Supplementary Methods for Majoros et al., "Predicting Gene Structure Changes Resulting from Genetic Variants via Exon Definition Features."

S1. Gene structure prediction in real genes modified with premature stop codons

To systematically investigate the behavior of traditional gene structure predictors on genes with possible loss-of-function mutations in personal genomes, we ran a leading HMM-based gene finder (Stanke et al., 2006) on 19,000 human genes modified to include a premature stop codon. First, the HMM was applied to the genomic sequence of each gene (including 1000 nucleotides of flanking sequence on either side) without modification, and the gene structure predicted by the HMM was noted; no comparison to the annotated gene structure was made. Then a stop codon was inserted at a randomly-selected location in the predicted coding segment, while ensuring that the inserted stop codon did not inadvertently create a canonical donor or acceptor splice site consensus. Then the HMM was run again to obtain a second prediction, this time using the sequence with the premature stop codon.

The splice pattern predicted for the modified sequence was compared the the splice pattern predicted for the unmodified sequence. As traditional gene finders key largely on codon biases in contiguous reading frames, we hypothesized that the gene finder would frequently modify its predicted splice pattern so as to omit the premature stop codon. We tabulated the number of times this occurred.

S2. Regularized logistic regression training

For logistic signal and content sensors we applied regularized (elastic net) logistic regression. We used *glmnet* version 2.0-2, with $\alpha = 0.5$ to interpolate equally between L_1 and L_2 regularization, with regularization strength λ selected to minimize the mean cross-validation error (Friedman et al., 2010). This value of α was chosen after observing in preliminary runs on individual HG00096 that $\alpha = 0.5$ produced higher prediction accuracy as validated via RNA-seq (AUC = 0.79) than pure *lasso regression* ($\alpha = 0$, AUC = 0.76) and pure *ridge regression* ($\alpha = 1.0$, AUC = 0.77). For content sensors, all 4096 hexamers were used as features. Hexamer counts were extracted from all reading frames on the sense strand of training exons and introns, and regularized logistic regression was applied to learn a weight for each hexamer for classifying exons versus introns. These weights were taken together as the model for Φ_{exon}. To obtain Φ_{intron}, the weights of the exon model were negated.

For signal sensors, a fixed window around each splice site was used to define indicator variables at each position within the window. For donor splice sites, 6 bp of sequence was included to the left (5') of the 2 bp consensus, as well as 12 bp of sequence to the right (3') of the consensus; only sequences GT, GC, and AT were accepted as valid consensuses. For

acceptor splice sites, 20 bp of sequence left (5') of the consensus and 2 bp right (3') of the consensus were included in the window; consensuses AG and AC were considered valid. At each window position we set four indicator variables, one for each of the four possible nucleotides, to indicate which nucleotide was present at that position in the current training or test case. For example, if nucleotide C was present at that position in the training case, the indicator for C was set to 1 and the indicators for A, G, and T were set to 0. A pseudocount of 0.1 was added to all counts. Regularized logistic regression (α = 0.5) was applied to learn a vector of weights for the indicators within the window. Separate models were learned for donor splice sites and acceptor splice sites.

S3. Evaluation of SGRF prediction accuracy using Geuvadis data

RNA-seq data from LCLs for 150 individuals in the Geuvadis project was used to validate the SGRF predictions. For each individual and each gene in GENCODE v19, phased variants were used to construct explicit haplotype sequences for the gene, as previously described (Majoros et al., 2017). Insertion/deletion variants were used to infer an alignment between the reference sequence and the personal genome. For each annotated isoform of each gene, the isoform was projected onto the personal genomic sequence using the inferred alignment. The SGRF was then applied to produce predicted splice forms in the personal genome.

Predictions were made separately using the human logistic model, the *Arabidopsis* logistic model, and the minigene weights from Rosenberg et al. (2015). RNA-seq data was then aligned to the personal genome by TopHat2 (Kim et al., 2013), with the projected annotations and pooled predictions provided to TopHat2 as annotations. Alignment to the personal genome rather than to the reference was performed in order to reduce *reference bias* in read-mapping (Degner et al., 2009). For each prediction, novel splice junctions that did not occur in any annotated isoform of the gene were considered to be consistent with RNA-seq if at least one spliced read exactly matched both coordinates of the splice junction. A threshold of one read was selected due to the low coverage of the Geuvadis data, and was justified by previous results in which higher thresholds produced similar accuracy estimates (Majoros et al., 2017).

Transcripts not expressed in LCLs at an FPKM of at least 3 were omitted from the analysis, as these may unfairly penalize predictors. Splicing changes predicted to induce nonsense-mediated decay (NMD) were also omitted from the evaluation, as NMD reduces transcript levels, which may in turn unfairly penalize predictors. True positive rate and false positive rate were calculated by tabulating predictions at the individual splice-junction level and were then used to construct ROC curves. Because a very large number of decoy splice sites will exist in a typical sequence, resulting in an excessively large number of possible splice junctions, evaluating classification accuracy on the set of all possible junctions would vastly inflate the number of true negatives and thereby inflate estimates of classification accuracy. The set of junctions (real and decoy) used for evaluation was therefore limited to the union of all junctions predicted by any predictor (including junctions assigned a probability of zero). All predictors were evaluated on this same union set of junctions. Reported ROC curves reflect only the accuracy of novel junctions not occurring in any annotated isoform of the gene.

