GLOBAL ANALYSIS OF HUMAN MRNA FOLDING DISRUPTIONS IN SYNONYMOUS VARIANTS DEMONSTRATES SIGNIFICANT POPULATION CONSTRAINT

Jeffrey B.S. Gaither, Grant E. Lammi, James L. Li, David M. Gordon, Harkness C. Kuck, Benjamin J. Kelly, James R. Fitch and Peter White[#*]

Computational Genomics Group, The Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, Ohio, USA

[#] Department of Pediatrics, College of Medicine, The Ohio State University, Columbus, Ohio, USA

* Corresponding author

Mailing address:

The Institute for Genomic Medicine

Nationwide Children's Hospital

575 Children's Crossroad

Columbus, OH 43215. USA

Phone: (614) 355-2671; Fax: (614) 355-6833

E-mail: peter.white@nationwidechildrens.org

Keywords: synonymous mutation, mRNA stability, RNA folding, RNA secondary structure, genetic disease, Spark, gnomAD, transcriptomics

TABLE OF CONTENTS

ABSTRACT

Background. In most organisms the structure of an mRNA molecule is a crucial determinant of its speed of translation, half-life, splicing propensities and final configuration as a protein. Synonymous mutations which distort this wildtype mRNA structure may be pathogenic as a consequence. However, current clinical guidelines classify synonymous or "silent" single nucleotide variants (sSNVs) as largely benign unless a role in RNA splicing can be demonstrated.

Results. We developed novel software to conduct a global transcriptome study in which RNA folding statistics were computed for 469 million SNVs in 45,800 transcripts using an Apache Spark implementation of the ViennaRNA software package in the cloud. Focusing our analysis on the subset of 17.9 million sSNVs we discover that variants predicted to disrupt mRNA structure have lower rates of incidence in the human population. Given that the community lacks tools to evaluate the potential pathogenic impact of sSNVs, we introduce a "Structural Predictivity Index" (SPI) to quantify this constraint due to mRNA structure.

Conclusion. Our findings support the hypothesis that sSNVs may play a role in human genetic diseases due to their effects on mRNA structure. The SPI score and our computed Vienna metrics provide a means of gauging the structural constraint operating on any sSNV. Given that up to 75% of patients with a suspected rare genetic disease lack a molecular diagnosis, our score has the potential to enable discovery of novel etiologies in human genetic disease. Our RNA Stability Pipeline as well as Vienna structural metrics and SPI scores for all human synonymous SNPs can be downloaded from GitHub https://github.com/nch-igm/rna-stability.

INTRODUCTION

While next generation sequencing (**NGS**) has accelerated the discovery of new functional variants in syndromic and rare monogenic diseases, many more disease-causing genes and novel genetic etiologies remain to be discovered (Wright *et al.* 2015; Deciphering Developmental Disorders Study 2017). Accurate molecular genetic diagnosis of a rare disease is essential for patient care (Wright *et al.* 2018),yet today's best molecular tests and analysis strategies leave 60-75% of patients undiagnosed (Yang *et al.* 2013; Yang *et al.* 2014b; Ellingford *et al.* 2016; Hegde *et al.* 2017; Worthey 2017). Current clinical practice for sequence variant interpretation focuses primarily on missense, nonsense or canonical splice variants (Richards *et al.* 2015), with numerous bioinformatics prediction algorithms and databases developed for functional prediction and annotation of non-synonymous single-nucleotide variants (nsSNVs) that impact protein function through changes in the underlying coding sequence (Alfares *et al.* 2018). However, these algorithms are inadequate to infer pathogenicity in non-protein-altering variants such as intronic or synonymous variants, which are under different and weaker evolutionary constraints (Gelfman *et al.* 2017). While the potentially pathogenic impact of non-synonymous single nucleotide variants (**nsSNVs**) that change the protein sequence are well understood, we have limited knowledge in regard to the role that synonymous SNVs (**sSNVs**) may have in human health and disease.

Synonymous variants result in codon changes that do not alter the amino acid sequence of the translated protein and as such were referred to as "silent" variants as they were initially considered to have no functional impact. However, there is a growing body of evidence demonstrating that synonymous codons have vital regulatory roles (Fahraeus *et al.* 2016; Lee *et al.* 2017; Ramanouskaya and Grinev 2017; Vaz-Drago *et al.* 2017; Hanson and Coller 2018) among the most important of which is their contribution to RNA structure.

