mainly driven by nt changes in the exon.

Combinatorially Complete Mapping of a Genotype Space

The identification of substantial changes in exon inclusion associated with only 12 mutations presents an opportunity to study the extent to which mutations have independent effects. The evolution of this exon occurred through one of millions of possible evolutionary paths that connect the genotype of the primate ancestor with the current human *FAS* exon 6 genotype. Would the effect of each mutation have been the same had it occurred at a different step in evolution?

To quantify the extent to which each of these mutations has effects on splicing that are either constant or context dependent, we designed a library of exon 6 variants in which all 12 mutations could randomly occur as single, double, triple, and higher-order combinations, a total of 3,072 genotypes (= $2^{10} \times 3$; 10 positions can have 2 different nt and one can have 3 different nt, Figure 1B).

We cloned the library into a minigene cassette covering *FAS* exons 5–7, transfected it into HEK293 cells, and quantified how often each particular genotype was included in the final mature mRNA relative to every other genotype in the library by RT-PCR and deep sequencing (Figure 1C). The linear relationship between enrichment scores (ESs) and PSI (r^2 = 0.92, Figure S1F) allows a PSI value for each genotype to be estimated (Table S3). ESs were generally well correlated across 9 biological replicates (Pearson's *r* between 0.57 and 0.74, Figure S1G). Correlations were much stronger for genotypes with a standard deviation of <10 PSI units (*r* between 0.97 and 0.98, Figure S1G). We focus in the main text on this high confidence subset of the data (n = 794 genotypes; analyses of all 3,072 genotypes are shown in supplemental figures with similar conclusions).

The PSI values of all genotypes range from 0% to 100% and follow a bimodal distribution, with 48% of genotypes having a PSI above 80% with a mode close to the PSI of the ancestral exon (96%), and 20% of genotypes having a PSI below 20% with a mode near 0% (Figures 1D, 1E, S1H, and S1I).

Mutations Have Non-independent Effects on Splicing

We first tested whether mutations have the same effect irrespective of the starting genotype in which they occur (Figure 2A). For example, the mutation T19G occurs in the ancestral genotype as well as in 11 genotypes that differ by 1 nt from the ancestor, 54 genotypes that differ by 2 nt, and so on (Figure 2B).

All 12 mutations had effects that changed substantially in different starting genotypes (Figures 2C and S2A). While

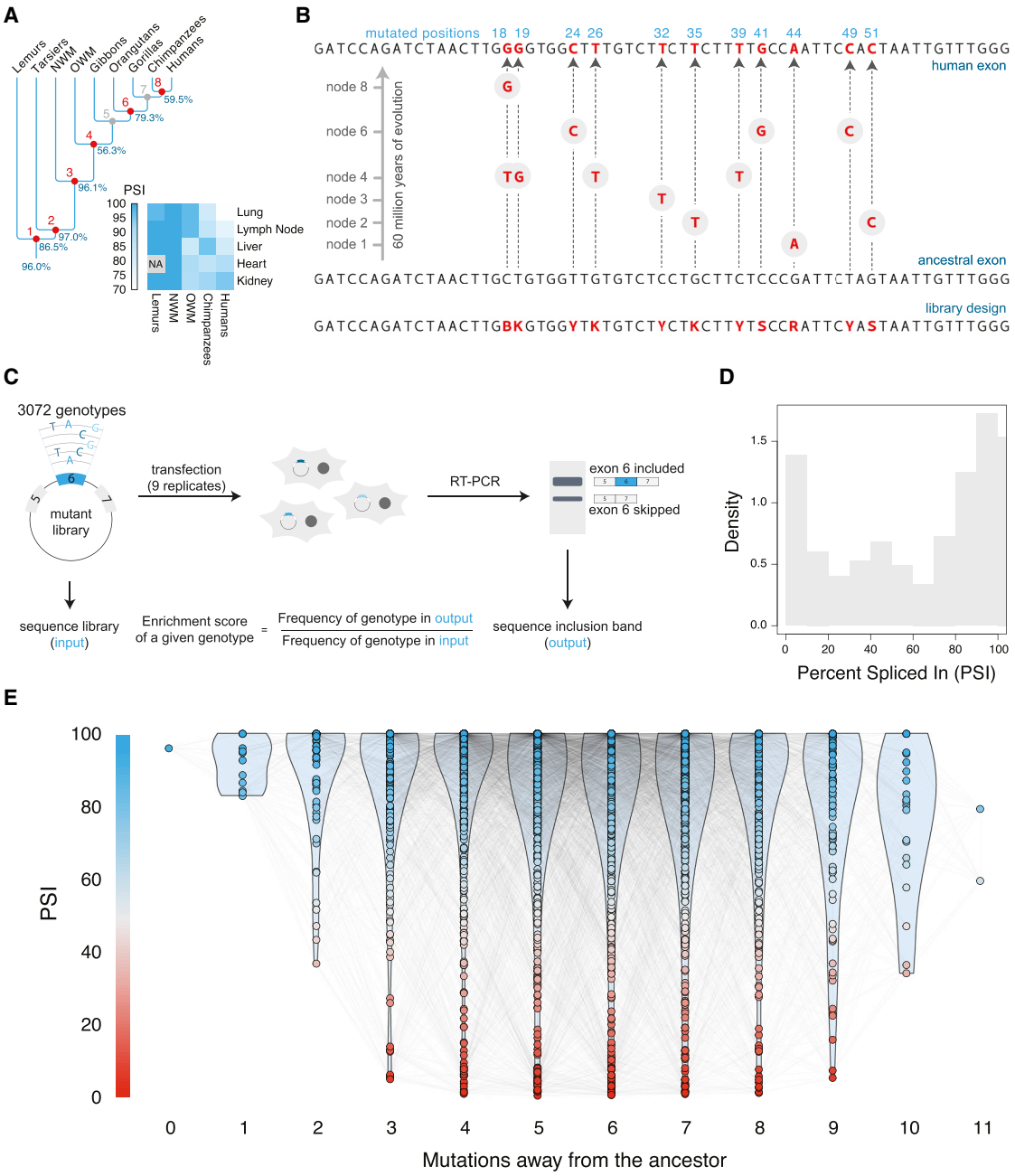


Figure 1. Analyzing Mutation Effects in a Combinatorially Complete Subset of a Genotype Space

(A) Primate species tree. Inset heatmap shows the percent spliced-in (PSI) of *FAS* exon 6 and its orthologs in RNA-seq data from different tissues in different species. % indicates the PSI of minigene transcripts with exon genotypes corresponding to each node.

(B) Mutations since the last common ancestor of primates. Nodes indicate inferred evolutionary intermediates. The sequence below the ancestral exon shows the design of our library.

(C) Experimental protocol.

(D) Distribution of PSI values.

(E) Genotype network of the library. Genotypes (nodes) are connected when they differ by 1 nt. The distribution of PSI values for each Hamming distance away from the ancestral sequence is shown as a vertical violin plot. PSI values estimated to be >100% are plotted at 100%.

mutations tended to display quantitatively different effects in the same direction (e.g., toward more inclusion), some mutations showed qualitatively different effects in different contexts: T19G promotes skipping in 134 exon genotypes and inclusion in 253 (one-sample Wilcoxon rank-sum test, false discovery rate [FDR] <0.05, n = 1,536 tests).

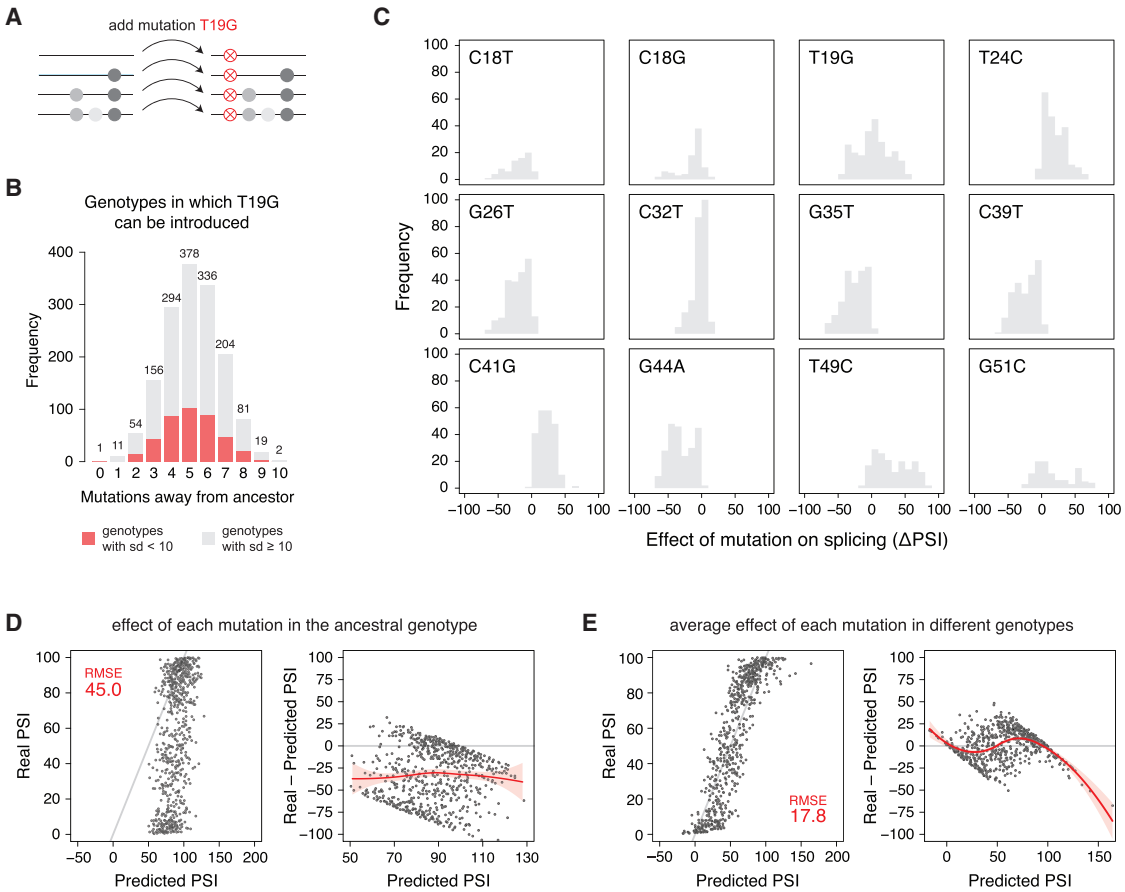


Figure 2. Mutations Have Non-independent Effects on Alternative Splicing

(A) T19G (red) can be introduced in genotypes containing different additional mutations (gray).

(B) Number of genotypes where T19G occurs for each Hamming distance away from the ancestral sequence.

(C) Distributions of mutation effects.

(D) Using the effect of mutations on the ancestral sequence to predict their effect in other contexts leads to poor prediction. Left: observed versus predicted PSI values. Right: residual plot with loess trend line and 95% confidence band.

(E) Predictive model that uses the average effect of mutations in different genotypes. Plots as in D.

Consistent with mutations having effects that change in different genotypes, quantifying the effect of each mutation in the ancestral exon and combining these effects using a linear model with 12 parameters (one parameter for each of the 12 mutations in our dataset) gave very poor prediction of the inclusion of the exons in the library, with a root-mean-square error (RMSE) of 45.0 PSI units (Figure 2D). Including the restriction that PSI values cannot be predicted to be above 100 or below 0 only moderately improved the predictions (RMSE = 43.7 PSI units, Figures S2B–S2D).

Considering the average effect of each mutation across all genotypes substantially improved the predictions, but important deviations from the real values remained (10-fold cross-validation RMSE = 17.8 PSI units; RMSE = 16.8 when bounding predictions between 0% and 100%, Figures 2E and S2E–S2G). The effects of mutations on exon 6 inclusion are therefore context dependent, with the effect of a mutation in a single genotype providing limited prediction of its effects in other closely related genotypes.

Mutation Effects Scale Non-monotonically with Starting Inclusion Levels

To investigate why the effects of individual mutations change across the dataset, we studied the relationship between the inclusion level of an exon before (starting PSI) and after (final PSI) a mutation is made (Figure 3A). Exon inclusion is a bounded function ranging from 0% to 100%. Thus, one simple model is that mutations have a constant effect on splicing that saturates when 0% or 100% inclusion is obtained (model 1, Figures 3B and 3C). Other models include: a fractional effects model where the distance to 100% inclusion or skipping is always reduced by a certain factor (model 2, Figures 3B and 3C), a diminishing returns model where the effects of mutations progressively decrease as they approach the limits of exon inclusion or skipping (model 3, Figures 3B and 3C), or a model where mutations push inclusion to the limits irrespective of the starting genotype (model 4, Figures 3B and 3C).

Surprisingly, plotting the effect of each mutation against the PSI of the genotype in which it occurs reveals that the

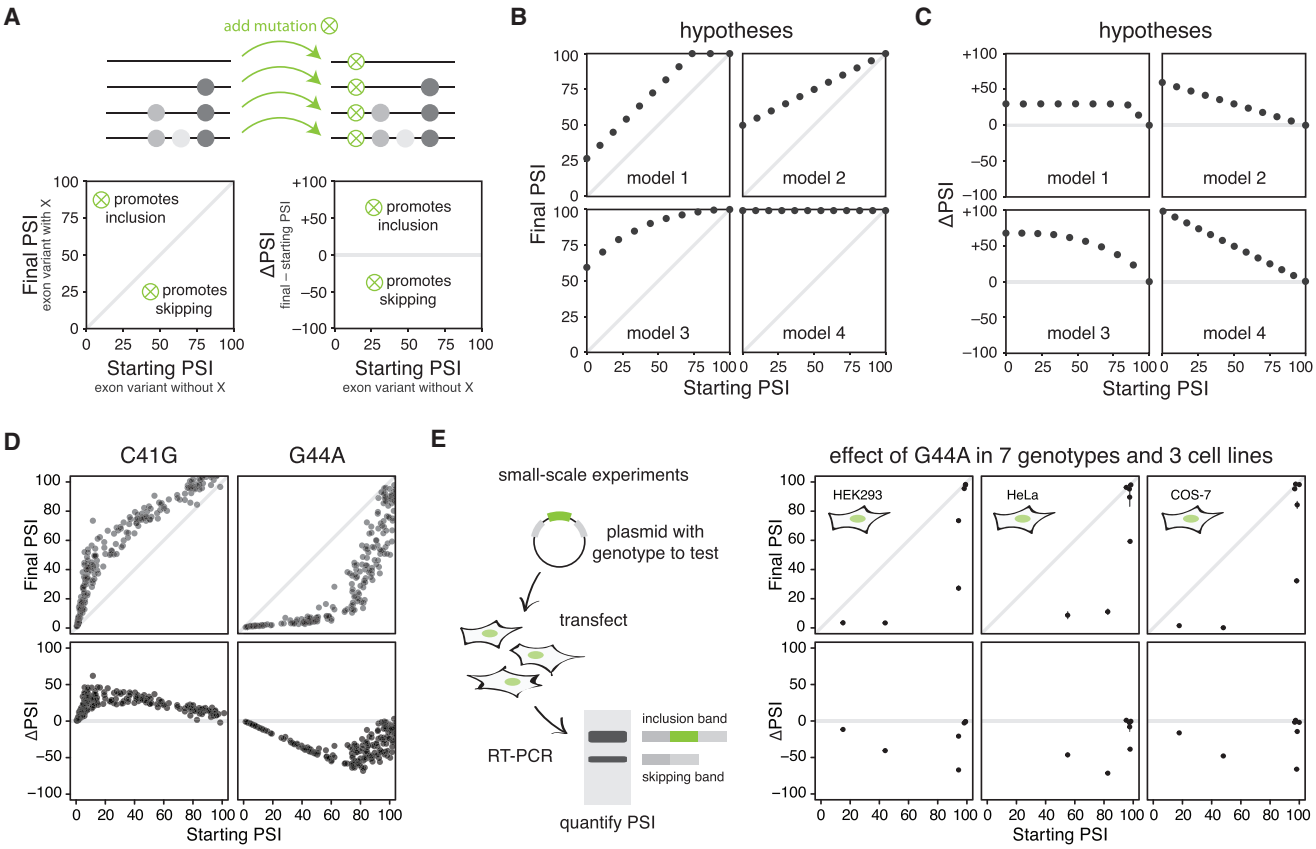


Figure 3. Non-linear Scaling of Mutation Effects as a Function of the Starting PSI

(A) Mutation X can be introduced in different genetic backgrounds, which might change its effect. To visualize how the effect of the mutation depends on the PSI of the exon in which it is introduced, the final PSI of a genotype with X (or ΔPSI) is plotted as a function of the PSI of the same genotype without X (starting PSI).
(B and C) Models for how the PSI at which a mutation is introduced (starting PSI) affects the final PSI (B) or the change in PSI (C) upon introducing the mutation.
(D) Relationship between final and starting PSI for 2 splicing mutations. The effect of inclusion-promoting C41G is smaller at both low and high starting PSIs. The effect of the skipping-promoting G44A is smaller at low and high starting PSIs, with the maximum effect size occurring at intermediate starting PSIs.
(E) Non-linear scaling for G44A in different genotypes in 3 cell lines.

relationship between the effect of the mutation and the starting PSI is non-monotonic, first increasing to a maximum and then decreasing again as the starting PSI changes from 0% to 100% (Figures 3D and S3). We confirmed this relationship in 3 different cell types for a mutation retested in 7 different exons with different starting PSIs (Figures 3E and S4A–S4C; Table S4). Thus, irrespective of the effect size or direction of effect of the mutation, there is a global scaling of mutation effects that involves the gradual reduction of mutation effects when the starting PSI is closer to either complete inclusion or skipping, with maximum effects at specific intermediate starting PSI values.

This means that, as expected, an inclusion-promoting mutation will have a small effect when introduced in an exon with a PSI near 100%. However, that same mutation will also have a small effect when introduced in an exon with a PSI near 0%, even though such an exon could allow for large increases in inclusion.

Moreover, mutations do not have their maximum effect at the same starting PSI level (for example, always when starting from 50% inclusion). Rather, the starting PSI at which each mutation has its maximum effect is a property of that mutation, with mutations of smaller maximum effect having their strongest effect at PSIs closer to 50% and mutations of larger maximum effect having their strongest effects closer to full inclusion or skipping (see below, Figure 4F).

Mathematical Modeling Suggests Competition as the Origin of Non-monotonicity in the Effects of Mutations

To understand why the effects of mutations show this surprising scaling behavior, we built a mathematical model for exon inclusion. Although it ignores the molecular details, this model captures one essential component of splicing decisions, which is competition between splice sites. For example, the 3′ splice sites of introns 5 and 6 compete with each other for pairing with the 5′ splice site of intron 5. These competitions result in mutually exclusive, unique outcomes for each individual mRNA molecule.