S4. Classification of whole exons and introns

To assess whether learned hexamer weights reflected coding features, noncoding features, or both, we used the logistic regression model alone (without the SGRF) to classify an equal number of exons versus introns shortened to the same length; splice sites and 10 bp flanking them were removed from all introns prior to shortening them to match exon lengths. Testing was performed only on exons and introns not included in the training set. The standard logistic function was used to compute $P(exon|sequence)$ and classification was performed under the rule that $P > 0.5$ dictates an exon prediction and $P < 0.5$ dictates an intron prediction.

For evaluation of prediction of minigene splicing outcomes, spliced read counts were obtained from Rosenberg et al. (2015) for the randomized sequences between alternate splice donor sites on the minigene. Sequences with appreciable numbers of read counts splicing at locations other than the two main splice sites, SD1 and SD2, were discarded. Sequences for which $count(SD1) > 2count(SD2)$ were counted as negative cases of exon definition (i.e., the exon most often failed to extend to the further splice site, SD2), and sequences for which $count(SD2) > 2count(SD1)$ were counted as positive cases of exon definition (i.e., the exon most often succeded at extending to the further splice site, SD2). The logistic model trained on human annotations was tested by using the standard logistic function as above to classify the randomized sequences as positives or negatives; these predictions were evaluated against the known classifications based on spliced read counts as described above. True positive rate and false positive rate were calculated accordingly to produce an ROC curve.

S5. Simulating creation and destruction of splice sites

In order to investigate the propensity of a simple mutation process to create or destroy splice sites, a previously-estimated context-dependent DNA substitution matrix (Allen et al., 2013) was used to simulate mutations jointly conditional on the nucleotide being mutated and the two immediately flanking nucleotides (one nt 5' of the mutated site and one nt 3' of the mutated site). Simulations were performed for 1,991 genes on human chromosome 1, with 1,000 mutations being applied per gene; each mutation was reversed prior to sampling the next mutation, so that each mutated sequence had an edit distance of exactly 1 from the original genomic sequence.

Each mutation was assessed as to its ability to create or destroy a splice site. For annotated splice sites, if a mutation changed a canonical donor (GT, GC, or AT) or acceptor (AG or AC) splice consensus to a non-consensus, the site was counted as disrupted. In addition, if a mutation did not change the consensus but did modify a flanking position such that the splice-site score dropped below the threshold of the logistic sensor (chosen to admit 99% of training splice sites), it was also counted as disruption of the splice site.

For mutations that did not disrupt a splice site, three criteria levels were applied to determine whether the mutation could create a *de novo* splice site. For the least stringent level, if the mutation changed a non-consensus 2 bp sequence to any consensus sequence for either donor or acceptor splice sites, the mutation was counted as creating a *de novo* splice site. For the second stringency level, if a non-consensus 2 bp sequence was changed to a

consensus sequence, the corresponding logistic signal sensor was applied to a window (Supplementary Methods section S2) containing the consensus and the logistic sensor was used to classify the sequence as a splice site or a non-splice site. Only sites in which both a non-consensus 2 bp sequence was transformed by the mutation into a consensus sequence and the logistic sensor classified the window as a splice site were counted as creation events. For the most stringent level, a mutation was required to create a 2 bp splice site consensus, be classified by the signal sensor as a splice site, and occur in a favorable exon definition context to be counted as a *de novo* splice site. Exon definition potential was measured using the logistic content sensors. For *de novo* splice sites occurring in annotated introns, the intervening sequence between the *de novo* splice site and the exon it would extend was evaluated using the exon content sensor. For *de novo* splice sites occurring in annotated exons, the propensity to shorten the exon was measured by applying the intronic content sensor to the sequence that would become intronic. Thresholds for signal and content sensors were selected so as to admit 99% of training sequences. Three simulations were performed using the same random number seed, one simulation per stringency level.

S6. Integration of SGRF into ACE+

Our previously-described software, ACE (Majoros et al., 2017), interprets proposed splicing changes in terms of their effects on encoded proteins, enabling prediction of loss of function via nonsense-mediated decay (NMD), nonstop decay, loss of coding potential, and protein modification or truncation. Incorporating the SGRF into the ACE framework enables putative splicing changes predicted by the SGRF to be evaluated in terms of their likely effect on gene function, and provides a means of ranking ACE predictions based on their probability under the SGRF. We named the resulting software ACE+ (Supplementary Fig. S7). ACE+ constructs personal genomes from a phased VCF file containing SNPs, multinucleotide variants, short insertions or deletions (indels), and short copy-number variants (CNVs). It then projects reference annotations onto the alternate sequence using a coordinate transformation and applies the SGRF to predict splicing changes. Finally, ACE+ interprets the resulting changes in terms of the potential for loss of function of the final gene product. Variants within each gene are then re-interpreted as to their likely effect, based on changes in the annotation predicted by the SGRF. For example, coding variants may become noncoding variants and vice-versa. Outputs are provided in a highly-structured format called Essex (Majoros et al., 2017). An API for parsing the structured output and extracting arbitrary features is provided in perl, python, and C++.