Messenger RNA (**mRNA**) is a single-stranded molecule that adopts three levels of structure: the *primary sequence* forms base pairs among its own nucleotides to build the *secondary structure*, which further folds through covalent attractions to form the *tertiary structure* (**FIGURE 1**) (Silverman 2008). While the tertiary structure of mRNA is challenging to model and poorly understood, sophisticated tools exist to compute the ensemble of possible secondary structures and determine the optimal structure for a given mRNA strand (Lorenz *et al.* 2011).

Studies first published in 1999 indicated that stable mRNA secondary structures are often selected for in key genomic regions across all kingdoms of life (Seffens and Digby 1999; Katz and Burge 2003; Chamary and Hurst 2005; Gu *et al.* 2010). Synonymous variants impacting RNA structure can alter global RNA stability, where stable mRNAs tend to have longer half-lives and less stable RNA molecules may be more rapidly degraded resulting in lower protein levels (Duan and Antezana 2003; Wan *et al.* 2012; Lazrak *et al.* 2013; Hunt *et al.* 2014; Shah *et al.* 2015; Bevilacqua *et al.* 2016). The stability of an mRNA transcript affects translational initiation and can determine how quickly a given protein is translated (Seffens and Digby 1999; Katz and Burge 2003; Chamary and Hurst 2005; Yang *et al.* 2014a; Presnyak *et al.* 2015; Bazzini *et al.* 2016). Recent studies strongly linked mRNA structure to protein confirmation and function, with synonymous codons acting as a subliminal code for the protein folding process (Plotkin and Kudla 2011; Chaney and Clark 2015; Presnyak *et al.* 2015; Faure *et al.* 2016; McCarthy *et al.* 2017; Hanson and Coller 2018). mRNA structure can also facilitate or prevent miRNAs and RNA-binding proteins from attaching to specific structural motifs (Fernandez *et al.* 2011; Brummer and Hausser 2014; Savisaar and Hurst 2017; Dominguez *et al.* 2018). Given these multiple mechanisms, when synonymous variants are ignored, we are almost certainly missing novel plausible explanations for genetic disease.

The role of mRNA structure in human health and disease, however, is poorly understood and relatively few pathogenic variants impacting mRNA folding have been described (Duan and Antezana

2003; Wan *et al.* 2012; Hunt *et al.* 2014; Bevilacqua *et al.* 2016). A structure-altering sSNV in the dopamine receptor DRD2 was shown to inhibit protein synthesis and accelerate mRNA degradation (Duan *et al.* 2003). A sSNV in the *COMT* gene, implicated in cognitive impairment and pain sensitivity, was shown *in vitro* to constrain enzymatic activity and protein expression (Nackley *et al.* 2006). A sSNV discovered in the OPTC gene of a glaucoma patient resulted in decreased protein expression *in vivo* (Acharya et al. 2007). In cystic fibrosis patients, a sSNV in the *CFTR* gene was linked to decreased gene expression (Bartoszewski *et al.* 2010). Additionally, a silent codon change, I507-ATC→ATT, contributes to CFTR dysfunction by a change in mRNA secondary structure that alters the dynamics of translation leading to misfolding of the CFTR protein (Lazrak *et al.* 2013; Shah *et al.* 2015). Two sSNVs in the NKX2-5 gene decreased the mRNA's transactivation potential in a yeast-based assay (Reamon-Buettner et al. 2013). In hemophilia B, the sSNV c.459G>A in factor IX impacts the transcript's secondary structure and reduces extracellular protein levels (Simhadri *et al.* 2017), and both synonymous and nonsynonymous SNVs were shown more likely be deleterious when occurring in a stable region of mRNA in hemophilia associated genes F8 and Duchenne's Muscular Dystrophy (Hamasaki-Katagiri *et al.* 2017).

We hypothesize that these reported instances of mRNA structure playing a role in disease represent only the tip of the iceberg and that many undiagnosed genetic disorders might also be influenced by disruptions to mRNA structures. As such, the goals of this study were the creation of metrics to predict a sSNV's pathogenicity due to its effects on mRNA structure and to utilize these metrics to test the hypothesis that synonymous variants predicted to have disruptive impacts on RNA stability would show significant constraint in the human population. In successfully doing so we hope to provide the genetics research community with tools to identify novel genetic etiologies in both monogenic genetic disorders and more complex human disease, thus leading to improved diagnosis and the possibility of novel prevention and treatment approaches.