In the model, the probability that a given exon is included at a specific time **t** is the probability of first arrival in a Poisson process (i.e., the probability that a splice site is recognized by the

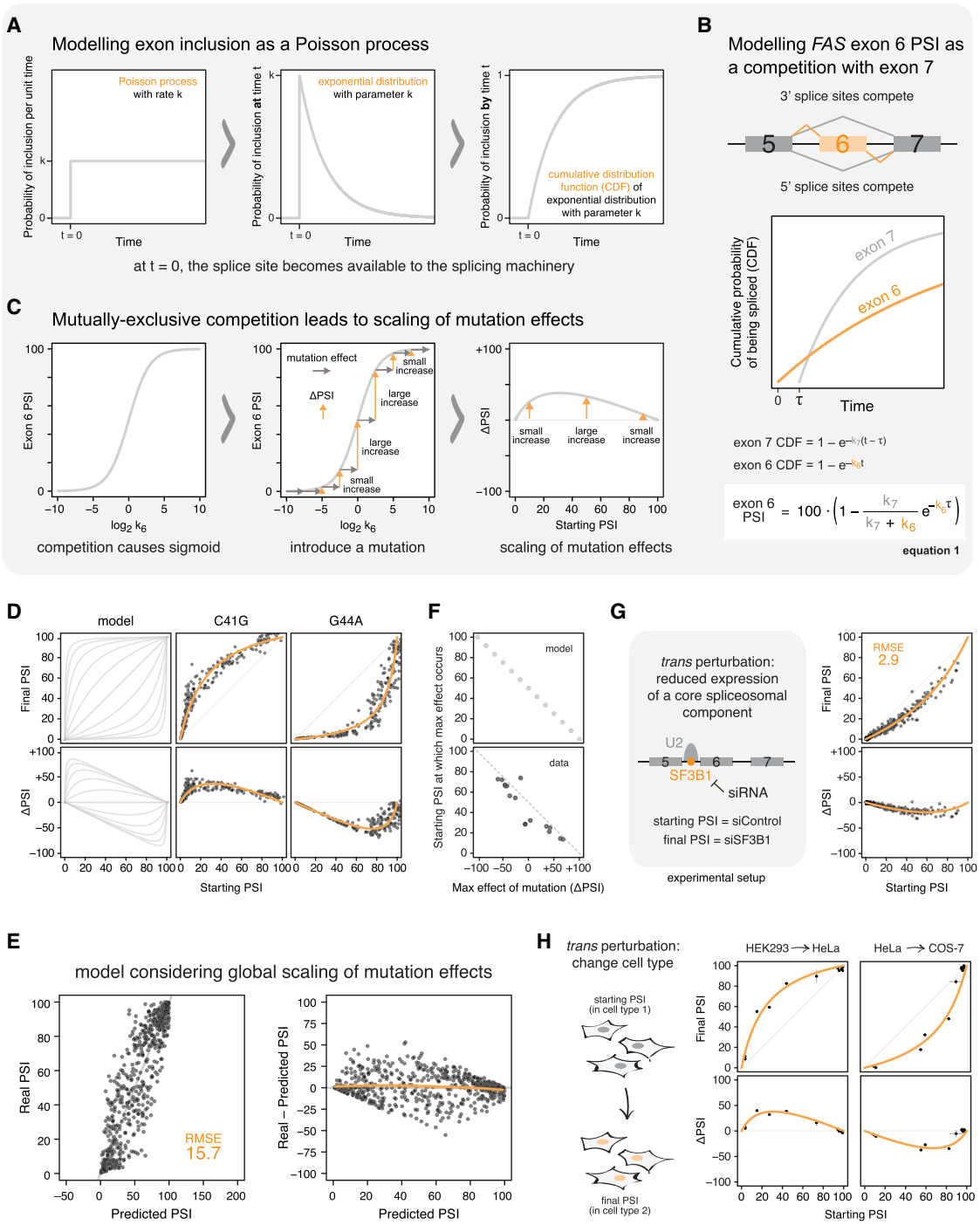


Figure 4. Splice-Site Competition Leads to Non-monotonic Global Scaling of Mutation Effects

(A) Exon inclusion was modeled as a Poisson process, where the probability of inclusion remains constant at k per unit time while the exon is not spliced in.

(B) 3′ and 5′ splice sites compete. The PSI of exon 6 is given by the probability that it is included in the mature mRNA before exon 7. Modeling inclusion of both exons as competing Poisson processes results in Equation 1.

(C) The relationship between exon 6 PSI and fold changes in its splicing efficiency parameter k_6 is sigmoidal.

(D) Dependence of the final PSI and ΔPSI on the starting PSI, for different values of (A) (mutation effect), and fitting the model to the data (also Figure S3).

(E) Predicting PSI with a model that considers the scaling of mutation effects. Left: observed versus predicted PSI. Right: residual plot with loess trend line and 95% confidence band.

(legend continued on next page)

splicing machinery remains constant over time, until it is eventually recognized) with parameter **k** (Figure 4A). Without splice-site competition, the probability of exon 6 inclusion at a specific time point therefore follows an exponential distribution with parameter $\mathbf{k_6}$ (orange curve in Figure 4B, see STAR Methods), and the probability of exon 7 splicing to exon 5 follows an exponential distribution with parameter $\mathbf{k_7}$ (gray curve in Figure 4B). If exons 6 and 7 compete for splicing (Figure 4B), the PSI of exon 6 is given by:

$$exon\ 6\ PSI = 100 \cdot \left(1 - \frac{k_7}{k_7 + k_6} e^{-k_6 \tau}\right) \quad \text{(Equation 1)}$$

where τ is the time delay between the splice sites flanking exon 6 becoming available to the splicing machinery and the 3′ splice site preceding exon 7 (3′ splice site of intron 6) becoming available, when competition between alternative splice sites for pairing to a common splice site takes place (full derivation in Data S1).

Importantly, Equation 1 implies a sigmoidal relationship between exon 6 PSI and fold changes in $\mathbf{k_6}$ (Figures 4C and S5A), so changes in $\mathbf{k_6}$ do not always result in the same PSI change (ΔPSI). Instead, a mutation that alters $\mathbf{k_6}$ results in small ΔPSI if that exon has very low levels of inclusion, larger ΔPSI if the exon has intermediate levels of inclusion, and small ΔPSI when the exon displays high levels of inclusion (middle panel in Figure 4C), thus providing a rationale for the observed non-monotonicity in the effects of mutations.

If we fix $\mathbf{k_7}$ to 1 and τ to 0, the relationship between the ΔPSI caused by a mutation and the PSI at which this effect is observed is given by:

$$\Delta PSI = 100 \cdot \left(\frac{A \cdot Starting\ PSI}{100 - Starting\ PSI + A \cdot Starting\ PSI}\right) - Starting\ PSI \quad \text{(Equation 2)}$$

where the mutation introduces an **A**-fold change in $\mathbf{k_6}$ (i.e., **A** is a parameter describing the molecular effect of a mutation; see Data S1).

Equation 2 describes a relationship between the starting PSI and the change in PSI (right panel in Figure 4C) very similar to that observed for the empirical data (Figures 4D and S3). Introducing a time delay between exon 6 and exon 7 synthesis changes the shape of these curves, but the non-monotonic behavior remains (Figure S5C).

In summary, the seemingly non-intuitive scaling of mutation effects may simply arise because the mutual exclusivity of splice-site choice results in a sigmoidal relationship between the change in the efficiency of exon 6 recognition and the final PSI.

Global Scaling Contributes Substantially to Genetic Prediction

We quantified the extent to which the global scaling law improves genetic prediction across all genotypes in our dataset. This model has the same number of parameters as the simple model –one for each mutation representing its parameter **A**. The cross-validation RMSE decreases from 17.8 to 15.7 PSI units (Figures 2E and 4E), and the systematic biases in the predictions observed with the simple linear model are reduced (Figures 2E, 4E, and S5D).

The Starting PSI at which the Maximum Effect of a Mutation Occurs Is Inversely Related to the Effect Size of the Mutation

If the phenotypic effect of a mutation (ΔPSI) at a given starting PSI is known, Equation 2 can be used to determine the molecular effect **A** of that mutation. Its ΔPSI at any starting PSI can then be calculated. The relationship between the maximum effect size of a mutation and the starting PSI at which this effect size is observed is expected to be (see Data S1):

$$Starting\ PSI\ where\ max\ effect\ occurs = 50 - \frac{1}{2} \cdot Max\ effect \quad \text{(Equation 3)}$$

in good agreement with the behavior of mutations in our dataset (r^2 = 0.93, Figure 4F). Thus, using a single parameter—the **A** parameter in Equation 2—the global scaling law determines both the starting PSI at which the mutation will have its maximum effect, as well as the ΔPSI for all other starting PSIs.

Trans Perturbations Also Cause Non-monotonic Changes in Inclusion

The global scaling law predicts how the effect of a mutation on splicing depends on the initial inclusion level, but the law should also be valid for any other perturbation altering the efficiency of exon 6 inclusion ($\mathbf{k_6}$). We reduced the expression of the splicing factor SF3B1, resulting in reduced exon 6 inclusion (Figure S4D; Tejedor et al., 2015). We used a library of human *FAS* exon 6 variants containing all 189 single mutations (Julien et al., 2016) to confirm that the consequence of SF3B1 depletion also scales non-monotonically with the starting PSI (Figure 4G). Thus, the scaling law applies not only to the effects of mutations within the exon, but also to the consequences of reducing the activity of a *trans*-acting factor.

Differences between Cell Types Also Display Global Scaling

To test whether scaling is observed under other conditions that induce differences in splicing patterns, we compared the PSI of 14 exon 6 genotypes (Table S4) in 3 different cell lines.

(F) Relationship between starting PSI at which the maximum effect of a mutation occurs and the maximum ΔPSI effect. Top: model behavior. Bottom: dataset.
(G) Scaling predicts the effect of reduced SF3B1 on the inclusion of *FAS* exon 6 variants.
(H) 14 genotypes were transfected in 3 cell lines and their PSIs determined using RT-PCR assays. Scaling predicts the effect of changing the cell type on the inclusion of these genotypes.

Comparing exon PSI between pairs of cell lines (where the PSI in one cell line is the starting PSI and in the other the final PSI) also displayed a scaling effect (Figure 4H). The scaling law can therefore be used to predict how variants of an exon differentially respond to a complex perturbation.

Additional Specific Interactions Are Sparse and Occur between Proximal Mutations

After considering global scaling, the PSI values still show substantial deviance from our predictions (Figures 4E and S4D), suggesting that specific interactions may occur between mutations. We tested whether the behavior of each mutation was different in the presence and absence of every other mutation and identified 7 interactions (Figures 5A and 5B), which can be classified into different qualitative types (Figure 5C; Weinreich et al., 2005). Magnitude epistasis happens when the magnitude (but not the direction) of the effect changes in a given genetic background. Sign epistasis occurs when the direction of a mutation effect changes. Masking epistasis takes place when the effect of a mutation disappears. We found 1 example of sign epistasis (C18G-T19G), 2 of masking epistasis (C32T-G35T and T49C-G51C), and 4 of magnitude epistasis (C18T-T19G, T24G-G26T, C39T-G41C, and C39T-G44A). These epistatic effects were consistent across a large number of genotypes (Figures 5D and S6A).

We validated the C18G-T19G interaction by transfecting minigenes containing these mutations into HEK293 and COS-7 cells (Figures S6B and S6C). As in our library, T19G promotes inclusion in the absence of C18G but skipping in its presence, also when integrated at a single locus in the genome (Figure S6E).

The 7 pairwise interactions were all between mutations within 6 nt of each other (7 out of 12 [58.3%] pairs within a hexamer interact compared to 0 of the other 53 pairs [0%], $p = 1.138 \times 10^{-6}$, Fisher's exact test). This is likely due to effects on binding of a *trans*-acting splicing factor (4–7 nt being the common binding site size for typical RNA-binding domains, Daubner et al., 2013), or 2 *trans*-acting factors with adjacent or overlapping binding sites.

Both Global Scaling and Specific Interactions Are Important for Accurate Genetic Prediction

After building a model considering both global scaling and the 7 pairwise interactions, the 10-fold cross-validation RMSE decreased to 8.0 PSI units (Figures 6A and 6B), less than half the RMSE of the models that did not consider either global scaling or specific interactions (Figures 2E and 4E). Moreover the prediction error when considering both global scaling and specific interactions does not increase with the number of mutations in the genotypes whose PSI is being predicted (Figure 6C). Thus, when both global scaling and specific interactions are considered, we can accurately predict the combined effect of up to 10 mutations.

Scaling of Mutation Effects in the Alternative Splicing of *WT1* Exon 5

To test whether global scaling applies to other exons, we analyzed data from a mutant library of *WT1* exon 5 (Ke et al., 2018). The inclusion levels of 141 single and 414 double mutants were varied by introducing 10 different exon splicing regulatory sequences (ESRs) in exon positions 5–10 (Figure 7A). We compared the ESs of mutants in the presence of the WT ESR (starting ES) and in the presence of other ESRs (final ES). The effect of changing an ESR inside *WT1* exon 5 scales as predicted by the scaling law (Figure 7B).

Evidence for Global Scaling in Exons throughout the Genome

Changes in *FAS* exon 6 PSI due to differences in the levels and activities of splicing factors between cell types undergo scaling (Figure 4H). To test whether other exons follow this behavior, we compared the inclusion of all exons in the genome across 4 pairs of conditions: 2 tissues (brain and skin), 2 cell lines (HepG2 and human umbilical vein endothelial cell [HUVEC]), the presence or absence of mutations in SF3B1, and 2 different developmental states (Figure 7C). The PSI of exons with high or low inclusion levels in one condition (e.g., the first tissue) tend to be similar in the other condition (the second tissue), whereas larger ΔPSIs are observed for exons with intermediate inclusion levels, consistent with global scaling.

Scaling of the Effects of sQTLs

To globally study the effect of changes in *cis* on the inclusion of exons across the genome, we performed a splicing quantitative trait locus (sQTL) analysis to find variants in a gene associated with altered inclusion of one of its exons (see STAR Methods). We compared the effects of 193,812 putative sQTLs (defined with Bonferroni-corrected $p < 0.05$) in different tissues across 635 humans (Battle et al., 2017). Since exon PSI often changes across tissues (Pan et al., 2008), this allowed us to compare the effect on inclusion of the same genetic variants at different PSIs for 4,418 alternative cassette exons. sQTLs had smaller effects in tissues with low or high PSI and larger effects in tissues with intermediate PSI (Figure 7D). For example, the PSI of ASPH exon 3 increases in the presence of SNP rs2350919. While the increase could in principle be more readily detected in tissues where the exon is mostly skipped, the PSI increase is more evident in tissues with an intermediate PSI (Figure 7E). Likewise, the PSI of PPA2 exon 6 decreases in the presence of SNP rs7672469, but this decrease is greater in tissues with more intermediate PSI levels (Figure 7E).

Global Scaling in Alternative 5′ and 3′ Splice-Site Choice

Although the mathematical model was built to describe the behavior of alternative exons, it should apply to any molecular process involving a mutually exclusive competition, like alternative splice-site selection. Indeed, the effect of mutations on an alternative 5′ splice-site choice (Rosenberg et al., 2015) depends on the starting splice-site usage (PSU) levels (Figures 7F, 7G, and S7A). To more globally assess this, we compared the PSU of thousands of alternative 3′ and 5′ splice sites in the genome across the same four pairs of conditions shown in Figure 7B, confirming that differences in PSU also scale non-monotonically (Figures 7H and S7B).

Taken together, therefore, global scaling is seen in many different datasets, for different types of splice-site choices, and for both *cis* and *trans* perturbations, including endogenous genes.

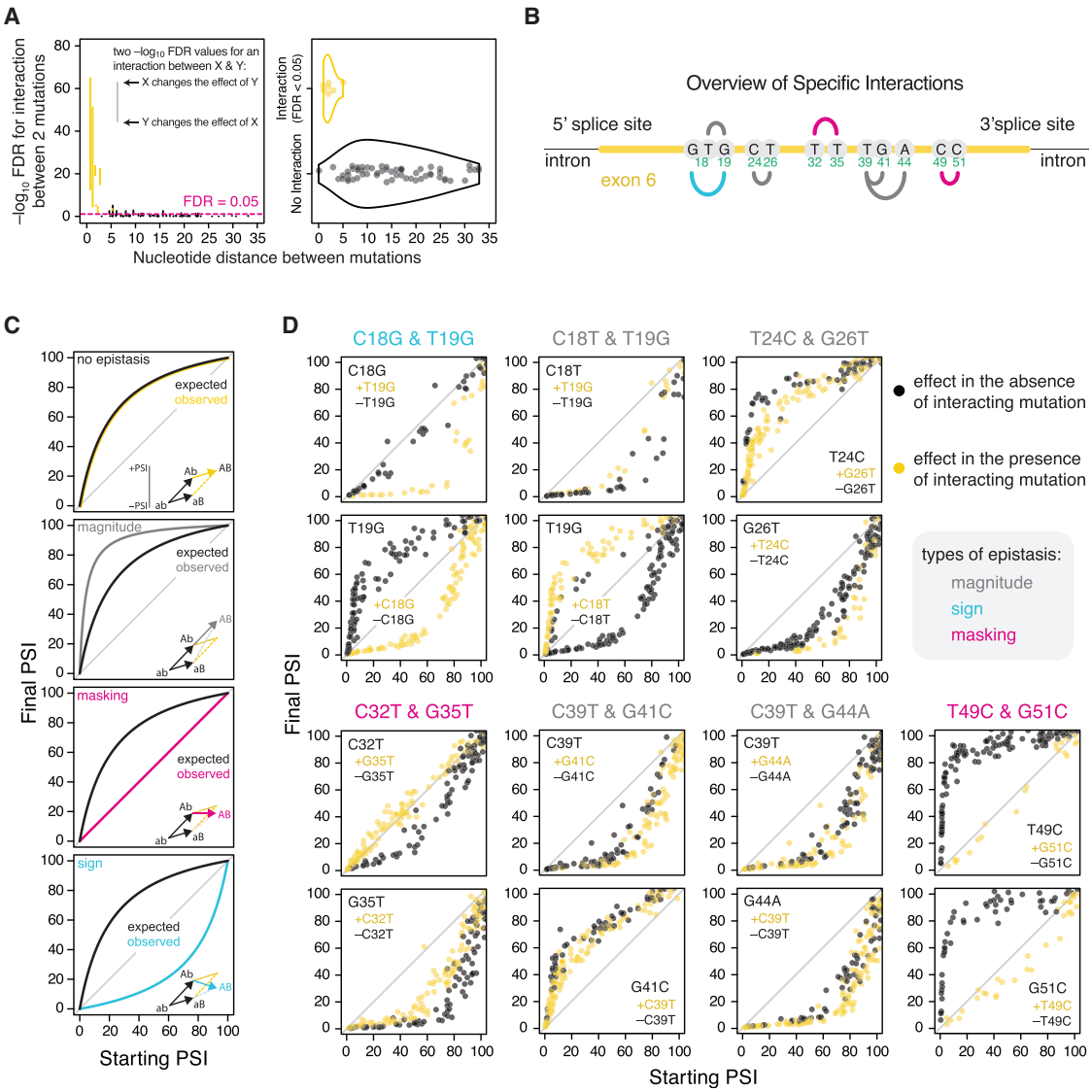


Figure 5. Specific Pairwise Interactions between Proximal Mutations

(A) Interacting mutations are proximal in the linear sequence of the exon.

For every pair of mutations, X and Y, 2 tests were performed to confirm whether the parameter (A) (Equation 2) of X is influenced by the presence of Y, and vice versa. An interaction was called if both tests were significant at an FDR <0.05. Left: bars spanning the two $-\log_{10}$ FDR values for each potential interaction. Yellow bars indicate both FDR values <0.05. Jitter was added to the x axis so overlapping bars can be visualized separately. Right: violin plots showing the distribution of distances between mutations that display a significant interaction (yellow) and between those that do not (black).

(B) The 7 interactions found. Colors indicate type of epistasis as in (C).

(C) Categories of epistasis.

(D) Behavior of mutations in the presence (yellow) or absence (black) of their interaction partner.

DISCUSSION

Exploring a Combinatorially Complete Genotype Space for the Evolution of an Alternatively Spliced Exon

Mutation libraries containing a complete subset of genotype space provide an opportunity to analyze the behavior of a mutation in thousands of closely related genetic contexts. Here, we have used this approach to systematically investigate how exonic mutations that occurred in the evolution of an alternatively spliced cassette exon—*FAS* exon 6—influence the inclusion of that exon in an mRNA.

Although many sequence and structural features have been shown to regulate alternative splicing (Barash et al., 2010; Chasin, 2007; Ke et al., 2011), a full understanding of the splicing code is far from being achieved (Xiong et al., 2015). Deep mutational scans of alternative exons have revealed the high density of information encoded in the sequence of individual exons (Braun et al., 2018; Julien et al., 2016; Ke et al., 2018; Soemedi

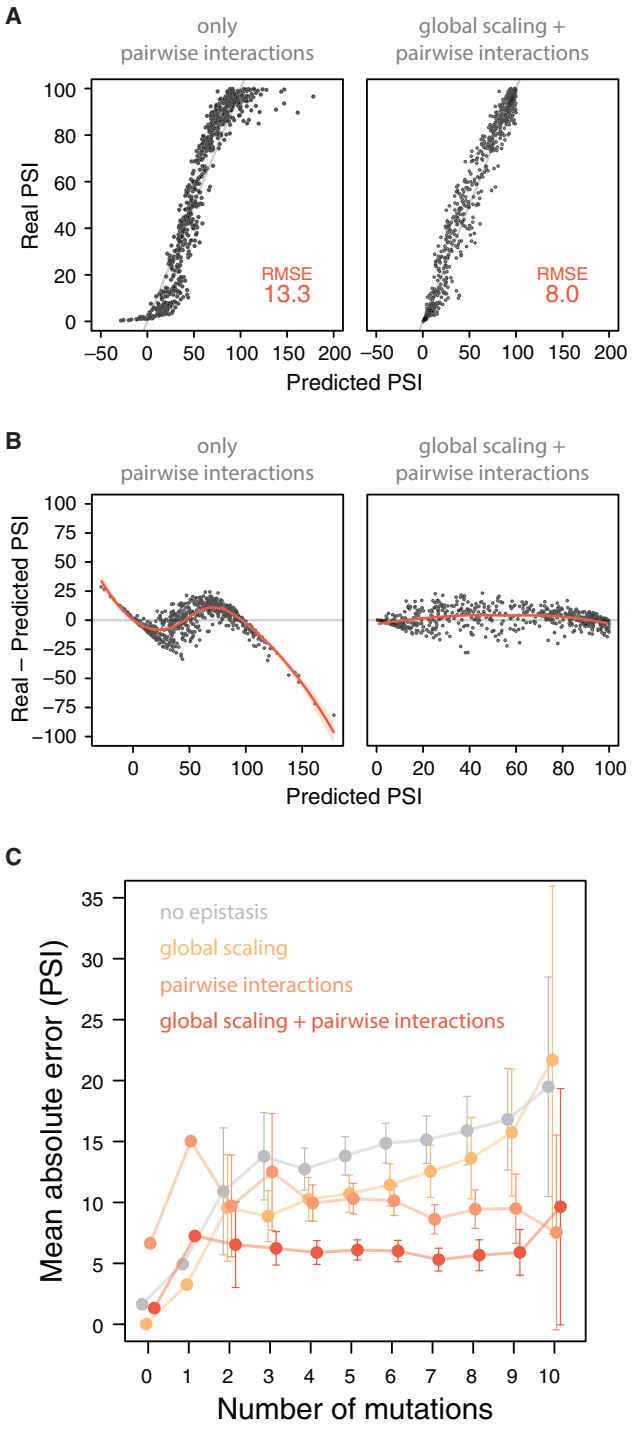


Figure 6. Combining Pairwise Interactions with Global Scaling to Achieve Accurate Genetic Prediction

(A) Real versus predicted PSI for a model that only considers pairwise interactions and one that also considers global scaling.

(B) Residuals plots with loess trend lines and 95% confidence bands, for the models shown in (A).

(C) Mean absolute error of different model predictions versus the number of mutations (relative to the ancestral sequence) in the genotype. Error bars indicate the 95% confidence intervals of the mean absolute error.

et al., 2017), and insights into the mechanisms behind the behavior of mutations in these assays have provided insights into how splicing regulatory information is decoded.

We found that the same mutation in closely related genetic contexts consistently has different effects on exon inclusion. However, these effects can be accurately predicted because they (1) follow a mathematically defined scaling law, and (2) display well-defined epistatic interactions with other proximal mutations. Specifically, mutations alter exon inclusion in a way that scales non-monotonically with the current level of exon inclusion. For any mutation, the impact on splicing is smallest when the current inclusion level is close to 0% or 100%, and increases progressively toward intermediate inclusion levels. In addition, the inclusion level at which a particular mutation has maximum effect is inversely and linearly related to the strength of the mutation (Figure 4F).

Non-monotonic Scaling May Arise Because Splice-Site Selection Is a Mutually Exclusive Molecular Event

Since the scaling law applies to mutations in different regions of the exon, likely having different molecular effects, it must arise from a general feature of splicing. We thus used a minimal mathematical model that captures the essence of splicing as an all-or-none molecular event at the level of individual mRNAs. This suggests that non-monotonic scaling arises because splice-site selection is a mutually exclusive molecular event with the competition between splicing to competing splice sites generating a sigmoidal relationship between the efficiency of one splicing outcome (e.g., inclusion of an exon) and the final percentage of isoform production.

Implications of the Scaling Law for Splicing Regulation

The scaling law has a number of practical implications. First, the evolution of a constitutive exon into an alternative exon will likely require multiple nt substitutions to *escape* the "inertia" of the scaling law that minimizes the effects of mutations when the exon is near full inclusion. Second, the same *trans*-acting perturbation can have different effects on different target exons, depending on their PSI. Knocking down SF3B1 leads to effects that scale non-monotonically with the starting exon inclusion level, and an hnRNP H knockdown has also been reported to have effects that scale according to the starting PSI (Braun et al., 2018).

In addition, the recent approval of Nusinersen (Spinraza), which corrects a splicing defect to treat spinal muscular atrophy (Talbot and Tizzano, 2017), has revitalized efforts to target splicing for therapeutic benefit. The scaling law can help to predict target splicing events or cellular states most sensitive to treatments as well as dose-response curves.

Specific Interactions Occur between Proximal Mutations

Pairwise interactions between mutations are sparse (7 out of 65 possible interactions), and all occur between mutations separated by <6 nt. As the binding sites of common RNA binding domains are 4–7 nt long (Daubner et al., 2013), this suggests mutations are having non-independent effects on the binding of individual or overlapping or adjacent *trans* factors. While our

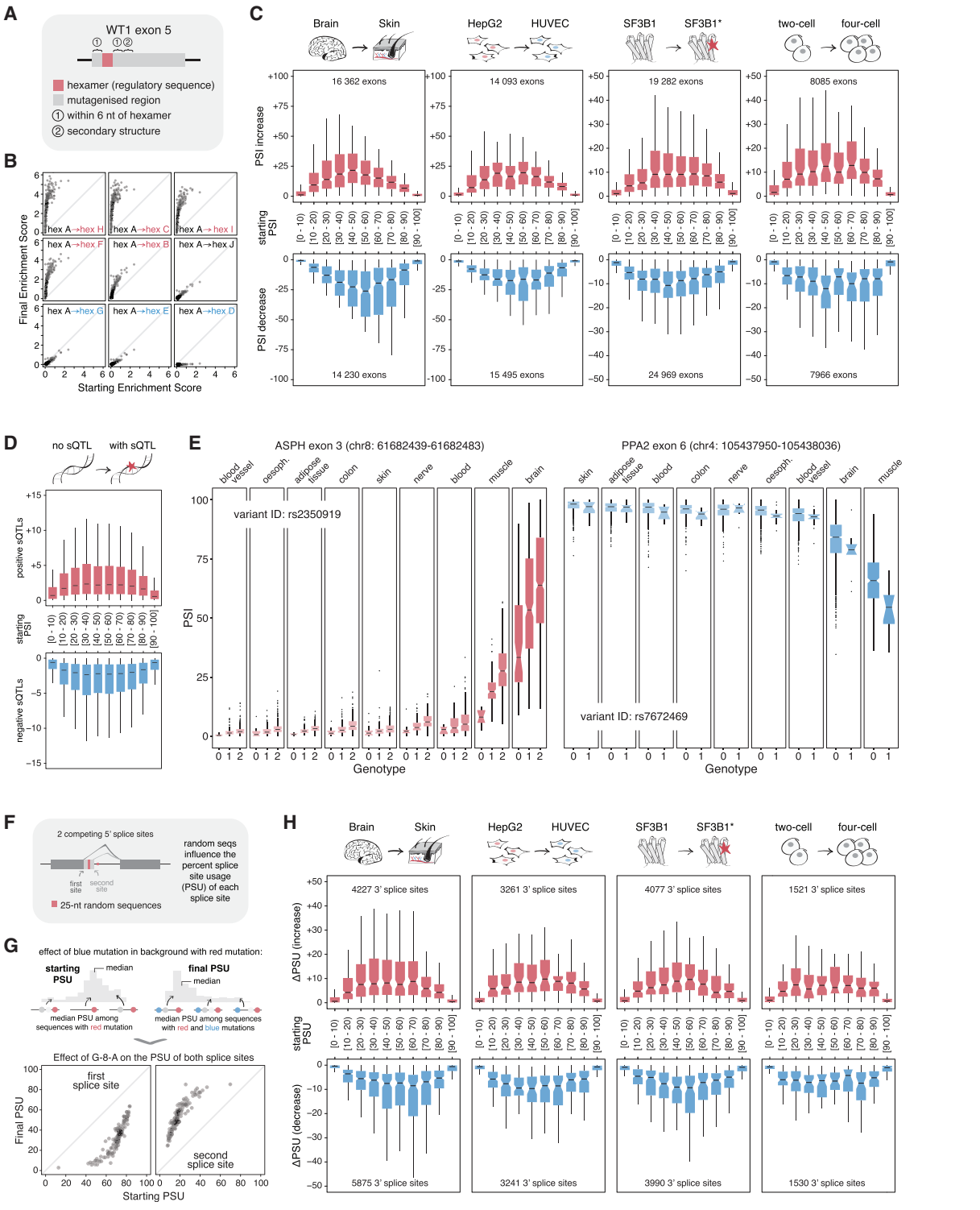


Figure 7. Generalizing the Scaling Law

(A) Mutant library of *WT1* exon 5 in the presence of different splicing regulatory sequences (hexamers).

(B) Comparing the enrichment scores (ESs) of *WT1* exon 5 genotypes in the presence of a wild-type hexamer sequence A (starting ES) with those in the presence of hexamer sequences B–J (final ES). To avoid epistatic effects, mutations within 6 nt from the hexamer and mutations in the region forming a secondary structure were removed from this analysis. Inclusion-promoting hexamers are labeled in red, skipping-promoting hexamers in blue, and neutral hexamers in black.

(legend continued on next page)

previous report (Julien et al., 2016) suggested long-distance epistatic interactions, most of these are likely a consequence of the general nonlinearity introduced into the landscape by the scaling law (Figure S5E; see below).

Three of the strongest interactions can illustrate potential mechanisms behind these epistatic effects. C32T and G35T are found in a region of the exon that binds to PTB (polypyrimidine tract-binding protein), which decreases *FAS* exon 6 inclusion (Izquierdo et al., 2005). These mutations increase skipping and each is predicted to increase the affinity for PTB binding (Figure S6F), but the double mutant results in less skipping than expected from the sum of the single-mutant effects (Figures 5D and S6A). One potential explanation is that, while each mutation increases PTB binding, binding affinity is no longer rate limiting for splicing regulation by PTB beyond a certain threshold. Previously, it has been shown that PTB can bind to a site centered on position 33 or to a site centered on position 36 (Mickleburgh et al., 2014). Whereas C32T is predicted to increase binding centered on position 33, G35T is predicted to increase binding centered on position 36 (Figure S6F). This suggests another model where positive epistasis arises because both strengthened sites cannot bind to PTB at the same time (Figure S6G). Such mechanisms may represent general causes of diminishing returns or antagonistic epistasis between *cis*-regulatory mutations.

T19G promotes inclusion in the absence of C18G but promotes skipping in the presence of C18G, an example of sign epistasis. In contrast, C18G has no effect in the absence of T19G but promotes skipping in the presence of T19G (Figures 5D and S6A). An explanation could be that T19G prevents the binding of a repressor, C18G has no effect on the binding of this repressor, and the double mutant creates a new binding site for the same or a new repressor (Figure S6H).

Finally, in the T49C–G51C interaction (Figures 5D and S6A), where each individual mutation promotes inclusion and the double mutant does not increase exon inclusion further, either mutation may be sufficient to prevent the binding of a regulatory factor. Introducing the second mutation therefore has no further effect (Figure S6I).

Global Scaling and Sparse Pairwise Interactions Are Sufficient for Accurate Genetic Prediction

A long-standing goal of genetics has been to accurately predict changes in phenotype from changes in genotype (Lehner, 2013). In particular, the extent to which pairwise and higher-order combinations of mutations are important for genetic prediction is not clear (Sailer and Harms, 2017; Weinreich et al., 2013). Here, we found that for the alternative splicing of a model exon, a relatively simple model can provide accurate genetic prediction for genotypes with up to 10 mutations. The model only contains a small number of specific second-order interactions but scales the effects of the individual mutations according to a global scaling law. If not explicitly accounted for, this global scaling will result in more complicated models with "phantom" pairwise and higher-order epistasis terms (Figure S5E). For example, a cross-validation RMSE of 11.0 PSI units can be achieved using a Lasso regression model with 98 parameters including 35/65 (53.8%) second- and 53/210 (25.2%) third-order interactions (STAR Methods).

Non-monotonic Scaling of Mutation Effects May Occur Quite Widely Because of Mutually Exclusive Molecular Events

Our mathematical model was built with alternative splicing in mind, but it can be used to simulate any other process involving a competition between mutually exclusive molecular events. Such competitions are common in biology, for example, between transcription factors binding to the same site (Darieva et al., 2010) or between 2 alternative protein interaction partners in a signal transduction cascade (Kiel et al., 2013). Moreover, sigmoidal relationships between molecular parameters and cellular phenotypes can be generated by many additional molecular mechanisms, for example, by cooperativity in molecular recruitment (Ackers et al., 1982) and in the folding (Tokuriki and Tawfik, 2009) and allostery (Ackers and Holt, 2006) of individual proteins. The scaling of mutation effects identified here could therefore be a widespread occurrence.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - *E. coli* strain for library subcloning
  - HEK293, HeLa and COS-7 cells for transfection
  - HEK293 FlpIn cells
- METHOD DETAILS
  - Phylogenetic analysis and ancestral sequence reconstruction
  - Analysis of FAS exon 6 inclusion across species
  - Combinatorially-complete library design and synthesis
  - Library amplification
  - Library subcloning
  - Input library

(C) Boxplots showing how genome-wide exon inclusion levels compare across 4 pairs of conditions, each involving a change in concentration or activity of splicing regulators. Notches display the confidence interval around the median, calculated as $1.58 * (\mathrm{IQR}/\sqrt{n})$, where IQR = interquartile range and n = number of data points. Upper whiskers extend from the 75th percentile to the largest value no further than 1.5 × IQR from the 75th percentile. Lower whiskers extend from the 25th percentile to the lowest value no further than 1.5 × IQR from the 25th percentile.
(D) How sQTL effects depend on the starting PSI.
(E) The effect of inclusion-promoting variant rs2350919 on ASPH exon 3 and that of skipping-promoting variant rs7672469 on PPA2 exon 6 both depend on the starting PSI.
(F) Mutant library of an intron with alternative 5′ splice sites (Rosenberg et al., 2015).
(G) Scaling of mutation effects in alternative splice-site choice. Scatterplots show the effect of G8A on the PSU of the 2 splice sites shown in (F).
(H) Boxplots showing how genome-wide alternative 3′ splice-site usage levels compare across the same 4 pairs of conditions described in (A).

- Cell transfection and generation of output libraries
- Sequencing
- Data processing and calculation of PSI values
- Cell culture and siRNA co-transfection
- Generation of output libraries for the siRNA experiment
- Sequencing the siRNA experiment
- Western blots
- Experimental validation of the sign epistasis interaction between C18G and T19G
- Assessment of the effects of mutants in HEK293 FlpIn cells
- Genetic prediction of combined mutation effects
- Simulated biallelic genotype landscape and "phantom" interactions
- Building a Lasso regression model with higher-order interactions and without global scaling
- PTB-Binding Motif Analysis
- Scaling of splicing perturbation effects in WT1 exon 5
- Genome-wide comparison of exon inclusion levels in two different tissues
- Genome-wide comparison of exon inclusion levels in two different cell lines
- Genome-wide comparison of exon inclusion levels in the presence and absence of mutations in SF3B1
- Genome-wide comparison of exon inclusion levels in two developmental stages
- sQTL analysis
- Alternative splice site usage in a mutant library
- Genome-wide alternative splice site usage across four pairs of conditions

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Quantification of FAS exon 6 inclusion in different primate species
  - Quantification of PSI values in the combinatorially-complete library
  - Experimental evaluation of PSI values
  - Estimating the maximum mutation effect size of each single mutation
  - Fitting the model to the Starting PSI – Final PSI datasets
  - Data Processing and Calculation of Enrichment Scores in the siRNA Libraries
  - Determining the PSI of Genotypes in the siRNA Libraries
  - Calculating RMSE values for the genetic prediction models
  - Statistical tests
  - Quantifying genome-wide exon inclusion exon inclusion levels in two different tissues
  - Quantifying genome-wide exon inclusion exon inclusion levels in two different cell lines
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures, six tables, and one data file and can be found with this article online at https://doi.org/10.1016/j.cell.2018.12.010.

ACKNOWLEDGMENTS

We thank Diego Garrido Martín for guidance in the sQTL analysis and Yamile Márquez for guidance with VAST-TOOLS. Work in B.L.'s is supported by a European Research Council (ERC) Consolidator grant (616434), the Spanish Ministry of Economy and Competitiveness (BFU2017-89488-P and SEV-2012-0208), the AXA Research Fund, the Bettencourt Schueller Foundation, Agencia de Gestio d'Ajuts Universitaris i de Recerca (AGAUR, SGR-831), the EMBL Partnership, and the CERCA Program/Generalitat de Catalunya. P.B.-C. was funded in part by a Severo Ochoa PhD fellowship and J.M.S. by an EMBO long-term fellowship (ALTF 857-2016). Work in J.V.'s laboratory is supported by Fundación Botín, Banco de Santander through its Santander Universities Global Division, ERC AdvG 670146, AGAUR, Spanish Ministry of Economy and Competitiveness (BFU 2014-005153, BFU 2017 89308-P, and SEV-2012-0208), the EMBL Partnership, and the CERCA program/Generalitat de Catalunya. The Genotype-Tissue Expression (GTEx) data used for the analyses described in this manuscript were obtained from the GTEx Portal on May 8, 2018 and dbGaP accession number phs000424.v7.p2 on May 8, 2018. The GTEx Project was supported by the Common Fund of the Office of the Director of the NIH and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

AUTHOR CONTRIBUTIONS

P.B.-C., B.M., J.V., and B.L. conceived and designed the study. B.M. generated the experimental data. P.B.-C. analyzed the data. P.B.-C. and J.M.S. built the mathematical model. B.L. and J.V. supervised the study. P.B.-C. and B.L. wrote the paper with input from J.M.S., B.M., and J.V.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 26, 2018
Revised: August 29, 2018
Accepted: December 7, 2018
Published: January 17, 2019

REFERENCES

Ackers, G.K., and Holt, J.M. (2006). Asymmetric cooperativity in a symmetric tetramer: Human hemoglobin. J. Biol. Chem. *281*, 11441–11443.

Ackers, G.K., Johnson, A.D., and Shea, M.A. (1982). Quantitative model for gene regulation by lambda phage repressor. Proc. Natl. Acad. Sci. USA *79*, 1129–1133.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53–59.

Battle, A., Brown, C.D., Engelhardt, B.E., and Montgomery, S.B.; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis &Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; eQTL manuscript working group (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Braun, S., Enculescu, M., Setty, S.T., Cortés-López, M., de Almeida, B.P., Sutandy, F.X.R., Schulz, L., Busch, A., Seiler, M., Ebersberger, S., et al. (2018). Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. Nat. Commun. *9*, 3315.

Cascino, I., Fiucci, G., Papoff, G., and Ruberti, G. (1995). Three functional soluble forms of the human apoptosis-inducing Fas molecule are produced by alternative splicing. J. Immunol. *154*, 2706–2713.

Chasin, L.A. (2007). Searching for splicing motifs. Adv. Exp. Med. Biol. *623*, 85–106.

Daguenet, E., Dujardin, G., and Valcárcel, J. (2015). The pathogenicity of splicing defects: Mechanistic insights into pre-mRNA processing inform novel therapeutic approaches. EMBO Rep. *16*, 1640–1655.

Darieva, Z., Clancy, A., Bulmer, R., Williams, E., Pic-Taylor, A., Morgan, B.A., and Sharrocks, A.D. (2010). A competitive transcription factor binding mechanism determines the timing of late cell cycle-dependent gene expression. Mol. Cell *38*, 29–40.

Daubner, G.M., Cléry, A., and Allain, F.H.T. (2013). RRM-RNA recognition: NMR or crystallography…and new findings. Curr. Opin. Struct. Biol. *23*, 100–108.

Diss, G., and Lehner, B. (2018). The genetic landscape of a physical interaction. eLife *7*, 1–31.

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

Dolatshad, H., Pellagatti, A., Fernandez-Mercado, M., Yip, B.H., Malcovati, L., Attwood, M., Przychodzen, B., Sahgal, N., Kanapin, A.A., Lockstone, H., et al. (2015). Disruption of SF3B1 results in deregulated expression and splicing of key genes and pathways in myelodysplastic syndrome hematopoietic stem and progenitor cells. Leukemia *29*, 1092–1103.

Domingo, J., Diss, G., and Lehner, B. (2018). Pairwise and higher-order genetic interactions during the evolution of a tRNA. Nature *558*, 117–121.

Dvir, S., Velten, L., Sharon, E., Zeevi, D., Carey, L.B., Weinberger, A., and Segal, E. (2013). Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. Proc. Natl. Acad. Sci. USA *110*, E2792–E2801.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Förch, P., Puig, O., Kedersha, N., Martínez, C., Granneman, S., Séraphin, B., Anderson, P., and Valcárcel, J. (2000). The apoptosis-promoting factor TIA-1 is a regulator of alternative pre-mRNA splicing. Mol. Cell *6*, 1089–1098.

Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D., and Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. Nat. Methods *7*, 741–746.

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. *27*, 221–224.

Hastie, T., and Stuetzle, W. (1989). Principal curves. J. Am. Stat. Assoc. *84*, 502–516.

Havens, M.A., Duelli, D.M., and Hastings, M.L. (2013). Targeting RNA splicing for disease therapy. Wiley Interdiscip. Rev. RNA *4*, 247–266.

Holmqvist, E., Reimegård, J., and Wagner, E.G.H. (2013). Massive functional mapping of a 5′-UTR by saturation mutagenesis, phenotypic sorting and deep sequencing. Nucleic Acids Res. *41*, e122.

Izquierdo, J.M., Majós, N., Bonnal, S., Martínez, C., Castelo, R., Guigó, R., Bilbao, D., and Valcárcel, J. (2005). Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. Mol. Cell *19*, 475–484.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI BLAST: A better web interface. Nucleic Acids Res. *36*, W5–W9.

Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J., and Lehner, B. (2016). The complete local genotype-phenotype landscape for the alternative splicing of a human exon. Nat. Commun. *7*, 11558.

Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. *21*, 1360–1374.

Ke, S., Anquetil, V., Zamalloa, J.R., Maity, A., Yang, A., Arias, M.A., Kalachikov, S., Russo, J.J., Ju, J., and Chasin, L.A. (2018). Saturation mutagenesis reveals manifold determinants of exon definition. Genome Res. *28*, 11–24.

Kiel, C., Verschueren, E., Yang, J.S., and Serrano, L. (2013). Integration of protein abundance and structure data reveals competition in the ErbB signaling network. Sci. Signal. *6*, ra109.

Kinney, J.B., Murugan, A., Callan, C.G., Jr., and Cox, E.C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc. Natl. Acad. Sci. USA *107*, 9158–9163.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. *34*, 1812–1819.

Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. Trends Genet. *27*, 323–331.

Lehner, B. (2013). Genotype to phenotype: Lessons from model organisms for human genetics. Nat. Rev. Genet. *14*, 168–178.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, C., Qian, W., Maclean, C.J., and Zhang, J. (2016). The fitness landscape of a tRNA gene. Science *352*, 837–840.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr., Kinney, J.B., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. *30*, 271–277.

Merkin, J., Russell, C., Chen, P., and Burge, C.B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science *338*, 1593–1599.

Mickleburgh, I., Kafasla, P., Cherny, D., Llorian, M., Curry, S., Jackson, R.J., and Smith, C.W.J. (2014). The organization of RNA contacts by PTB for regulation of FAS splicing. Nucleic Acids Res. *42*, 8605–8620.

Olson, C.A., Wu, N.C., and Sun, R. (2014). A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Curr. Biol. *24*, 2643–2651.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat. Biotechnol. *27*, 1173–1175.

Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.I., Cooper, G.M., et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. *30*, 265–270.

Peng, X., Thierry-Mieg, J., Thierry-Mieg, D., Nishida, A., Pipes, L., Bozinoski, M., Thomas, M.J., Kelly, S., Weiss, J.M., Raveendran, M., et al. (2015). Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHPRTR). Nucleic Acids Res. *43*, D737–D742.

Phillips, P.C. (2008). Epistasis—The essential role of gene interactions in the structure and evolution of genetic systems. Nat. Rev. Genet. *9*, 855–867.

Poelwijk, F.J., Krishna, V., and Ranganathan, R. (2016). The context-dependence of mutations: A linkage of formalisms. PLoS Comput. Biol. *12*, e1004771.

Puchta, O., Cseke, B., Czaja, H., Tollervey, D., Sanguinetti, G., and Kudla, G. (2016). Network of epistatic interactions within a yeast snoRNA. Science *352*, 840–844.

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172–177.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell *163*, 698–711.

Sailer, Z.R., and Harms, M.J. (2017). High-order epistasis shapes evolutionary trajectories. PLoS Comput. Biol. *13*, e1005541.

Saraiva-Agostinho, N., and Barbosa-Morais, N.L. (2018). psichomics: Graphical application for alternative splicing quantification and analysis. Nucleic Acids Res. Published online October 2, 2018. https://doi.org/10.1093/nar/gky888.

Sarkisyan, K.S., Bolotin, D.A., Meer, M.V., Usmanova, D.R., Mishin, A.S., Sharonov, G.V., Ivankov, D.N., Bozhanova, N.G., Baranov, M.S., Soylemez, O., et al. (2016). Local fitness landscape of the green fluorescent protein. Nature *533*, 397–401.

Schafer, S., Miao, K., Benson, C.C., Heinig, M., Cook, S.A., and Hubner, N. (2015). Alternative splicing signatures in RNA-seq data: Percent spliced in (PSI). Curr. Protoc. Hum. Genet., 11.16.1–11.16.14.

Shalem, O., Sharon, E., Lubliner, S., Regev, I., Lotan-Pompan, M., Yakhini, Z., and Segal, E. (2015). Systematic dissection of the sequence determinants of gene 3′ end mediated expression control. PLoS Genet. *11*, e1005147.

Shendure, J., and Akey, J.M. (2015). The origins, determinants, and consequences of human mutations. Science *349*, 1478–1483.

Soemedi, R., Cygan, K.J., Rhine, C.L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W.G. (2017). Pathogenic variants that alter protein code often disrupt splicing. Nat. Genet. *49*, 848–855.

Talbot, K., and Tizzano, E.F. (2017). The clinical landscape for SMA in a new therapeutic era. Gene Ther. *24*, 529–533.

Tapial, J., Ha, K.C.H., Sterne-Weiler, T., Gohr, A., Braunschweig, U., Hermoso-Pulido, A., Quesnel-Vallières, M., Permanyer, J., Sodaei, R., Marquez, Y., et al. (2017). An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. *27*, 1759–1768.

Tejedor, J.R., Papasaikas, P., and Valcárcel, J. (2015). Genome-wide identification of Fas/CD95 alternative splicing regulators reveals links with iron homeostasis. Mol. Cell *57*, 23–38.

Tokuriki, N., and Tawfik, D.S. (2009). Stability effects of mutations and protein evolvability. Curr. Opin. Struct. Biol. *19*, 596–604.

Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. Science *347*, 1260419–1260419.

Veidenberg, A., Medlar, A., and Löytynoja, A. (2016). Wasabi: An integrated platform for evolutionary sequence analysis and data visualization. Mol. Biol. Evol. *33*, 1126–1130.

Weinreich, D.M., Watson, R.A., and Chao, L. (2005). Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evolution *59*, 1165–1174.

Weinreich, D.M., Lan, Y., Wylie, C.S., and Heckendorn, R.B. (2013). Should evolutionary geneticists worry about higher-order epistasis? Curr. Opin. Genet. Dev. *23*, 700–707.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. *20*, 1131–1139.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614–620.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit polyclonal anti-SF3B1	Abcam	Catalogue #: ab39578
Mouse monoclonal anti-GAPDH	Abcam	Catalogue #: ab8245
Rabbit ECL IgG, HRP-Linked Whole Ab	GE Healthcare	Catalogue #: NA9340
Mouse ECL IgG, HRP-Linked Whole Ab	GE Healthcare	Catalogue #: NA931
Bacterial and Virus Strains		
Stellar competent cells (*E. coli* HST08 strain)	Clontech	Catalogue #: 636766
Chemicals, Peptides, and Recombinant Proteins		
Accuprime Pfx DNA polymerase	ThermoFisher Scientific	Catalogue #: 12344024
GoTaq flexi DNA polymerase	Promega	Catalogue #: M7806
Lipofectamine 2000	ThermoFisher Scientific	Catalogue #: 11668027
Opti-MEM I reduced serum medium	ThermoFisher Scientific	Catalogue #: 31985-047
DMEM Glutamax	ThermoFisher Scientific	Catalogue #: 61965059
Foetal bovine serum	ThermoFisher Scientific	Catalogue #: 10270
Penicillin-Streptomycin	ThermoFisher Scientific	Catalogue #: 15070063
SYBR safe DNA gel stain	ThermoFisher Scientific	Catalogue #: S33102
Opti-MEM	ThermoFisher Scientific	Catalogue #: 13778150
DMEM F12	ThermoFisher Scientific	Catalogue #: 31330038
complete protease inhibitor	Roche	Catalogue #: 11697498001
Western Lightning Plus ECL chemiluminescence reagent	PerkinElmer	Catalogue #: NEL105001EA
Opti-MEM I reduced serum medium without phenol red	ThermoFisher Scientific	Catalogue #: 11058021
DMEM + GlutaMAX	ThermoFisher Scientific	Catalogue #: 61965-059
Hygromycin B	ThermoFisher Scientific	Catalogue #: 10687-010
Blasticidin	ThermoFisher Scientific	Catalogue #: A1113903
Doxycycline	CONDA	Catalogue #: 631311
Critical Commercial Assays		
In-Fusion HD cloning kit	Clontech	Catalogue #: 639649
Plasmid DNA purification maxi kit	Quiagen	Catalogue #: 50912163
Gel extraction kit	Quiagen	Catalogue #: 50928704
Maxwell LEV 16 simplyRNA tissue kit	Promega	Catalogue #: AS1280
Whatman Protran 0.2 um nitrocellulose	GE Healthcare	Catalogue #: 106000001
Kodak BioMax MR film	Sigma-Aldrich	Catalogue #: Z353949
Deposited Data		
Raw sequencing reads for combinatorially complete mutant library	This paper	GEO: GSE111316 (replicates split into different fastq files) ENA: PRJEB24588 (replicates in the same fastq file)
Raw sequencing reads for doped mutant library in the presence of siRNA against SF3B1 or control siRNA	This paper	GEO: GSE111316 (replicates split into different fastq files) ENA: PRJEB24588 (replicates in the same fastq file)
Tissue-specific RNA-Seq data from nonhuman primates	Peng et al., 2015	http://www.nhprtr.org/
Human Protein Atlas tissue-specific RNA-Seq data	Uhlen et al., 2015	http://www.proteinatlas.org/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Human Illumina Body Map tissue-specific RNA-Seq data	Illumina	https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/
Mus musculus tissue-specific RNA-Seq data	Merkin et al., 2012	GEO: GSE41637
Homo sapiens (GRCh38) genome sequence	Ensembl	ftp://ftp.ensembl.org/pub/release-84/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.toplevel.fa.gz
Pan troglodytes genome sequence	Ensembl	ftp://ftp.ensembl.org/pub/release-84/fasta/pan_troglodytes/dna/Pan_troglodytes.CHIMP2.1.4.dna.toplevel.fa.gz
Papio anubis genome sequence	Ensembl	ftp://ftp.ensembl.org/pub/release-84/fasta/papio_anubis/dna/Papio_anubis.PapAnu2.0.dna.toplevel.fa.gz
Macaca mulatta genome sequence	Ensembl	ftp://ftp.ensembl.org/pub/release-84/fasta/macaca_mulatta/dna/Macaca_mulatta.MMUL_1.dna.toplevel.fa.gz
Callithrix jacchus genome sequence	Ensembl	ftp://ftp.ensembl.org/pub/release-84/fasta/callithrix_jacchus/dna/Callithrix_jacchus.C_jacchus3.2.1.dna.toplevel.fa.gz
Microcebus murinus genome sequence	Ensembl	ftp://ftp.ensembl.org/pub/release-84/fasta/microcebus_murinus/dna/Microcebus_murinus.micMur1.dna.toplevel.fa.gz
Mus musculus genome sequence	Ensembl	ftp://ftp.ensembl.org/pub/release-84/fasta/mus_musculus/dna/Mus_musculus.GRCm38.dna.toplevel.fa.gz
Homo sapiens (GRCh38) genome annotations	Ensembl	ftp://ftp.ensembl.org/pub/release-84/gtf/homo_sapiens/Homo_sapiens.GRCh38.84.gtf.gz
Pan troglodytes genome annotations	Ensembl	ftp://ftp.ensembl.org/pub/release-84/gtf/pan_troglodytes/Pan_troglodytes.CHIMP2.1.4.84.gtf.gz
Papio anubis genome annotations	Ensembl	ftp://ftp.ensembl.org/pub/release-84/gtf/papio_anubis/Papio_anubis.PapAnu2.0.84.gtf.gz
Macaca mulatta genome annotations	Ensembl	ftp://ftp.ensembl.org/pub/release-84/gtf/macaca_mulatta/Macaca_mulatta.MMUL_1.84.gtf.gz
Callithrix jacchus genome annotations	Ensembl	ftp://ftp.ensembl.org/pub/release-84/gtf/callithrix_jacchus/Callithrix_jacchus.C_jacchus3.2.1.84.gtf.gz
Microcebus murinus genome annotations	Ensembl	ftp://ftp.ensembl.org/pub/release-84/gtf/microcebus_murinus/Microcebus_murinus.micMur1.84.gtf.gz
Mus musculus genome annotations	Ensembl	ftp://ftp.ensembl.org/pub/release-84/gtf/mus_musculus/Mus_musculus.GRCm38.84.gtf.gz
M227 and M228 position weight matrices	Ray et al., 2013	http://cisbp-rna.ccbr.utoronto.ca/
WT1 exon 5 mutant enrichment scores table	Ke et al., 2018	https://genome.cshlp.org/content/suppl/2017/12/14/gr.219683.116.DC1/Supplemental_Table_S2.xlsx
GTEx junction read counts file	Battle et al., 2017	https://gtexportal.org/home/datasets
RNA-Seq data from HepG2 cells	ENCODE Project Consortium, 2012	GEO: GSM2308416 and GSM2308417
RNA-Seq data from Huvec cells	ENCODE Project Consortium, 2012	GEO: GSM2072423 and GSM2072424

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
RNA-Seq data from myelodysplastic syndrome	Dolatshad et al., 2015	GEO: GSE63569
Single-cell RNA-Seq data from human embryos as different developmental stages	Yan et al., 2013	GEO: GSE36552
GTEx genotype matrix file	Battle et al., 2017	dbGap: phs000424.v7.p2
Homo sapiens (GRCh37) genome annotations for sQTL analysis with GTEx data	GENCODE	ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/gencode.v19.annotation.gff3.gz
Processed read counts file for an alternative 5′ splice site mutant library	Rosenberg et al., 2015	https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE74070&format=file
Experimental Models: Cell Lines		
HEK293	ATCC	ATCC CRL-1573
HeLa	ATCC	ATCC CCL-2
COS-7	ATCC	ATCC CRL-1651
HEK293 Flp-In T-Rex cells	ThermoFisher Scientific	Catalogue #: R780-07
Oligonucleotides		
Combinatorially complete library of *FAS* exon 6 mutants: TGTCCAATGTTCCAACCTACAGGATCCAGATCTAACT TGBKGTGGYTKTGTCTYCTKCTTYTSCCRATTCYAST AATTGTTTGGGGTAAGTTCTTGCTTTGTTCAAACTGC AGATTGAAATAACTTGGGAAGTAG	IBA Gmbh	N/A
Forward primer for library amplification FAS_i5_GC_F: TGTCCAATGTTCCAACCTACAG	This paper	N/A
Rverse primer for library amplification FAS_i6_GC_R: CTACTTCCCAAGTTATTTCAATCTG	This paper	N/A
Forward primer for ampliseq sequencing of input library FAS_i5_TR_F: AAAATGTCCAATGTTCCAACC	This paper	N/A
Forward primer for ampliseq sequencing of input library FAS_i5_TR_R: TGCAGTTTGAACAAAGCAAGA	This paper	N/A
Forward primer for ampliseq sequencing of output library FAS_e5_BR_F: CAGCAACACCAAGTGCAAAG	This paper	N/A
Reverse primer for ampliseq sequencing of output library FAS_e5_BR_R: TGCATGTTTTCTGTACTTCCTTTC	This paper	N/A
Primers used for RT-PCR (PT1): GTCGACGACACTTGCTCAAC	This paper	N/A
Primers used for RT-PCR (PT2): AAGCTTGCATCGAATCAGTAG	This paper	N/A
Mixture of siRNA oligonucleotides against SF3B1 (On-TARGETplus SMARTpool siRNA against SF3B1)	Dharmacon	Catalogue #: L-020061-0005
siRNA control oligonucleotide AAGGUCCGGCUCCCCCAAAUG	Sigma-Aldrich	N/A
Recombinant DNA		
pCMV FAS wt minigene exon 5-6-7	This paper	N/A
Doped library of FAS exon 6 single/double mutants	Julien et al., 2016	N/A
Software and Algorithms		
Blastn	Johnson et al., 2008	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Seaview	Gouy et al., 2010	http://doua.prabi.fr/software/seaview
Clustal Omega		http://doua.prabi.fr/software/seaview
PAGAN	Veidenberg et al., 2016	http://wasabiapp.org/software/pagan/
TimeTree	Kumar et al., 2017	http://www.timetree.org/
STAR v2.5.2a	Dobin et al., 2013	https://github.com/alexdobin/STAR/releasesh
SAMtools v1.3.1	Li et al., 2009	https://sourceforge.net/projects/samtools/files/samtools/1.3.1/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PSI.sh script	Schafer et al., 2015	https://github.com/pablo-baeza/Baeza_et_al_2018/blob/master/001_Exon_inclusion_levels_in_different_animals/PSI.sh
ImageJ v1.47	NIH	https://imagej.nih.gov/ij/download.html
SABRE demultiplexer	github/najoshi	https://github.com/najoshi/sabre
PEAR merger	Zhang et al., 2014	https://cme.h-its.org/exelixis/web/software/pear
Seqtk	Heng Li	https://github.com/lh3/seqtk
FASTX-toolkit	Hannon lab	http://hannonlab.cshl.edu/fastx_toolkit
PrimerX	Carlo Lapid and Yimin Gao	http://www.bioinformatics.org/primerx/
VAST-TOOLS	Tapial et al., 2017	https://github.com/vastgroup/vast-tools
BCFTools	Genome Research Limited	http://www.htslib.org/
R v3.3.3	The R Foundation	https://www.r-project.org/
Analog package in R	CRAN	https://cran.r-project.org/package=analogue
Caret package in R	CRAN	https://cran.r-project.org/package=caret
Glmnet package in R	CRAN	https://cran.r-project.org/package=glmnet
GGplot2 package in R	CRAN	https://cran.r-project.org/package=ggplot2
Psichomics package in R	Bioconductor	https://bioconductor.org/packages/release/bioc/html/psichomics.html
Lmtest package in R	CRAN	https://cran.r-project.org/package=lmtest
Other		
Resource web page with all the scripts needed to reproduce the computational analyses	This paper	https://github.com/lehner-lab/Scaling_Law

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Ben Lehner (ben.lehner@crg.eu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

E. coli strain for library subcloning

E. coli cells used to build the combinatorially-complete library (see Library subcloning subsection in the METHOD DETAILS) were Stellar competent cells (636766, Clontech) and grown for 18 hours in LB medium containing ampicillin. Cells were cultured at 37°C.

HEK293, HeLa and COS-7 cells for transfection

Mycoplasma-free tested HEK293 cells (CRL-1573, ATCC), HeLa cells (CCL-2, ATCC) or COS-7 (CRL-1651, ATCC) cells were grown in Lipofectamine 2000 (11668027, ThermoFisher Scientific) and Opti-MEM I Reduced Serum Medium (31985-047, ThermoFisher Scientific). Six hours post-transfection, the cell culture medium was replaced with DMEM Glutamax (61965059, ThermoFisher Scientific) containing 10% FBS and Pen/Strep antibiotics, and cells were allowed to grow for 48 hours. HEK293 and HeLa cells are female, COS-7 cells are male. All cells were grown at 37°C. After receiving the cells from ATCC, they were not authenticated.

HEK293 FlpIn cells

HEK293 Flp-In T-Rex cells (ThermoFisher Scientific, R780-07) were grown in 6-well plates using Lipofectamine 2000 with Opti-MEM I reduced serum medium without phenol red (ThermoFisher Scientific, 11058021). After 6 hours, the medium was changed to DMEM + GlutaMAX (ThermoFisher Scientific, 61965-059) in the presence of 10% fetal bovine serum and pre-strep antibiotics. Stable transfectants were selected with 100 μg/ml hygromycin B (ThermoFisher Scientific, 10687-010) and 5 μg/ml blasticidin (ThermoFisher Scientific A1113903). Polyclonal populations were grown in the presence of 5 μg/ml blasticidin for 10 days. For each genotype, six individual clones were picked with a pipette tip and moved to wells in a 6-well plate. They were grown in the presence of 5 μg/ml blasticidin and 100 μg/ml hygromycin B. HEK293 Flp-In T-Rex cells are female. Cells were grown at 37°C. Cells were not authenticated.