RESULTS

Massively parallel generation of RNA stability metrics

Global assessment of sSNVs is truly a big data problem as it requires generation and evaluation of several raw values for each of hundreds of millions of positions within the genome. To address this challenge and successfully predict the mRNA-structural effects of every possible sSNV, we developed novel software built upon the Apache Spark framework (**FIGURE 2**). Apache Spark is a distributed, open source compute engine that drastically reduces the bottleneck of disk I/O by processing its data in memory whenever possible (Zaharia *et al.* 2012). This leads to a 100x increase in speed and allows for more flexible software design than can be achieved in the traditional Hadoop MapReduce paradigm. Spark is well suited to address many of the challenges faced in analyzing big genomics data in a highly scalable manner and adoption is growing steadily, with applications such as SparkSeq (Wiewiorka *et al.* 2014) for general processing, SparkBWA (Abuin *et al.* 2016) for alignment and VariantSpark for variant clustering (O'Brien *et al.* 2015). By developing a solution within this framework, we eliminate significant computational hurdles standing in the way of large-scale analysis of sSNVs.

We used the RefSeq database (Release 81, GRCh38) as the source for all known human coding transcript sequences. At each position within a given transcript, four 101-base sequence windows were built, differing only in their central nucleotide, which was set to the reference nucleotide or one of the three possible alternate bases. Using Apache Spark in the Amazon Web Services (AWS) Elastic Map Reduce (EMR) service, we developed a massively parallel implementation of the ViennaRNA Package to analyze the four possible sequences. This enabled us to examine changes in mRNA folding that result from any given polymorphism, and thereby obtain ten metrics which quantified the SNV's effect on mRNA secondary structure (see **SUPPLEMENTARY TABLE 1**). First, we utilized RNAfold to obtain predicted free energies for both mutant and wildtype sequences, which we compared directly to obtain

four metrics describing the sSNV's effect on mRNA stability. Next, we fed the predicted structures from RNAfold into the Vienna programs RNApdist and RNAdistance to obtain 6 additional metrics quantifying the change in base-pairing and ensemble diversity due to each SNV. We performed this procedure for all 469 million possible SNVs in 45,800 transcripts.

After pre-processing we assigned each sSNV a classification based on the most deleterious role it played in any transcript, in decreasing order of deleteriousness: start loss, stop gain, start gain, stop loss, missense, synonymous, 5 prime UTR, 3 prime UTR. We then focused on the set of 22.9 million synonymous variants. While non-synonymous variants also play a role in mRNA structure, we chose to exclude 63.8 million nsSNVs from the subsequent analysis as their impact on conserved amino acid sequences would make it difficult to discern constraint at the mRNA structural level. We also filtered out variants implicated in splicing or lacking annotations needed in future steps, leaving us with a core dataset of 17.9 million sSNVs (see **METHODS** for details, **FIGURE 2** for a summary of our computational pipeline, and **SUPPLEMENTARY TABLE 2** for a record of the number of SNVs filtered at each stage). Of the 10 mRNA-structural metrics computed for each sSNV we adopted three as the primary focus for our analysis: dMFE, CFEED, and dCD. The metric dMFE (delta Minimum Free Energy) measures the change in overall mRNA stability imputed by the sSNV, while CFEED (Centroid Free Energy Edge Distance) gives the number of base pairs that vary between the mutant and wildtype centroid structures. The metric dCD (delta Centroid Distance) measures the sSNV's effect on the diversity of the mRNA's structural ensemble.

To test whether certain sSNVs are under constraint due to their effect on mRNA structure, RNA folding metrics from our Vienna pipeline were combined with population frequencies from the Genome Aggregation Database (gnomAD), containing aggregate WGS and WES data from a total of 138,632 unrelated human individuals (Lek *et al.* 2016). Our expectation was that SNVs with disruptive structural properties would be found less frequently in human population. Constrained variants were defined as those

absent from gnomAD, versus un-constrained variants as being those with exon minor allele frequency >0, a strategy similar to that employed by other groups (Gronau *et al.* 2013; Huang *et al.* 2017).