METHOD DETAILS

Phylogenetic analysis and ancestral sequence reconstruction

The sequence corresponding to the human genomic region covering *FAS* exon 5 to *FAS* exon 7 was downloaded from the UCSC genome browser and Blastn (Johnson et al., 2008) was employed to identify the orthologous sequences in primates, colugos and treeshrews. By aligning each genomic sequence (containing exons and introns) to its corresponding mRNA sequence (containing only exons), the orthologs of exon 6 in different species were identified and their limits defined. A multiple sequence alignment of all the exon sequences was built with the Seaview software (Gouy et al., 2010) implementation of the Clustal Omega algorithm.

The sequences of the *FAS* exon 6 evolutionary intermediates were inferred using a maximum-likelihood algorithm implemented by the PAGAN software from the Wasabi suite (Veidenberg et al., 2016). This program requires a guide tree that can be provided by the user. For this, a phylogenetic tree containing all the species in our multiple sequence alignment was downloaded from TimeTree (Kumar et al., 2017). Tarsier and lemur sequences contain insertions and deletions (indels) not found in primates more closely related to humans (Figure S1A). The reconstructed evolutionary intermediates ancestral to tarsiers and lemurs contained these insertions and deletions. However, these indels are not present in mammals more distantly related to humans such as colugos and treeshrews (Figure S1A). Therefore, these indels were manually removed from the reconstructed sequences.

Analysis of FAS exon 6 inclusion across species

Tissue-specific RNA-Seq data from nonhuman primates was downloaded from the Nonhuman Primate Reference Transcriptome Resource (www.nhprtr.org, Peng et al., 2015). Tissue-specific RNA-Seq data from humans was downloaded from the Human Protein Atlas (www.proteinatlas.org, Uhlen et al., 2015) and the Human Illumina Body Map (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/). The *PSI.sh* pipeline described by Schafer et al. (2015) was used with the RNA-Seq aligner STAR v2.5.2a (Dobin et al., 2013) and SAMTools (Li et al., 2009) to measure the inclusion levels of *FAS* exon 6 in each sample (for more details, see the QUANTIFICATION AND STATISTICAL ANALYSIS section).

Combinatorially-complete library design and synthesis

A sequence library was designed to include the 63-nucleotide-long ancestral exon sequence, with 10 positions allowing for any of the two nucleotides observed at those specific sites throughout evolution, and one position allowing for the three nucleotides observed at that position, flanked by 22 nucleotides of the 3′ end of the human intron 5 and 50 nucleotides of the 5′ end of the human intron 6:

5′-TGT CCA ATG TTC CAA CCT ACA GGA TCC AGA TCT AAC TTG **BK**G TGG **Y**T**K** TGT CT**Y** CT**K** CTT **YTS** CC**R** ATT C**Y**A **S**TA ATT GTT TGG GGT AAG TTC TTG CTT TGT TCA AAC TGC AGA TTG AAA TAA CTT GGG AAG TAG −3′

Letters in bold indicate the varied positions. B stands for either of C, G or T; K stands for either of T or G; Y stands for either of C or T; S stands for either of G or C; R stands for either of A or G. The library was synthesized and purified by Reverse Phase HPLC by IBA Gmbh, and ordered on the 0.2 μmol scale.

Library amplification

Accuprime Pfx (12344024, ThermoFisher Scientific) was used following the manufacturer's instructions to amplify 20 ng of single-stranded library DNA for 25 cycles with the following flanking intronic primers: FAS_i5_GC_F and FAS_i6_GC_R ("Primers used for library amplification" in Table S5).

Library subcloning

The amplified library was recombined with pCMV FAS wt minigene exon 5-6-7 (Förch et al., 2000) using the In-Fusion HD Cloning kit (639649, Clontech) in a 1:8 vector:insert optimized ratio and transformed into Stellar competent cells (636766, Clontech) to maximize the number of individual transformants, which was around 800,000 individual clones per library. The library was amplified by growing for 18 hours in LB medium containing ampicillin and the final plasmid library was purified using the Quiagen plasmid maxi kit (50912163, Quiagen) and quantified with a NanoDrop spectrophotometer.

Input library

20 ng of the library was amplified in triplicates using GoTaq flexi DNA polymerase (M7806, Promega) for 25 cycles with three pairs of intronic primers FAS_i5_BR_F and FAS_i5_BR_R ("Primers used for Ampliseq sequencing" in Table S5), leading to a 135 nucleotide PCR band that was gel-purified and sequenced. Each pair of primers had a distinct 8-mer barcode sequence to discriminate between technical replicates ("Primers used for Ampliseq sequencing" in Table S5).

Cell transfection and generation of output libraries

For each of the nine experimental replicates, 10 ng of the library were transfected into 250 000 HEK293 cells in one well of a 6-well plate using Lipofectamine 2000 (11668027, ThermoFisher Scientific) and OPTIMEM I Reduced Serum Medium (31985-047, ThermoFisher Scientific). Six hours post-transfection, the cell culture medium was replaced with DMEM Glutamax (61965059,

ThermoFisher Scientific) containing 10% FBS and Pen/Strep antibiotics. 48 hours post-transfection, total RNA was isolated using the automated Maxwell LEV 16 simplyRNA tissue kit (AS1280, Promega). cDNA was synthetized from 500 ng total RNA using Superscript III (18080085, Life Technologies), and amplified with one of the nine pairs of barcoded FAS_e5_BR_F and FAS_e5_BR_R primers ("Primers used for Ampliseq sequencing" in Table S5) and GoTaq flexi DNA polymerase (M7806, Promega). Each pair of primers had a distinct 8-mer barcode sequence to distinguish the nine experimental replicates ("Primers used for Ampliseq sequencing" in Table S5). PCR products were run on a 2% agarose gel and the band corresponding in size to the amplification product of exon inclusion was excised, purified using the Quiaquick Gel extraction kit (Quiagen, 50928704) and quantified with a NanoDrop spectrophotometer.

Sequencing

Equimolar quantities of three independent amplifications of the input library and equimolar quantities of the purified inclusion band (output library) of each of the nine replicates were pooled and sequenced at the EMBL Genomics Core Facility where Illumina Ampliseq PCR-free libraries were prepared and run on a single lane of an Illumina HiSeq3000. In total, 191.6 million paired-end reads were obtained. The median sequencing coverage for all 3072 genotypes in the input was between 657 and 2069 reads. In the output, the median sequencing coverage was between 389 and 1447 reads depending on the replicate. Raw sequencing data has been submitted to GEO with accession number GSE111316, and to the European Nucleotide Archive with accession number PRJEB24588.

Data processing and calculation of PSI values

The barcodes associated with each experimental replicate (Table S5) were used to demultiplex the raw sequencing files with the SABRE software (https://github.com/najoshi/sabre). Paired-end reads from each replicate were then merged using PEAR (https://cme.h-its.org/exelixis/web/software/pear, Zhang et al., 2014) with the following arguments for the input reads: -m 116 -n 116 -v 116; and the following arguments for the output reads: -m 119 -n 119 -v 115. When necessary, reads were reverse complemented and trimmed using the Seqtk trim tool (https://github.com/lh3/seqtk) with parameters –b 26 and –e 27 for the input reads and –b 26 –e 30 for the output reads. Finally, the FASTQ/A Collapser from the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit) was employed along with a custom python script to count how many times each genotype occurred in a sequencing file. Genotypes containing a mutation not included in the original library design (which allows for a maximum of 3072 different genotypes) was considered a sequencing error and discarded in downstream analyses (this occurred with 14.2% of the reads in the input and 17.2% of the reads in the output). The mean read count among the 3072 allowed genotypes was 1893 in the input and 1355 in the output. For all other genotypes, the mean read counts was 6 in the input and 9 in the output without any one genotype being present more often than expected by a Poisson-distributed random variable. Read counts were then used to calculate enrichment scores and PSI values as described in the QUANTIFICATION AND STATISTICAL ANALYSIS section.

Cell culture and siRNA co-transfection

This section refers to the SF3B1 knock-down experiment (Figure 4G), where the doped library of FAS exon 6 mutants was transfected in the presence of a control siRNA or in the presence of an siRNA against SF3B1. Mycoplasma-free tested HEK293 cells (ATCC, CRL-1573) were sub-cultured in DMEM Glutamax (ThermoFisher Scientific, 61965059) containing 2.5mM glutamine, 15mM HEPES, 10% FBS (ThermoFisher Scientific, 10270) and Pen-Strep (ThermoFisher Scientific, 15070063). 250,000 HEK293 cells were plated in 6-well plates (NUNC, 140675). 10 ng of a FAS exon 6 doped library (Julien et al., 2016) were co-transfected in triplicates with 100nM On-TARGETplus SMARTpool siRNA against SF3B1 L-020061-0005 (Dharmacon) using 2 μL Lipofectamine 2000 (ThermoFisher Scientific, 11668027) per 1ml of total volume of transfection in OPTIMEM (ThermoFisher Scientific, 13778150). Five hours after treatment, the medium was replaced with DMEM-F12 containing 10% FBS and Pen-Strep. 72 hours after transfection, total RNA was extracted using the Maxwell LEV 16 simplyRNA Tissue kit (PROMEGA, AS1280). RNA quality was assessed with a NanoDrop spectrophotometre, and in parallel, protein extracts were prepared using RIPA buffer (1mM EDTA, 1.5mM MgCl2, 20mM TrisHCl pH7.5, 150 mM NaCl, 1% NP40) with 1x Complete protease inhibitor (Roche, 11697498001). As a control, 10 ng of plasmid DNA were co-transfected with 100nM control siRNA ordered from Sigma-Aldrich (sequence: AAGGUCCGGCUCCCCCAAAUG).

Generation of output libraries for the siRNA experiment

This section refers to the SF3B1 knock-down experiment (Figure 4G), where the doped library of FAS exon 6 mutants was transfected in the presence of a control siRNA or in the presence of an siRNA against SF3B1. 72 hours post-transfection, total RNA was isolated using the automated Maxwell LEV 16 simplyRNA tissue kit (AS1280, Promega) as described above for the combinatorially-complete library. cDNA was synthetized from 500 ng total RNA using Superscript III (18080085, Life Technologies), and amplified with one of the three pairs of barcoded FAS_e5_BR_F and FAS_e5_BR_R primers ("Primers used for Ampliseq sequencing – siRNA Experiments" in Table S5) and GoTaq flexi DNA polymerase (M7806, Promega). Each pair of primers had a distinct 8-mer barcode sequence to distinguish the nine experimental replicates ("Primers used for Ampliseq sequencing – siRNA Experiments" in Table S5). PCR products were run on a 2% agarose gel and the band corresponding in size to the amplification product of splicing inclusion was excised, purified using the Quiaquick Gel extraction kit (Quiagen, 50928704) and quantified with a NanoDrop spectrophotometre.

Sequencing the siRNA experiment

This section refers to the SF3B1 knock-down experiment (Figure 4G), where the doped library of FAS exon 6 mutants was transfected in the presence of a control siRNA or in the presence of an siRNA against SF3B1. Equimolar quantities of the purified inclusion band of each of the three control siRNA replicates, and equimolar quantities of the purified inclusion band of each of the three SF3B1 siRNA replicates were pooled and sequenced at the EMBL Genomics Core Facility where Illumina Ampliseq PCR-free libraries were prepared and run on one lane of an Illumina HiSeq2000. In total, 198.4 million paired-end reads were obtained. In the control siRNA experiment, the median sequencing coverage was between 11204 and 17146 and between 20 and 33 reads for single and double mutants, respectively. In the SF3B1 siRNA experiment, the median sequencing coverage was between 14036 and 16600 and between 28 and 41 reads for single and double mutants, respectively.

Western blots

This section refers to the SF3B1 knock-down experiment (Figure 4G), where the doped library of FAS exon 6 mutants was transfected in the presence of a control siRNA or in the presence of an siRNA against SF3B1. Protein extracts were fractionated by electrophoresis in 10% native acrylamide:bisacrylamide (30:0.8%) gels, and semi-dry transferred onto a 0.45 μM nitrocellulose membrane (Protran BA85 10401196, Whatman). The following primary antibodies were used for western blot analysis: rabbit polyclonal anti-SF3B1 (Abcam, ab39578) and anti-GAPDH mouse monoclonal (6CS) (Abcam, ab8245). The following secondary antibodies were then incubated with the membranes: ECL rabbit or mouse IgG, HRP-Linked Whole Ab (GEHealthcare, NA9340 or NA931). After extensive washes the bound antibodies were detected using the Western Lightning Plus ECL chemiluminescence reagent (PerkinElmer, NEL105001EA) and exposed to Kodak BioMax MR film (Sigma-Aldrich, Z353949).

Experimental validation of the sign epistasis interaction between C18G and T19G

To validate the sign epistasis interaction, the PSI of 4 genotypes were quantified: (1) human *FAS* exon 6, (2) human *FAS* exon 6 with the ancestral nucleotide at position 18, (3) human *FAS* exon 6 with the ancestral nucleotide at position 19, and (4) human *FAS* exon 6 with both ancestral nucleotides (''ancestral-like'' genotype). The four genotypes were also tested in the presence of mutation C48T, which decreases inclusion levels and consequently facilitates quantification of genotypes displaying high levels of exon inclusion. Inclusion levels were quantified in HEK293 and COS-7 cells following the protocol described in the ''Experimental evaluation of PSI values'' subsection of the QUANTIFICATION AND STATISTICAL ANALYSIS section.

Assessment of the effects of mutants in HEK293 FlpIn cells

To validate the sign epistasis interaction between C18G and T19G in transcripts derived from reporters integrated in a single genomic site, HEK293 Flp-In T-Rex cells (ThermoFisher Scientific, R780-07) were used following the manufacturer's instructions. Minigenes containing the genotypes described in the previous section were subcloned into the site-specific integration plasmid pcDNA/FRT/TO. 100 ng of pcDNA/PRT/TO_minigenes were co-transfected with 900 ng of the Flp recombinase expression plasmid pOG44 into cells grown in 6-well plates using Lipofectamine 2000 with Opti-MEM I reduced serum medium without phenol red (ThermoFisher Scientific, 11058021). After 6 hours, the medium was changed to DMEM + GlutaMAX (ThermoFisher Scientific, 61965-059) in the presence of 10% fetal bovine serum and pre-strep antibiotics. Stable transfectants were selected with 100 μg/ml hygromycin B (ThermoFisher Scientific, 10687-010) and 5 μg/ml blasticidin (ThermoFisher Scientific A1113903). Polyclonal populations were grown in the presence of 5 μg/ml blasticidin for 10 days. For each genotype, six individual clones were picked with a pipette tip and moved to wells in a 6-well plate. They were grown in the presence of 5 μg/ml blasticidin and 100 μg/ml hygromycin B. Gene expression was then induced using 1μg/ml of doxycycline (CONDA, 631311). Total RNA was extracted (as described below in the ''Experimental evaluation of PSI values'' subsection of the QUANTIFICATION AND STATISTICAL ANALYSIS section) and RT-PCR assays were carried out and the products analyzed on 6% acrylamide gels.

Genetic prediction of combined mutation effects

In order to predict the combined effect of different mutations on the inclusion of the exon, five different linear models were built using the *lm* function in R:

1. A model assuming the 12 mutations present independent linear effects, trained only on the 12 single mutant genotypes (relative to the ancestral sequence) as well as the ancestor (Figure 2D). Therefore, this model takes into account the effect of each of the 12 mutations in only one specific genetic background.
2. A model assuming the 12 mutations present independent linear effects, trained on the high-confidence subset of the dataset (Figure 2E). Therefore, this model takes into account the averaged effect of each single mutation across many different genetic contexts.
3. A model assuming the 12 mutations display independent effects that are nonlinear following the scaling law, trained on the high-confidence subset of the data (Figure 4E).
4. A model allowing for the seven pairwise interactions observed but not allowing for nonlinearities arising from the scaling law, trained on the high-confidence subset of the data (Figure 6A, left panel).
5. A model allowing for the seven pairwise interactions as well as the nonlinear behavior (Figure 6A, right panel).

Models 1 and 2 predicts the PSI of a genotype by using dummy variables representing the presence or absence of each of the 12 mutations in that genotype. Model 3 predicts the mutation effect ln(A) (equation S9, see Data S1) of a genotype using the same variables as model 1 or model 2. For this model, ln(A) was chosen instead of A because mutation effects are additive in logit space (equation S9). For a given genotype, the mutation effect A is calculated setting the starting PSI to 96% (the PSI of the ancestral exon) and the final PSI to that genotype's estimated PSI. Model 4 was built like model 2, but fitting additional dummy variables for the seven epistatic interactions identified. Model 5 was built like model 3, but allowing for the seven pairwise interactions.

When the starting PSI is set to 96%, final PSI values above 100% result in negative values of A (Equation 2). Since ln(A) is not defined for negative numbers, genotypes with an estimated PSI above 100% could not be used to build models 3 and 5. For different models to be adequately compared they should all be built using the same observations (genotypes). Therefore, models 2–5 (and the model using mutation effects on the ancestral background) were built after removing any data points with an estimated PSI above 100%.

Simulated biallelic genotype landscape and "phantom" interactions

This analysis was carried out to illustrate how the presence of global scaling can result in the detection of spurious epistatic interactions (Figure S5E). A biallelic genotype-phenotype landscape with 10 loci (labeled A – J) was built where the allele at each locus could either be 0 (the wild-type state) or 1 (the mutated state). The PSI of the wild-type exon was defined as 50% by setting its k_6 value to 1, fixing k_7 to 1 and τ to 0 (see Equation 1 and exon competition modeling section below).

Mutation at each locus introduced an additive effect x on the ln k_6 parameter of the wild-type exon (Figure 4C middle panel):

$$new\ k_6 = e^{\ln(old\ k_6)+x}$$

Introducing two mutations with effects x and y:

$$new\ k_6 = e^{\ln(old\ k_6)+x+y}$$

The effects of mutation at each locus were:

Mutation	Effect
A	+2.5
B	+2.0
C	+1.5
D	+1.0
E	+0.5
F	−0.5
G	−1.0
H	−1.5
I	−2.0
J	−2.5

Equation 1 was used to convert all these new k_6 values into PSI scores and the PSI distribution of the simulated dataset is shown in Figure S4F.

A Walsh-Hadamard transform was applied to a vector **w** containing all the PSI values in the simulated landscape:

$$e = H \cdot w$$

where **e** is a vector containing the Walsh coefficients, related to the epistatic effect of each genotype in the landscape averaged across all genetic backgrounds (Domingo et al., 2018, Poelwijk et al., 2016). The Walsh-Hadamard transformation matrix **H** can be built recursively with the following formula:

$$H_{n+1} = \begin{pmatrix} H_n & H_n \\ H_n & -H_n \end{pmatrix}$$

with $H_0 = 1$.

One can then use the inverse transformation H^{-1} on **e** to make predictions **w'** about the PSI values in the landscape.

$$w' = H^{-1} \cdot e$$

In this example, **w** = **w'** because **e** was originally built using **w**. To analyze the contribution of the 10^{th} order epistatic terms to our predictions we set their Walsh coefficients inside **e** to 0, and measured the RMSE between the real values (**w**) and the predicted

values (**w'**). The contribution of the 9th order epistatic terms was analyzed by setting their coefficients and those of higher order terms to 0, and so on. The results of this analysis are shown in Figure S5E.

Building a Lasso regression model with higher-order interactions and without global scaling

This section refers to the Lasso model described in the end of the Discussion section titled "Global scaling and sparse pairwise interactions are sufficient for accurate genetic prediction." To build a complex model of our combinatorially-complete dataset (including third-order interactions – higher order combinations not calculated because they greatly increased the computational burden of this analysis) without taking global scaling into account, the *glmnet* function of the *glmnet* package in R was used with its default parameters to perform lasso regression including second- and third-order interactions. The chosen model was the simplest model whose lambda fell within one standard error of the model with the smallest lambda value. This lambda was identified using the *cv.glmnet* function with default parameters. The RMSE value reported for this model corresponds to its aggregated 10-fold cross-validation RMSE score, calculated using the same procedure as above.

PTB-Binding Motif Analysis

To study how mutations in the URE6 region affect the strength of putative PTB binding sites (Figure S6F), motifs M227 and M228 were downloaded from the CISBP-RNA database (Ray et al., 2013). A Position Weight Matrix (PWM) score was calculated in different windows along the exon and plotted in Figure S6F.