Global constraint to maintain stability

Our study reveals a striking connection between a given sSNV's impact on mRNA structure and its frequency in the gnomAD database. We define the central variable Y to be Y=1 when a sSNV is present in gnomAD and Y=0 when the sSNV is absent. Synonymous variants that disrupt structure tend to have Y=0 (i.e. are absent from the gnomAD database), while those with limited impact on structure tend to have Y=1 (i.e. appear at least once in the gnomAD database). This central finding is summarized in **FIGURE 3**, which shows the proportion of synonymous SNVs with Y=1 at every value of the metrics dMFE, CFEED and dCD (note: here and throughout, dCD values are rounded to the nearest integer). The leading **FIGURE 3A** shows the correlation between Y and the stability metric dMFE. The bell-shaped distribution shows that Y=1 occurs most often among those sSNVs that maintain the mRNA's existing level of stability, i.e. those sSNVs with dMFE close to 0. When the sSNV either over-stabilizes the mRNA (low dMFE) or de-stabilizes it (high dMFE) the sSNV is depleted in the population roughly in proportion to the level of disruption.

FIGURE 3B shows an analogous plot for the structural disruption metric CFEED (see **SUPPLEMENTARY FIGURE 2** for an illustration of how CFEED is calculated). This plot appears to depict two separate trends, but actually shows a single pattern that alternates between high and low on successive values: the SNVs with CFEED=0,4,8,12... are enriched over those with CFEED=2,6,10,14... (CFEED can only take on even values because the destruction/creation of a base pair always requires two edits). One possible explanation for this duality is that when CFEED fails to be divisible by 4, there is necessarily a change in the total number of base-pairings in the mRNA centroid structure. Thus, sSNVs which conserve

the total number of base-pairs could be potentially favored. **FIGURE 3B** also supports the hypothesis that structurally disruptive sSNVs should appear less frequently in the population. We see that sSNVs which leave the centroid structure unchanged (i.e. CFEED=0) are roughly 20% more common than those sSNVs predicted to alter it. And within each of the two separate trends (that is, the multiples and non-multiples of 4) the population frequency declines as the number of centroid base-pairing changes grows from small to large.

Finally, sSNVs which either diversify the ensemble of mRNA structures (high dCD) or homogenize it (low dCD) are depleted in the population proportionately to their disruptions, as shown in **FIGURE 3C**. The symmetry in depletion between over- and under-diversifying sSNVs is surprisingly regular.

The relationship between the three metrics is illuminated by color-coding in **FIGURE 3.** We observe in **FIGURES 3A AND 3B** that disruptions in the magnitude of stability (|dMFE|) and base-pairing (CFEED) of a sSNV are markedly correlated, with the two metrics enriched for each other at extreme values (red coloring). **FIGURE 3C** depicts a clear relationship between diversity and stability, with those sSNVs diversifying the ensemble (high dCD) also tending to de-stabilize it (blue). This diversity-instability relationship is intuitive, as a destabilizing mutation "frees up" portions of the mRNA to assume new forms. Together, these observations validate the central hypothesis that sSNVs which disrupt mRNA structure should be constrained in human populations.

Variation of constraint with REF>ALT context

Since an mRNA's secondary structure is largely determined by its primary structure (i.e. by the sequence of nucleotides), we would expect the constraint in **FIGURE 3** to be partially dependent on sequence features around each sSNV. To fully determine the role of non-structural variables in the trends

of **FIGURE 3**, we first control for the most important sequence-variables, the REF and ALT of the sSNV. We divide our sSNVs into 14 classes (**TABLE 1):** 12 classes based on their reference and alternate alleles (e.g. A>C, C>G, T>C, etc.) and 2 additional classes based on potential loss of methylated cytosine (CpG>TpG or CpG>CpA, the latter of which results from a deamination on an antisense strand). Within each REF>ALT context we reconstruct the three plots of **FIGURE 3** and also perform weighted linear and quadratic regressions between the three different stability metrics and Y=1 (see **METHODS** for details). All significant results ($p < 0.005$) of this procedure appear in **TABLE 1**.

Looking at **TABLE 1A** (which shows the results for dMFE) we find that disruptions to mRNA stability are constrained across many of our sSNV classes. The fact that most of linear p-values are much smaller than the quadratic p-values indicates that in most contexts the dMFE-Y relationship is *linear,* in contrast to the bell-shaped relationship we see when considering global dMFE (**FIGURE 3A**). Therefore, the slope of the regression line indicates which direction of dMFE is enriched for Y=1. For example, in the context of G>T the negative normalized slope indicates that lower dMFE values (i.e. stabilizing) are less constrained (i.e. Y=1). The slope of the regression line (and the relationships it models) proves to depend largely on whether a context's REF and ALT nucleotides are "strong" (C,G) or "weak" (A,T) binders. We note from **TABLE 3A** that strong>weak mutations consistently have negative slopes (except in the irregular context G>A; see *Constraint for mRNA stability in non-CpG-transitional contexts*), while the two weak>strong contexts A>G and T>C have positive slopes.

In **TABLE 1B** we observe the constraint for our structural disruption metric CFEED. The results here are surprising – the contexts are split between positive and negative slopes. In support of our hypothesis, four of the sequence contexts display a negative slope, implying that sSNVs with high CFEED values are constrained. However, in contrast to our hypothesis, three of the sequence contexts have a positive slope, which implies that sSNVs in these contexts with high CFEED values are enriched. In the

case of CpG>TpG mutations the low quadratic p-value indicates that the pattern is actually bell-shaped, with both low and high CFEED values being depleted; but in G>A and C>A contexts the quadratic term is not significant. Actual plots of these patterns reveal that the ones in which CFEED is depleted are more striking (see **FIGURE 4** and **SUPPLEMENTARY FIGURES 4-5** for plots of Y vs. CFEED in all stability-significant contexts), but this peculiar result must still be addressed. We speak more on this topic in the **DISCUSSION.**

Finally, **TABLE 1C** shows mutation contexts that are significantly constrained against changes to ensemble diversity. We see that only a few contexts experience this constraint. But when significant, the constraint for diversity appears to inherit the bidirectionality of **FIGURE 3C** (with the quadratic term being the most significant and the linear fit being very poor). In these contexts, decreases and increases to ensemble diversity appear to be equally harmful.

CpG transitions have constraint against de-stabilization of their mRNA structures

The data in **TABLE 1** highlight that our observed constraint for mRNA structure is the greatest when considering CpG transitions. Since these variants (and their suppression) are crucial to the story of mRNA stability, it is important to have an appreciation of their role in a biochemical context. The dinucleotide CG (usually denoted CpG to distinguish this linear sequence from the CG base-pairing of cytosine and guanine) is capable of becoming methylated and then mutating by a process called "deamination" into a TG dinucleotide. While studies have demonstrated that methylated CpG residues are up to 40X times more likely to be deaminated than their unmethylated counterparts (Vinson and Chatterjee 2012), mechanisms exist to enzymatically repair CpG deaminations (Morgan *et al.* 2007; Bellacosa and Drohat 2015). In mammals 70-80% of CpGs are methylated, which makes a CpG transition almost 5x more common than any other mutation-type among mammals (see **SUPPLEMENTARY DATA TABLE 3**) (Li

and Zhang 2014). Possible explanations for the distribution and retention of CpGs in mammals have been extensively debated, with some arguing that the phenomenon is not even the result of selective forces (Cohen *et al.* 2011).

The nucleotides C and G also form the foundation of mRNA secondary structures. Most of the energy of an mRNA structure lies in its "stacks" of nucleotides with the average energy of a C-G pair in a stack around 65% stronger than that of any other base-pairing (Turner and Mathews 2010). Moreover, the self-complementarity of CpGs means that upstream and downstream instances can bind together and form a four-base stack which other base-pairs can then build around.

In the present study we find strong evidence that CpG transitions are constrained against de-stabilization of their mRNA structures. This striking trend is largely explained (in a statistical sense) by CpG content, i.e. number of CpG dinucleotides in the surrounding 120 nucleotides of the mRNA transcript (see "Depleter R^2 in **TABLE 1B**). We distinguish CpG>CpA versus CpG>TpG transitions (the former of these usually results from a CpG>TpG deamination on an anti-sense DNA strand), as these two mutation-types show a qualitatively different constraint for mRNA structure. **FIGURE 4** shows the performance of our three main metrics in CpG-transitional contexts. Most strikingly, we find that synonymous CpG>CpA and CpG>TpG mutations both show a steady constraint against de-stabilization (high dMFE) (**FIGURES 4A & 4B**). Fascinatingly, both contexts exhibit a cluster of outliers in the most destructive (i.e. most de-stabilizing region), suggestive of extreme constraint borne of significant structural disruption. Though the two plots exhibit the same basic shape, the context CpG>CpA of **FIGURE 4A** shows higher de-stabilizing tendencies (higher dMFE values) and also a stronger constraint (lower P(Y=1)).