Scaling of splicing perturbation effects in WT1 exon 5

This analysis relates to Figures 7A and 7B, where we show how the effect of changing a silencer/enhancer sequence in WT1 exon 5 scales as predicted by the scaling law. Table S2 was downloaded from Ke et al., 2018. This table contains enrichment scores for 141 single and 414 double mutant variants of *WT1* exon 5 in the presence of 10 different exon splicing regulatory sequences (ESR), which may act as exonic splicing enhancers or silencers. To compare the effect on exon inclusion of each ESR at different starting PSIs, we plotted the enrichment scores of mutants in the presence of the WT ESR (starting PSI condition) against the enrichment scores in the presence of all other ESRs (final PSI). To avoid epistatic interactions between a mutation and an ESR, we excluded mutations within 6 nucleotides of the ESR, as well as mutations in positions 16-23, which form a secondary structure in the presence of the WT ESR but not in the presence of the other ESRs.

Genome-wide comparison of exon inclusion levels in two different tissues

This analysis relates to Figure 7C, where we show how complex perturbations in *trans* factors affect exon inclusion levels as predicted by the scaling law. Genome-wide exon inclusion levels were quantified as described in the "Quantifying genome-wide exon inclusion exon inclusion levels in two different tissues" subsection of the QUANTIFICATION AND STATISTICAL ANALYSIS section. To compare exon inclusion levels across two different tissues, we chose brain and skin, as these are the tissues with the largest number of samples (1671 and 1203, respectively).

Genome-wide comparison of exon inclusion levels in two different cell lines

This analysis relates to Figure 7C. Two cell lines representing different lineages were selected among the ENCODE common cell types: HepG2 and Huvec (ENCODE Project Consortium, 2012). RNA-seq data from both cell lines were downloaded from GEO using accession numbers GSM2308416 and GSM2308417 (for HepG2), and GSM2072423 and GSM2072424 (for Huvec). Exon inclusion levels were quantified as described in the "Quantifying genome-wide exon inclusion exon inclusion levels in two different cell lines" Subsection of the QUANTIFICATION AND STATISTICAL ANALYSIS section.

Genome-wide comparison of exon inclusion levels in the presence and absence of mutations in SF3B1

This analysis relates to Figure 7C. We downloaded RNA-seq data from eight myelodysplastic syndrome (MDS) patients with SF3B1 mutations and four MDS patients without mutations in this splicing factor (Dolatshad et al., 2015; GEO series record GSE63569). Next, the data were processed with vast-tools as described in the QUANTIFICATION AND STATISTICAL ANALYSIS for the comparison between two cell lines, with wild-type SF3B1 representing the "starting PSI" condition, and mutated SF3B1 representing the "final PSI" condition.

Genome-wide comparison of exon inclusion levels in two developmental stages

This analysis relates to Figure 7C. Single-cell RNA-seq data from two-cell stage and four-cell-stage human embryos were downloaded (Yan et al., 2013; accessible for download from GEO series record GSE36552) and processed with vast-tools as described in the QUANTIFICATION AND STATISTICAL ANALYSIS for the comparison between two cell lines. Individual cells were not sequenced deep enough for an accurate alternative splicing analysis. Therefore, before running *vast-tools combine*, *vast-tools merge* was used to pull the *vast-tools align* outputs corresponding to different cells from the same embryo together into a new set of output files which were processed as described above, with the 2-cell-stage representing the "starting PSI" condition, and the 4-cell-stage representing the "final PSI" condition.

sQTL analysis

This analysis refers to Figures 7D and 7E where we find evidence of scaling of sQTL effects. The BCFTools (http://www.htslib.org) *view* command with the *–min-af* option set to *0.01:minor* was used to filter the GTEx genotype matrix file (*GTEx_Analysis_2016-01-15_v7_WholeGenomeSeq_635Ind_PASS_AB02_GQ20_HETX_MISS15_PLINKQC.vcf.gz*, provided via dbGap) for common variants within the GTEx cohort (minimum allele frequency $\geq$ 1%). The file was then further filtered using linkage disequilibrium pruning ($R^2 <$ 0.2, a threshold often used to select for distinct loci, see Battle et al., 2017; BCFTools command *+prune* with parameter *–max-LD 0.2*).

The GTEx junction read counts file was processed using the *Psichomics* library in R as described in the ''Quantifying genome-wide exon inclusion exon inclusion levels in two different tissues'' subsection of the QUANTIFICATION AND STATISTICAL ANALYSIS for the comparison of exon PSI levels in different tissues. The output of the *quantifySplicing* function was subset to include only samples for which genotype information was also available. Splicing events that could not be quantified in any of the remaining samples were removed from the dataset.

For each exon splicing event, each potential sQTL and each tissue, a generalized linear model was built to predict exon PSI assuming a binomial distribution (*glm* function in R with the *family* parameter set to *''binomial''*) and using the following variables as predictors: *genotype* (the number of copies of the sQTL candidate), *sex*, *age* and *ischemic time* (the time after death before the sample was processed). A second model was built without the genotype variable. The *lrtest* function from the *lmtest* package in R was then employed to compare both models with a likelihood ratio test and confirm whether introducing the genotype variable in the model significantly improves model predictions.

To focus on cis-sQTLs and to limit the computational burden of this analysis, we only tested variants inside the same gene as the splicing event of interest. Gene boundaries were determined according to the Gencode Release 19 (GRCh37.p13) annotations of the human genome (gencode.v19.annotation.gff3.gz, available for download at ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/gencode.v19.annotation.gff3.gz). A total of 9 375 981 tests were performed and the resulting p values were Bonferroni-corrected. sQTLs without a significant effect (Bonferroni-corrected p value < 0.05) in at least one tissue were removed from the dataset. A list of splicing events with their associated significant sQTLs is shown in Table S6.

The effect size of an sQTL in each tissue was given by the slope of the *genotype* term in a simple linear model where *genotype* was the only predictor for PSI. The intercept term for this model was used as an estimate of the exon PSI in the absence of its sQTL (i.e., the ''starting PSI'' condition). Although one could simply take the average PSI among all samples that lack the sQTL, some sQTLs were never absent (i.e., all samples had either one or two copies of the variant). Therefore, this method allows us to estimate (extrapolate) what the PSI would be in the absence of the sQTL, even if the data are not available. If the estimated PSI in the absence of the sQTL was above 100 (or below 0), this was manually fixed to 100 (or 0).

Alternative splice site usage in a mutant library

This analysis relates to Figures 7F–7H and S7A. The processed read counts file for the alternative 5′splice site mutant library (Rosenberg et al., 2015) was downloaded and extracted from GSE74070_RAW.tar (accessible from GEO series GSE74070). The file was then filtered with a custom bash script to include only those sequences with 100 or more reads. Read counts supporting usage of either the first splice site (SD_1), the second splice (SD_2), the cryptic splice site (SD_C) or intron retention (SD_0) were extracted for each sequence and used to calculate the percent splice site usage (PSU) of the first and second splice sites, as follows:

$$PSU\ of\ first\ splice\ site = 100 \cdot \left(\frac{SD_1}{SD_1 + SD_2 + SD_C + SD_0} \right)$$

$$PSU\ of\ second\ splice\ site = 100 \cdot \left(\frac{SD_2}{SD_1 + SD_2 + SD_C + SD_0} \right)$$

Since the 25-nucleotide mutagenised regions in this library (Figure 7E) are completely random and not based on any wild-type sequence, the sequence corresponding to the median splice site 1 PSU was taken as the wild-type sequence:

5′ TGCTTGGGGAGAAAGGGAAACACATTGCCGGGGGTCGACCCAGGTTCGTGAACGGGATCAAAGCCAACAAGTGCAGAGG TATTCTTATCACCTTCGTGGCT 3′

where the first 25 nucleotides correspond to the first 25-nucleotide-long random region and the last 25 nucleotides correspond to the second randomized region. All mutations were labeled and reported (Figures 7G and S7A) relative to this sequence.

To plot final versus starting PSU as we did for our mutant library, one would ideally require a combinatorially-complete genotype space. Such a space can be represented as a graph where edges connect genotypes (nodes) differing by one nucleotide substitution. Each point in a final versus starting PSI scatterplot would therefore represent a pair of genotypes connected by an edge. However, if our dataset can be represented by a dense graph with many edges, then the dataset from Rosenberg et al., 2015, would be represented by a very sparse graph almost without edges. Therefore, we cannot plot final versus starting PSU using the same method as previously.

Instead, to study the dependence of the effect of mutation A on the starting PSU in which it occurs, we processed the data as follows. The starting PSU was calculated as the median PSU values of all sequences containing every other individual mutation in the absence of mutation A (i.e., the median PSU of all sequences containing mutation B but not mutation A, the median of all sequences containing mutation C but not mutation A, and so on). For the final PSU, we took the median PSU of all sequences containing those other mutations in addition to mutation A (i.e., the median of all sequences containing both B and A, the median of all sequences containing both C and A, and so on).

Genome-wide alternative splice site usage across four pairs of conditions

This analysis relates to Figures 7H and S7B. To calculate genome-wide PSU levels in brain and skin, the Psichomics package in R was used as described in the ''Quantifying genome-wide exon inclusion exon inclusion levels in two different tissues'' subsection of the QUANTIFICATION AND STATISTICAL ANALYSIS section, but the *eventType* argument of the quantifySplicing function was now set to ''A5SS'' (for alternative 5′ splice site) or ''A3SS'' (for alternative 3′ splice site) instead of ''SE.'' For the other three comparisons, vast-tools was used as described in the ''Quantifying genome-wide exon inclusion exon inclusion levels in two different tissues'' subsection of the QUANTIFICATION AND STATISTICAL ANALYSIS section for the analysis of exon inclusion levels. Only splicing events whose IDs begin with HsaALTD (for alternative 5′ splice sites) or HsaALTA (for alternative 3′ splice sites) were considered for downstream analysis.

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification of FAS exon 6 inclusion in different primate species

The *PSI.sh* pipeline described by Schafer et al. (2015) was used with the RNA-Seq aligner STAR v2.5.2a (Dobin et al., 2013) and SAMTools (Li et al., 2009) to measure the inclusion levels of *FAS* exon 6 in each sample. If more than one sample for a specific species and tissue was available, a separate PSI value was estimated for each sample and the mean from all samples was calculated (full data is shown in Table S1). Whenever we had samples for more than one species within one of the clades from the phylogenetic tree shown in Figure 1A, the mean of their PSI values was taken. To measure the inclusion of the orthologous exon in mice (Table S2), tissue-specific RNA-Seq data generated by Merkin et al. (2012) was downloaded and processed using the same pipeline.

Quantification of PSI values in the combinatorially-complete library

An enrichment score (ES) for each genotype in the library was calculated as the ratio between its frequency in the output and the input libraries:

$$ES = \frac{Freqency_{output}}{Frequency_{input}}$$

The PSI of the ancestral exon was experimentally determined to be 96% (Figure S1B). To calculate the PSI of a genotype, the ratio of the genotype ES over the ancestral exon ES was multiplied by 96:

$$PSI = 96 * \frac{ES_{genotype}}{ES_{ancestral\ exon}}$$

Since the PSI estimate of the human wild-type sequence could have been affected by endogenous sequences amplified along with the library, the PSI of this genotype was set to 53.4%, the experimentally-determined value for this genotype that was quantified as described in the ''Experimental evaluation of PSI values'' section.

Experimental evaluation of PSI values

To confirm the accuracy of our PSI estimations, we quantified the inclusion of the 7 reconstructed evolutionary intermediates as well as that of 2 additional genotypes. To obtain these intermediates, ancestral exon oligonucleotides were ordered from IBA Gmbh on the 0.2 μmol scale. These oligonucleotides include the 63-nucleotide-long ancestral exon sequence flanked by 22 nucleotides of the 3′ end of the human intron 5 and 50 nucleotides of the 5′ end of the human intron 6:

5′-TGT CCA ATG TTC CAA CCT ACA G**GA TCC AGA TCT AAC TTG CTG TGG TTG TGT CTC CTG CTT CTC CCG ATT CTA GTA ATT GTT TGG G**GT AAG TTC TTG CTT TGT TCA AAC TGC AGA TTG AAA TAA CTT GGG AAG TAG −3′

Letters in bold indicate the exonic region. As described above for the library, this sample was purified by Reverse Phase HPLC, amplified and recombined with the pCMV FAS wt minigene exon 5-6-7 vector. Mutants were obtained using the Accuprime Pfx DNA polymerase (ThermoFisher Scientific, 12344024), following the manufacturer's instructions, and primers were designed with PrimerX (http://www.bioinformatics.org/primerx/). Individual mutants were verified by Sanger sequencing and transfected into Hek293 cells (or HeLa, or COS-7) in triplicates to quantify the ratio between exon 6 inclusion and skipping. For RT-PCR, minigene-specific primers PT1 and PT2 were used (''Primers used for RT-PCR'' in Table S3). These primers are complementary to a plasmid backbone sequence near the 3′end of the minigene transcripts) in order to avoid amplification of endogenous *FAS* RNAs.

RT-PCR products were fractioned by electrophoresis using 6% polyacrylamide gels in 1 x TBE and Sybr safe staining (ThermoFisher Scientific, S33102). The bands corresponding to exon inclusion or skipping were quantified using ImageJ v1.47 (NIH, USA).

Estimating the maximum mutation effect size of each single mutation

This analysis relates to Figure 4F, where we show how the maximum effect of a mutation is inversely related to the starting PSI where this maximum effect is observed. For each of the 12 single mutations, a Starting PSI – Final PSI dataset was built by matching genotypes without the given mutation (Starting PSI) to the corresponding genotypes containing the mutation (Final PSI). Low-complexity principal curves (Hastie and Stuetzle, 1989) were fitted to the Starting PSI – Final PSI datasets using the *prcurve* function (with the *complexity* argument set to 4 and the *method* argument set to ''pca'') from the *analog* package in R. These curves are a non-parametric fit to the data that try to describe the trends observed. Only data points with estimated PSI values below 100% and whose enrichment scores had a standard deviation below 10 PSI units (low-noise data points) were considered when fitting these curves. The maximum effect size predicted by a principal curve was used as an estimate of the corresponding mutation's maximum effect size (Figure 4F). Mutations C18G, T19G, C32T, T49C and G51C display two distinct behaviors that cannot be described with just one principal curve. In these cases, a principal curve was fitted for each of the two behaviors. The maximum effect sizes estimated using this non-parametric fit were in good agreement with the model behavior (Figure 4F).

Fitting the model to the Starting PSI – Final PSI datasets

Equation 2 can be rearranged such that the effect **A** of a mutation is described as:

$$A = \frac{100 \cdot Final\ PSI - Final\ PSI \cdot Starting\ PSI}{100 \cdot Starting\ PSI - Final\ PSI \cdot Starting\ PSI}$$

The mean effect Â of each of the 12 mutations in our library was calculated taking into account only the low-noise data points from each Starting PSI – Final PSI dataset. The **A** in Equation 2 was substituted by this estimated Â to generate curves that accurately describe the behavior of the 12 different mutations (Figures S3A–S3D).

Data Processing and Calculation of Enrichment Scores in the siRNA Libraries

This section refers to the SF3B1 knock-down experiment (Figure 4G), where the doped library of FAS exon 6 mutants was transfected in the presence of a control siRNA or in the presence of an siRNA against SF3B1. The sequencing data were processed using the pipeline described above for the combinatorially-complete library, with enrichment scores being calculated as the frequency in the output library (the sequenced siRNA experiments) over the frequency in the input library (the sequenced doped library before transfection; data taken from Julien et al., 2016).

Determining the PSI of Genotypes in the siRNA Libraries

This analysis refers to the SF3B1 knock-down experiment (Figure 4G), where the doped library of FAS exon 6 mutants was transfected in the presence of a control siRNA or in the presence of an siRNA against SF3B1. In the presence of control siRNA, the PSI of the WT human genotype was experimentally determined to be 53.4% (Figure S4D). The enrichment score of the human WT genotype was estimated by taking the mode of all the single mutant enrichment scores, which are expected to form a distribution around the WT enrichment score. This was done to achieve an accurate WT enrichment score estimate, which might be affected by endogenous sequences that could have been amplified along with the library. To calculate the PSI of a genotype, the ratio of that genotype's enrichment score (ES) over the estimated human WT ES was multiplied by 53.4:

$$PSI = 53.4 * \frac{ES_{genotype}}{Estimated\ ES_{WT\ human}}$$

In the presence of siRNA against SF3B1, the PSI of the WT human genotype was experimentally determined to be 35.6% (Figure S4D). To calculate the PSI of a genotype, the ratio of that genotype's ES over the estimated human WT ES was multiplied by 35.6:

$$PSI = 35.6 * \frac{ES_{genotype}}{Estimated\ ES_{WT\ human}}$$

To study the relationship between starting and final PSI, only genotypes one or two mutations away from the human WT sequence, with an ES standard deviation below 0.1, and with more than 3000 reads in original input library (Julien et al., 2016) were considered (a total of 448 different exon genotypes). The method of least-squares was used to fit Equation (7) to these points (Figure 4G).

Calculating RMSE values for the genetic prediction models

The RMSE value reported for each model in the ''Genetic prediction of combined mutation effects'' subsection of the METHOD DETAILS corresponds to the aggregated 10-fold cross-validation RMSE score. The observations used to build each model were split into ten folds using the *createFolds* function (with the *k* argument set to 10) from the *caret* package in R. Each fold was treated as a

validation set and the model fit on the remaining nine folds. The error (RMSE) was then calculated on the held-out fold. This resulted in ten different RMSE values that were then aggregated:

$$Aggregated\ RMSE = \sqrt{\frac{RMSE_1^2 + \ldots + RMSE_{10}^2}{10}}$$

These RMSE values were then used to compare how the different models explain our data. This process was repeated without selecting for the low-variance genotypes and the result is shown in the supplementary figures (Figures S2B, S4G, and S7A).

Statistical tests

To determine the genotypes where T-19-G promotes inclusion or skipping, the ΔPSI after adding T-19-G in each background was determined in all 9 biological replicates. For each background, a one-sample Wilcoxon rank sum test was used to compare the observed ΔPSI against 0. P values were corrected for multiple testing using the Benjamini-Hochberg procedure. Backgrounds with a positive ΔPSI and an FDR-corrected p value < 0.05 were considered to be backgrounds where T-19-G promotes inclusion. A negative ΔPSI with an FDR-corrected p value < 0.05 was considered to indicate a background where T-19-G promotes skipping.

To test whether two mutations (X and Y) interact, we took the starting PSI – final PSI dataset corresponding to mutation X and filtered it to include only low-variance genotypes. For each of the remaining starting-final PSI pairs, we estimated the mutation' parameter A using equation S9 from the supplementary text. A two-sample t test was used to compare the mean parameter A in those genotypes also containing mutation Y with the mean parameter A in genotypes without Y. This test was repeated across all 130 possible pairs of mutations such that every potential interaction was tested twice (i.e., a test for Y affecting the behavior of X and another test for X affecting the behavior of Y). The resulting p values were FDR-corrected using the Benjamini-Hochberg procedure. An epistatic interaction was called between a pair of mutations X and Y whenever X significantly affected the parameter A of Y (at an FDR < 0.05) and vice versa, when Y significantly changed the parameter A of X (FDR < 0.05). Although an FDR of 0.05 was chosen as the significance threshold, the same 7 epistatic interactions are identified when using a cut-off of 0.1 or 0.01.

In the sQTL analysis (see below) we tested the significance of genetic variants on the inclusion of an exon using a likelihood-ratio test. A total of 9 375 981 tests were performed and Bonferroni-corrected, as described below in the "sQTL Analysis" section.

Quantifying genome-wide exon inclusion exon inclusion levels in two different tissues

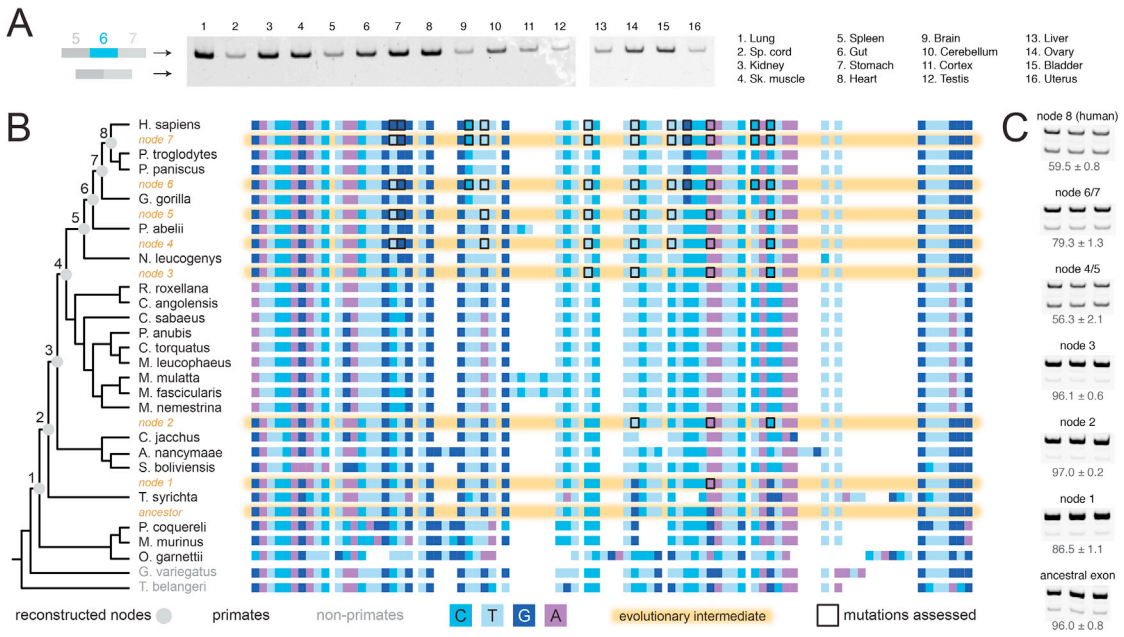
This analysis relates to Figure 7C, where we show how complex perturbations in *trans* factors affect exon inclusion levels as predicted by the scaling law. The *quantifySplicing* function from the *Psichomics* package in R (Saraiva-Agostinho and Barbosa-Morais, 2018) was used with the GTEx junction read counts file (GTEx_Analysis_2016-01-15_v7_STARv2.4.2a_junctions.gct.gz; available for download at https://www.gtexportal.org/home/datasets) to estimate the PSI of all alternative exons in each GTEx sample (Battle et al., 2017), based on the proportion of reads supporting exon inclusion over the reads supporting either inclusion or skipping. All estimates were calculated based on the *Psichomics* hg19/GRCh37 alternative splicing annotations. The *minReads* argument was set to 10 (such that a splicing event requires at least 10 reads for it to be quantified) and the *eventType* argument was set to "SE" (instructing the *quantifySplicing* function to quantify alternative exon events). The output of *quantifySplicing* was set to include only samples originating from either brain or skin tissue. For each exon splicing event, the mean PSI was calculated in brain and in skin. The difference in inclusion levels (ΔPSI) was determined as the mean PSI in skin (the "final PSI") minus the mean PSI in brain (the "starting PSI").

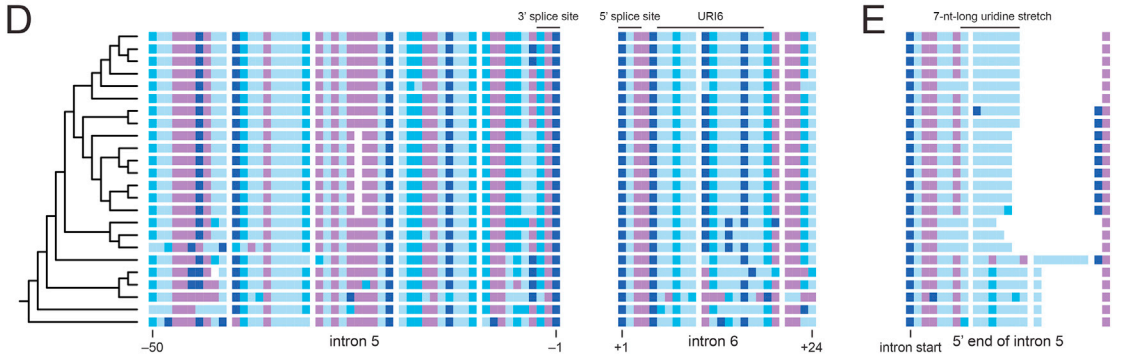
Quantifying genome-wide exon inclusion exon inclusion levels in two different cell lines

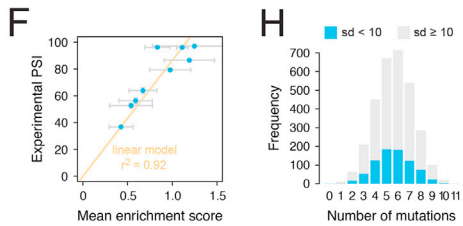
This analysis relates to Figure 7C. Vast-tools *align* and *combine* (Tapial et al., 2017) were used to quantify splicing events across the genome and the output was filtered with a custom bash script to include only exon skipping events. For all exon skipping events (with an ID starting with HsaEX) in each of the two cell lines compared, the mean PSI was taken across all samples with a vast-tools quality score greater than or equal to 'LOW'. If, for a given splicing event, either cell line had no samples above the quality threshold, the splicing event was removed from the dataset and not considered in the analysis. The difference in inclusion levels (ΔPSI) was determined as the mean PSI in HUVEC (the "final PSI") minus the mean PSI in HepG2 (the "starting PSI").

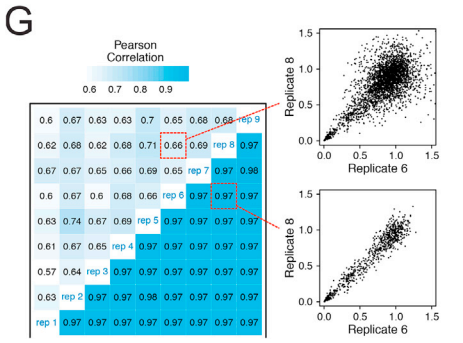
DATA AND SOFTWARE AVAILABILITY

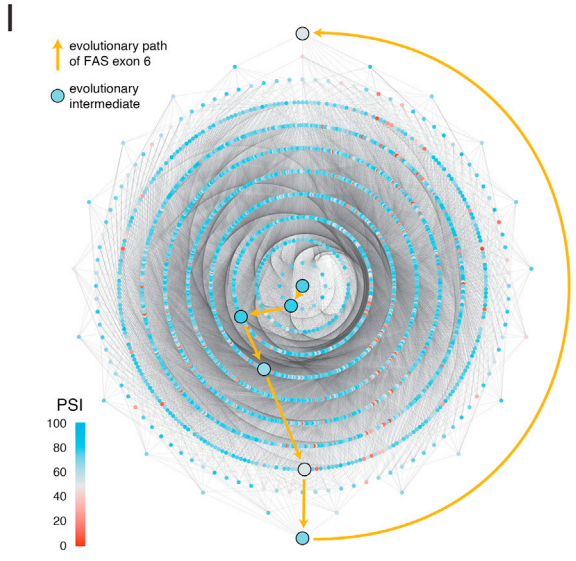
Raw sequencing data for (1) the phylogeny-based mutant library and (2) the doped library in the presence of siRNA against SF3B1 or control siRNA have been submitted to GEO with accession number GSE111316 and to the European Nucleotide Archive with accession number PRJEB24588. All scripts used in this study are available at https://github.com/lehner-lab/Scaling_Law











(legend on next page)

Figure S1. Analyzing Mutation Effects in a Combinatorially Complete Subset of a Genotype Space, Related to Figure 1

A. Pattern of mouse *FAS* exon 6 inclusion in different mouse tissues. The exon was constitutively included in all tissues examined.

B. Multiple sequence alignment of *FAS* exon 6 from different primates and the reconstructed evolutionary intermediates (highlighted in yellow). Mutations acquired throughout the course of evolution are shown surrounded by black squares.

C. Inclusion levels of human *FAS* exon 6 and the reconstructed evolutionary intermediates expressed from minigene constructs in HEK293 cells. Numbers indicate PSI values and standard deviations as quantified using ImageJ (see STAR Methods).

D. Multiple sequence alignment of intronic sequences flanking *FAS* exon 6, including the 3′ and 5′ splice sites. The phylogenetic tree shown to the left of the multiple sequence alignment is the same as in A.

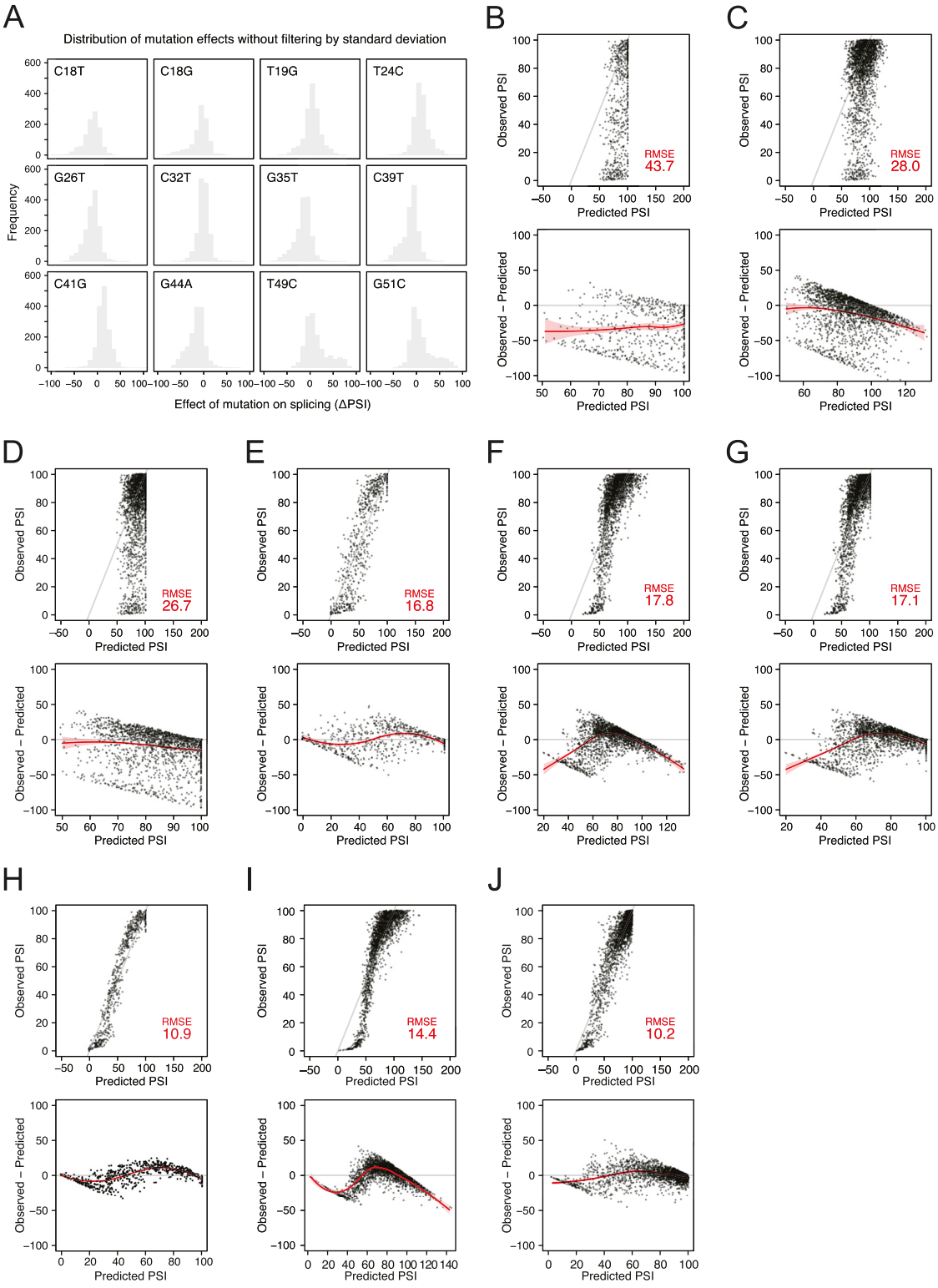
E. Multiple sequence alignment of the 5′ end of intron 5 showing how the 7-nucleotide stretch of uridines at the 5′ end of *FAS* intron 5 (see Figure S6D) is shorter in sequences from primates more distantly related to humans. Sequences are ordered by species as in A.

F. Linear relationship between PSI values from individual transfections and library ESs allows to build a linear model to predict PSI values from ESs.

G. Correlation between ESs across the nine replicates. Upper half of the heatmap shows correlation scores for all 3072 genotypes in the library. Lower half shows correlation scores for low-variance (standard deviation < 10 PSI units) subset of the library.

H. Bar plot showing the number of genotypes for each Hamming distance away from the ancestral sequence. The fraction in blue represents those genotypes with a standard deviation below 10 PSI units.

I. Alternative visualization of a combinatorially-complete subset of the genotype-phenotype landscape of FAS exon 6. Each node represents a genotype with genotypes connected by an edge if they differ by 1 nt. The larger circles connected by yellow arrows represent the evolutionary intermediates. Nodes are colored by PSI scores.



(legend on next page)

Figure S2. Mutations Have Non-independent Effects on Splicing (including genotypes with a standard deviation above 10 PSI units), Related to Figures 2 and 6

A. Distributions of mutation effects of all genotypes in the library.

B. Top: real versus predicted PSI for a model that uses the effect of individual mutations on the ancestral sequence to predict their effect in other contexts, and which introduces the restriction that predicted PSI values cannot be greater than 100 or smaller than 0. Predictions made on the low-variance subset of the data. Bottom: Residuals plot with loess trend line and 95% confidence band.

C. Top: real versus predicted PSI for a model that uses the effect of individual mutations on the ancestral sequence to predict their effect in other contexts. Predictions also made on genotypes with high variance. Bottom: Residuals plot with loess trend line and 95% confidence band.

D. Top: real versus predicted PSI for a model that uses the effect of individual mutations on the ancestral sequence to predict their effect in other contexts, and which introduces the restriction that predicted PSI values cannot be greater than 100 or smaller than 0. Predictions also made on genotypes with high variance. Bottom: Residuals plot with loess trend line and 95% confidence band.

E. Top: predictive model that uses the average effect of mutations in different genotypes and which also introduces the restriction that predicted PSI values cannot be above 100 or below 0. Model built on the low-variance subset of the data. Bottom: Residuals plot with loess trend line and 95% confidence band.

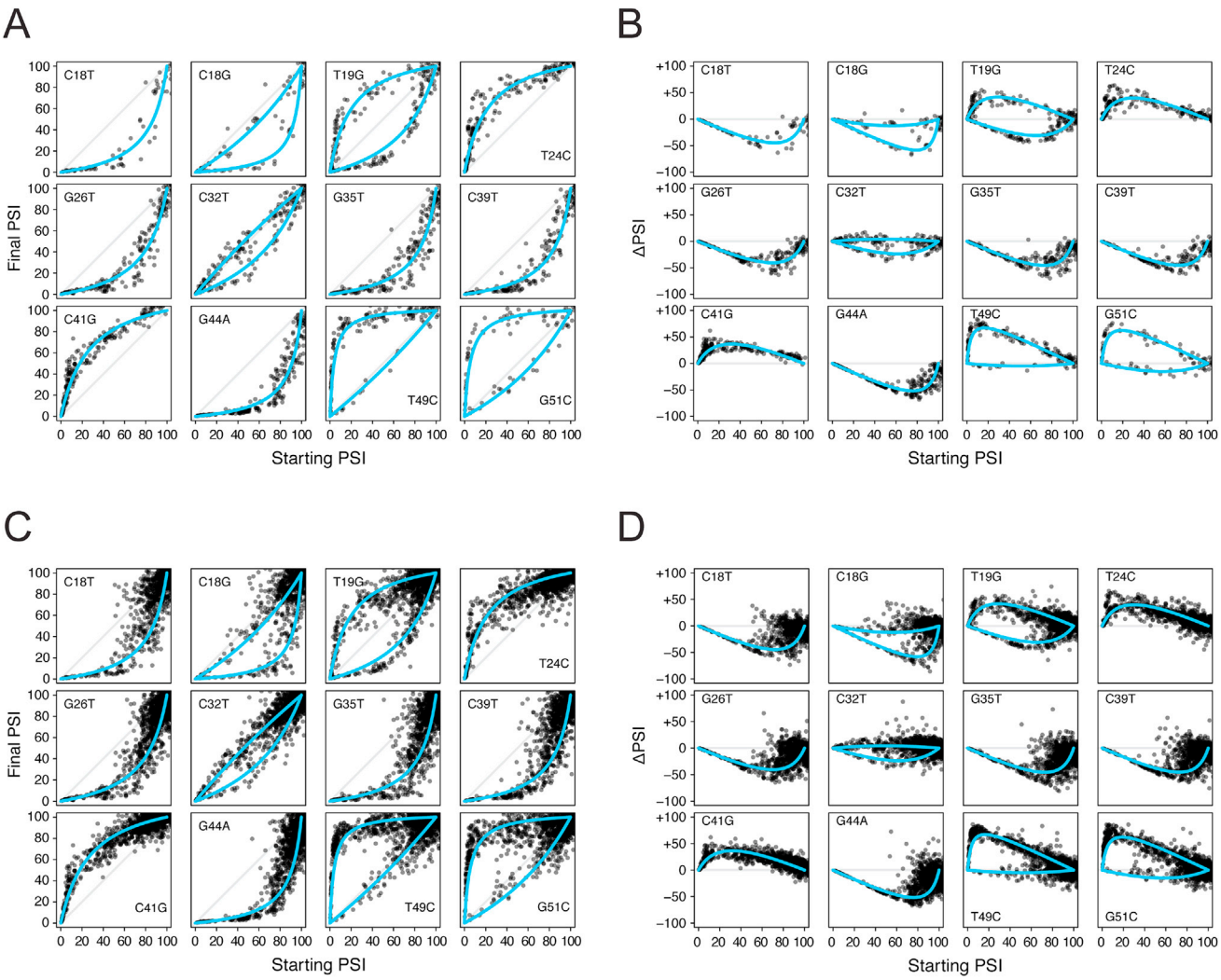
F. Top: real versus predicted PSI for a model that considers the average effect of mutations in different contexts. Model built also on genotypes with high variance. Bottom: Residuals plot with loess trend line and 95% confidence band.

G. Top: real versus predicted PSI for a model that considers the average effect of mutations in different contexts, and which introduces the restriction that predicted PSI values cannot be above 100 or below 0. Model built also on genotypes with high variance. Bottom: Residuals plot with loess trend line and 95% confidence band.

H. Top: real versus predicted PSI for a model that considers the average effect of mutations in different contexts, includes seven pairwise epistatic terms and introduces the restriction that predicted PSI values cannot be above 100 or below 0. Model built on the low-variance subset of the data. Bottom: Residuals plot with loess trend line and 95% confidence band.

I. Top: real versus predicted PSI for a model that considers the average effect of mutations in different contexts and includes seven pairwise epistatic terms. Model built also on genotypes with high variance. Bottom: Residuals plot with loess trend line and 95% confidence band.

J. Top: real versus predicted PSI for a model that considers the average effect of mutations in different contexts, includes seven pairwise epistatic terms and introduces the restriction that predicted PSI values cannot be above 100 or below 0. Model built also on genotypes with high variance. Bottom: Residuals plot with loess trend line and 95% confidence band.



E

cross-validation RMSE values

	C18T	C18G top	C18G bottom	C19G top	C19G bottom	T24C	G26T	C32T top	C32T bottom	G35T	C39T	C41G	G44A	T49C top	T49C bottom	G51C top	G51C bottom
sd < 10	15.4	6.4	34.2	9.4	10.5	10.0	15.2	4.5	9.7	28.8	19.3	8.1	24.6	7.2	6.0	7.0	8.6
any sd	20.2	11.0	26.2	10.6	14.0	11.0	19.2	9.9	12.9	22.6	20.8	9.0	23.0	8.4	12.9	11.0	12.0

Figure S3. The Global Scaling Law Describes the Effects of All Mutations in Our Dataset, Related to Figure 3D

A. Final PSI versus starting PSI corresponding to the indicated mutations for high-confidence datapoints. Blue lines show the fit of our mathematical model to the data. Two different curves were fitted to the data whenever a mutation displayed two distinct behaviors.

B. ΔPSI versus starting PSI corresponding to the indicated mutations for high-confidence datapoints. Blue lines calculated as in A.

C. Final PSI versus starting PSI corresponding to the indicated mutations for all datapoints. Blue lines are the same as those shown in A.

D. ΔPSI versus starting PSI corresponding to the indicated mutations for all datapoints. Blue lines are the same as those shown in B.

E. Cross-validation RMSE values for the fitted curves shown in A–D. Some mutations display two distinct behaviors (two different curves) in A-D. In these cases, 2 different RMSE values were calculated and labeled "Top" or "Bottom" depending on which curve they refer to.

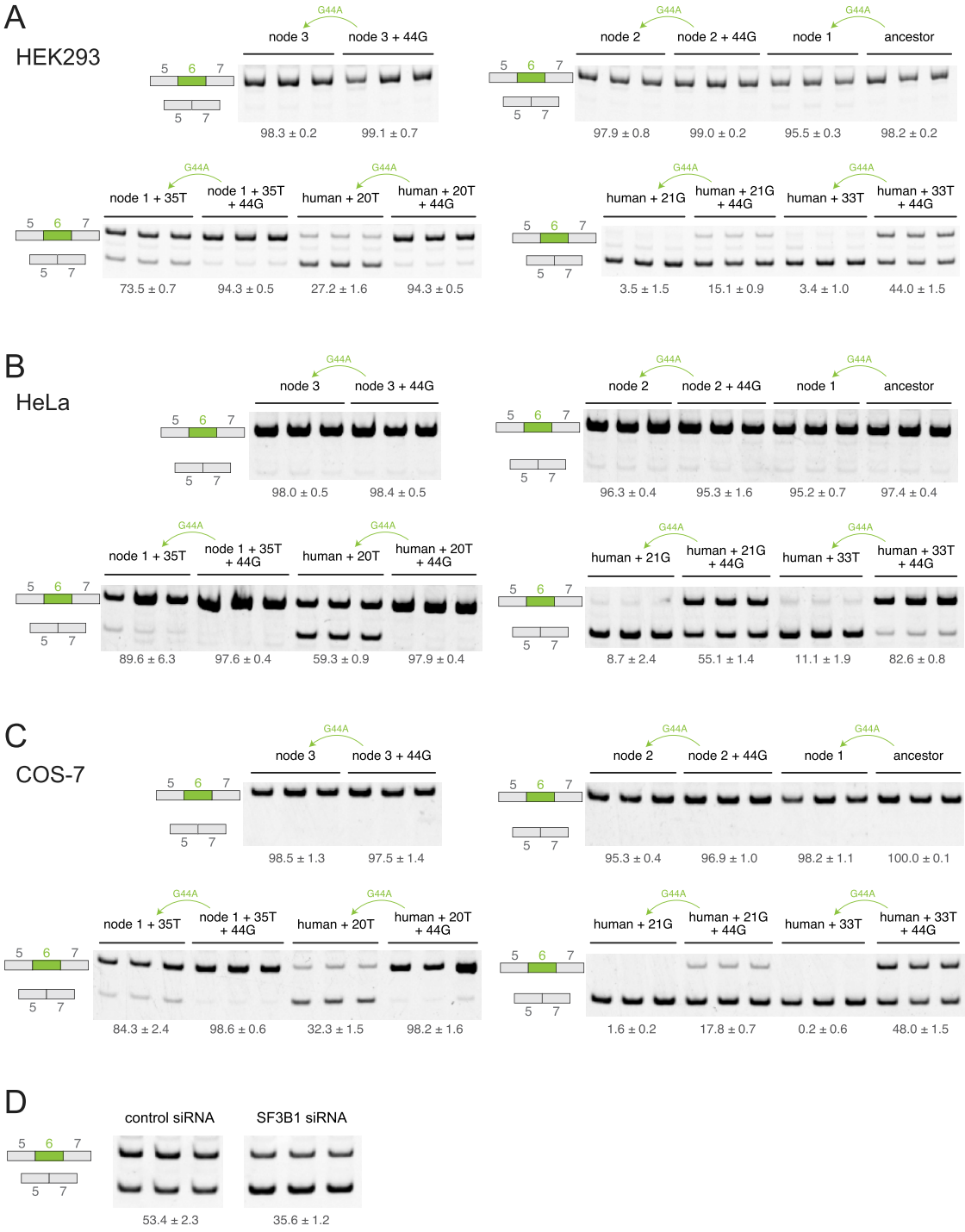


Figure S4. RT-PCR Assays, Related to Figure 3E

A. RT-PCR analysis of *FAS* exon 6 inclusion or skipping for 14 individual *FAS* exon 6 genotypes expressed in HEK293 cells. RNA from cells transfected with the corresponding constructs was purified and amplified by RT-PCR using primers corresponding to vector sequences. Products of amplification were separated by electrophoresis, stained and PSI values estimated using the ImageJ software.

B. RT-PCR analysis of *FAS* exon 6 inclusion or skipping for 14 individual *FAS* exon 6 genotypes expressed in HeLa cells. Analyses were as in A.

C. RT-PCR analysis of *FAS* exon 6 inclusion or skipping for 14 individual *FAS* exon 6 genotypes expressed in COS-7 cells. Analyses were as in A.

D. RT-PCR analysis of human *FAS* exon 6 inclusion or skipping under conditions of siRNA-mediated knock down of the splicing factor SF3B1, or control siRNA. Analyses were as in A.

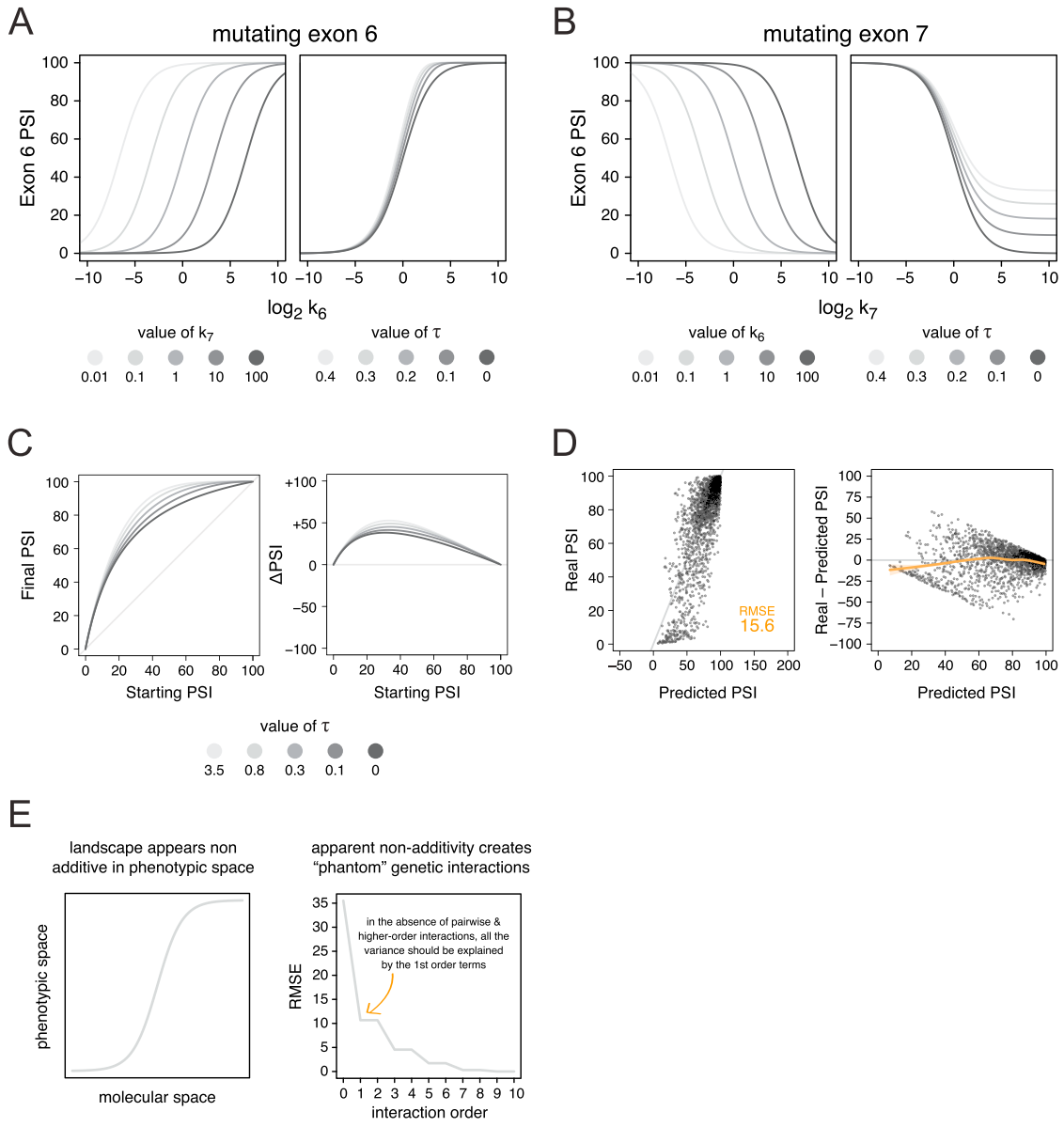


Figure S5. Analysis of the Mathematical Model, Related to Figure 4

A. Relationship between exon 6 PSI and fold-change in k_6 when fixing k_7 (left, $\tau = 0$) and τ (right, $k_7 = 1$) to different values.

B. Relationship between exon 6 PSI and fold-changes in k_7 when fixing k_6 (left, $\tau = 0$) and τ (right, $k_7 = 1$) to different values.

C. Effect of τ on the non-monotonic scaling of mutation effects when the mutation has a moderate inclusion-promoting effect.

D. Left: model considering the global scaling of mutation effects built on genotypes without filtering by their PSI standard deviation. Right: residuals plot with loess trend line and 95% confidence band.

E. Because of the global scaling law, a simulated biallelic landscape with 10 loci where all mutations behave additively in the underlying molecular space appears non-additive in phenotypic (PSI) space. Since the landscape is purely additive, single mutant effects should explain all the variance in the data. However, if the nonlinearlity in the landscape introduced by the global scaling law is not taken into account decomposing the effect of mutations on phenotype using the Walsh-Hadamard transformation (Domingo et al., 2018; Poelwijk et al., 2016; see STAR Methods) reveals "phantom" specific pairwise and higher order interactions.

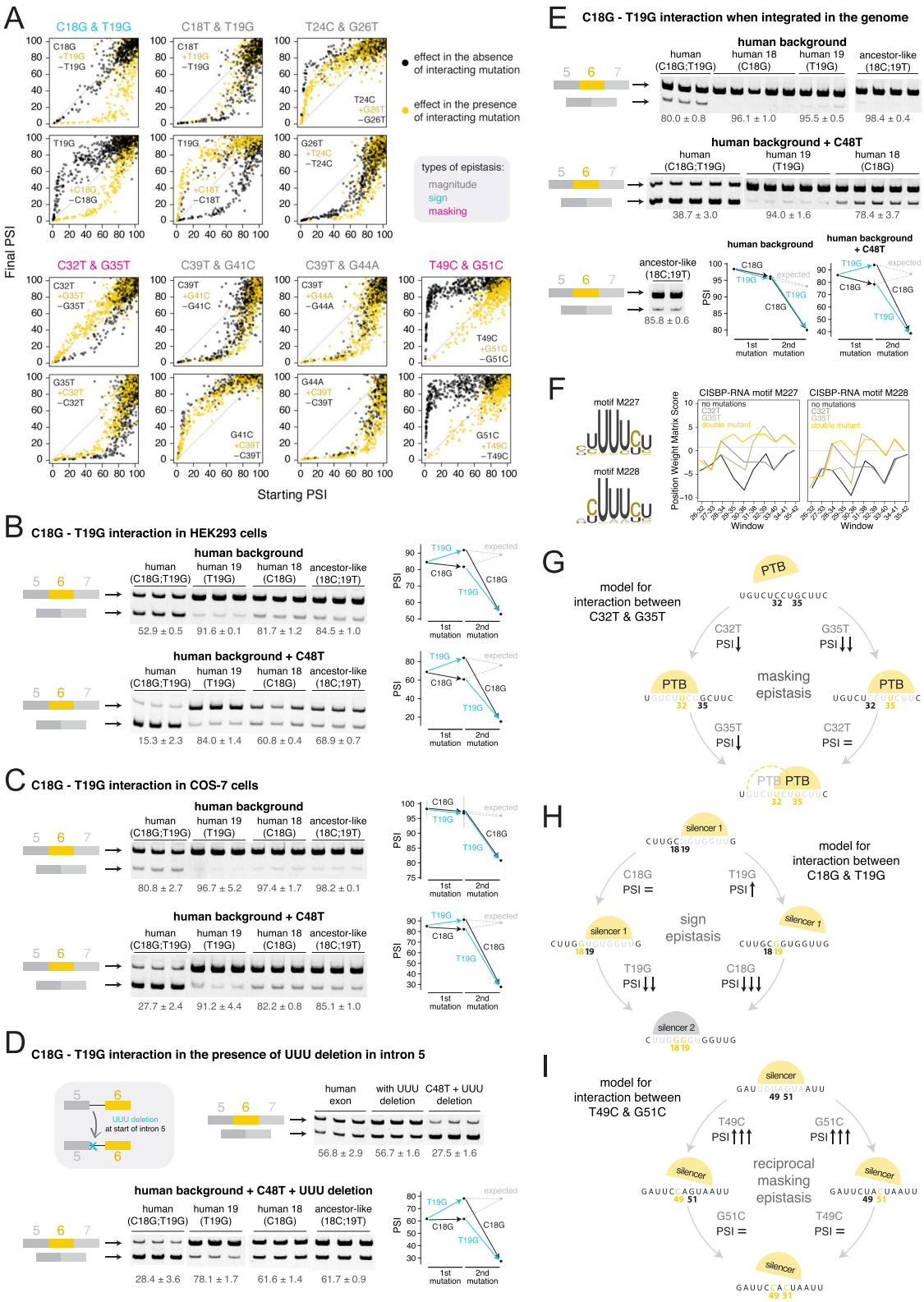


Figure S6. Pairwise Interactions, Related to Figure 5

A. Behavior of the indicated mutations (final versus starting PSI plots) as a function of the starting PSI in the presence (yellow) or absence (black) of its interaction partner, for all genotypes, without filtering by standard deviation). Relevant to Figure 5D.

(legend continued on next page)

B. Validating the interaction between C18G and T19G in HEK293 cells. This interaction was tested in the context of a human exon (top) as well as in the context of mutation C48T (bottom, which decreases the inclusion levels of *FAS* exon 6 and should therefore allow for a better assessment of mutation effects, bringing the ancestral-like exon to more intermediate levels of exon inclusion). T19G promotes inclusion when position 18 contains a C, but promotes skipping when position 18 contains a G. Analysis carried out as in Figure S4. Plots on the right represent PSI values in the ancestral-like exon (human sequence plus a C in position 18 and a T in position 19) or upon each mutation and gray error bars show the standard deviation.

C. Validating the interaction between C18G and T19G in COS-7 cells (fibroblast-like cell lines derived from African green monkey kidney tissue), in the context of the human exon (top gel) and in the context of mutation C48T (bottom gel). Without mutation C48T, the ancestral-like exon is included with a PSI of 98.2% and the effect of individual mutations could not be assessed accurately. However, the interaction becomes clear in the presence of C48T, when the ancestral-like exon is included at lower levels. Analysis was carried out as in B.

D. To explore the evolutionary relevance of the sign epistasis interaction, we tested whether it is independent of co-evolving intronic changes. The 5′ end of *FAS* intron 5 contains a uridine-rich region that can bind to TIA-1 and promote *FAS* exon 6 inclusion (Förch et al., 2000). Phylogenetic analysis suggests that three uridines were progressively gained in this region throughout the evolution of primates (Figure S1E). Top gel: Deleting 3Us does not affect *FAS* exon 6 inclusion. Bottom gel: Validating the interaction between C18G and T19G in HEK293 cells, in the presence of a 3-nucleotide deletion within the 7-nucleotide-long stretch of uridines found in the 3′ end of intron 5. The interaction was tested in the presence of mutation C48T as described in B.

E. Validating the interaction between C18G and T19G in HEK293-FlpIn cells, where the minigene is integrated in the genome at a single specific recombination site.

F. Motif analyses of the likelihood of PTB binding to different windows within the URE6 region of *FAS* exon 6.

G. The interaction between C32T and G35T could involve each mutation creating overlapping - and so mutually exclusive - binding sites for PTB. G35T creates a stronger PTB binding site and therefore, in the presence of G35T, C32T has no additional effect.

H. The interaction between C18G and T19G could be caused by T19G alone breaking the binding site of a splicing repressor protein, but creating a new repressor binding site in the presence of C18G.

I. The interaction between T49C and G51C could be due to either mutation completely breaking a splicing silencer. These mutations have no additional effect in the presence of the other mutation because the silencer is already lost.

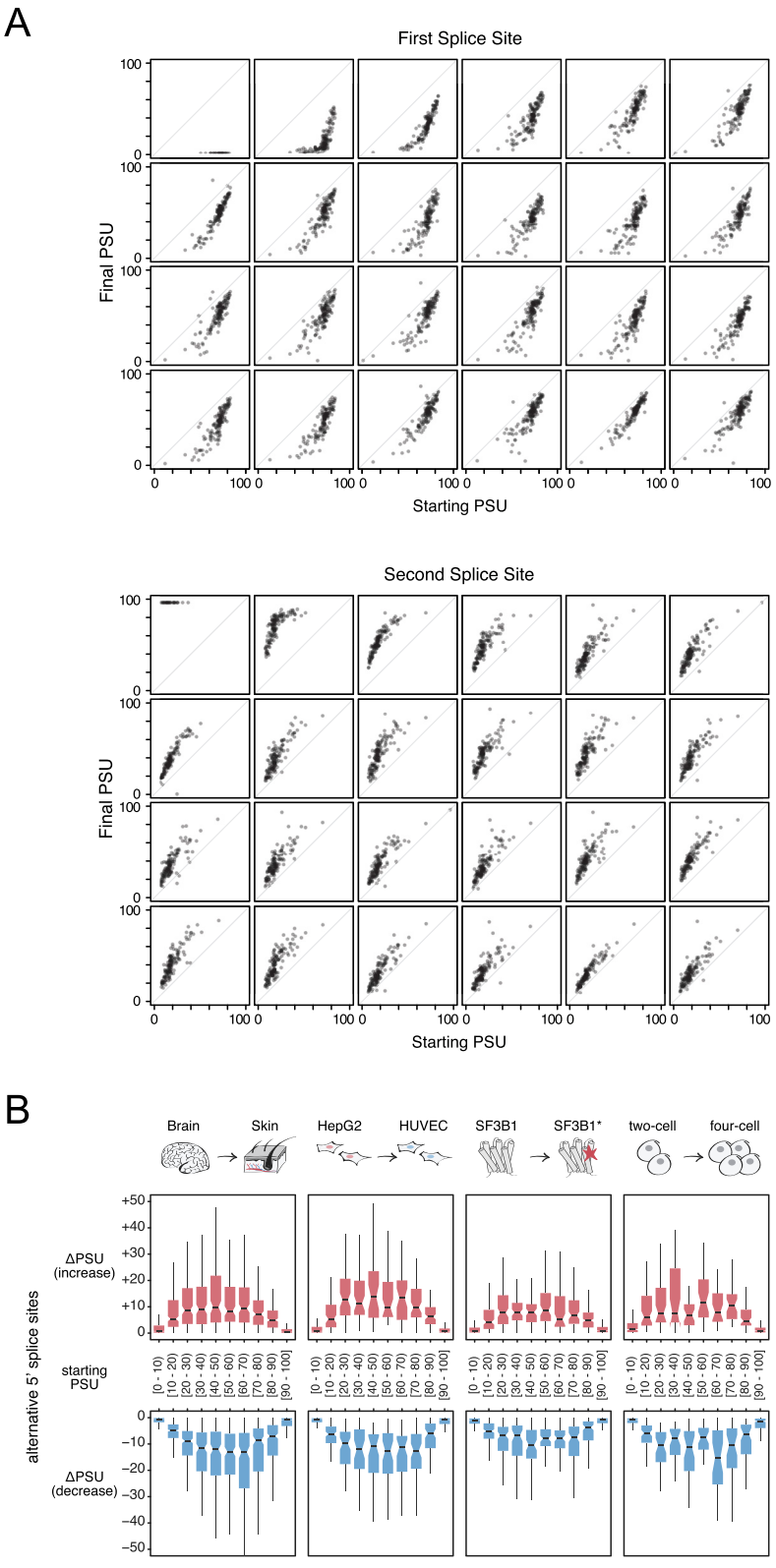


Figure S7. Global Scaling in Alternative Splice Site Usage, Related to Figure 7

A. Effect of 24 mutations from the alternative 5′ splice site library in Rosenberg et al., 2015, on splice site usage.

B. Boxplots showing how genome-wide alternative 5′ splice site usage levels compare across the four pairs of conditions from Figure 7A.