The behavior of the edge metric CFEED in these contexts is less clear-cut. In **FIGURE 4**C we see a clear pattern of constraint against mutations with high CFEED values; and the red coloring shows that such changes are, on average, de-stabilizing. But the constraint in the context CpG>TpG (**FIGURE 4D)** is

much less forceful (in fact, its quadratic p-value is much smaller than its linear) and the blue coloring by dMFE shows such mutations are on average neutral or even de-stabilizing. Finally, **FIGURES 4E & 4F** show that the basic pattern of constraint for diversity in **FIGURE 3C** is reproduced and is essentially unchanged for both types of CpG transition. The coloring again indicates that mutations CpG>CpA are much more destabilizing than their CpG>TpG counterparts.

The markedly greater constraint and tendency towards de-stabilization among CpG>CpA transitions suggests they are under different selective pressures than CpG>TpG transitions, despite being largely produced by the same biochemical mechanism (a CpG>TpG deamination on either a positive- or negative-sense strand – see **SUPPLEMENTARY DATA TABLE 3**). We speculate on this disparity in the **DISCUSSION.**

Constraint for mRNA stability in non-CpG-transitional contexts

We see the strongest constraint for mRNA structure in CpG transitions, but we observe an analogous pattern in most REF>ALT contexts (as indicated by **TABLE 1**). We can classify these remaining contexts based on whether their slopes in **TABLE 1A** are positive or negative. **SUPPLEMENTARY FIGURE 4** shows plots of contexts where dMFE and the gnomAD variable Y are negatively correlated. In such contexts the data are consistent with the hypothesis that sSNVs which de-stabilize mRNA are constrained. Notably, all these contexts are strong>weak (or strong>strong in the case of C>G), consistent with the principle that one purpose of such nucleotides is to maintain stability. The coloring by CFEED indicates that a change in either direction is likely to alter the mRNA secondary structures.

In **SUPPLEMENTARY FIGURE 5** we show the contexts where dMFE and Y vary positively, which amounts to the claim that stabilizing mutations are constrained in these contexts. Correspondingly, we note that two out of three of these contexts are weak>strong (and the third is the unusual context G>A

where SNPs that alter stability or diversity are actually enriched). The context T>C exhibits a notable constraint in either direction, an anomaly which we speculate on in the **DISCUSSION.**

Depleter variables

In **TABLE 1** we provide a "Depleter" for the connection between our RNA folding metrics and gnomAD frequencies for each mutational context. The name "Depleter" signifies that each such variable is chosen so as to correlate negatively with gnomAD (which is why these variables are given with +/- signs in **TABLE 1**). For example, the Depleter for dMFE in the context CpG>CpA is +CpG content, meaning that when CpG content increases in this context, the varaible Y is depleted.

The Depleter is chosen to be the variable that best explains the connection between the mRNA structural variable and Y in the given context. The proportion of connection explained is given by the field "Depleter R^2". For example, in the context CpG>CpA we can explain 78% of the dMFE-gnomAD connection using a model that relies only CpG content.

To determine which variable is most informative (and should therefore be called the Depleter) we compute an associated R^2 for a set of features of the sequence around the sSNV (the upstream/downstream nucleotides and the proportion of A, C, G, T, CpG or ApT [di]nucleotides in the surrounding 120 bases). Each of these features is used to build a simple logistic model to predict Y=1, and the predictions of the model are then compared to the actual proportion P(Y=1) at a value of the metric. For example, building a CpG-context-based model allows us to compute the quantity P(Y=1 | CpG content), and then we consider the difference:

$$\mathbf{E}(\,\mathbf{P}(\mathrm{Y}=1 \mid \mathrm{CpG\ content}) \mid \mathrm{dMFE}) \;-\mathbf{P}(\mathrm{Y}=1 \mid \mathrm{dMFE}\,)$$

Squaring this difference and taking a weighted sum over all values of dMFE in a context, we recover the variance left unexplained by a particular non-structural variable. We obtain an R^2 by comparing

unexplained variance to that obtained using a null model, and the Depleter is then the variable with the largest R^2 (see **METHODS** for more details). **TABLE 1** shows that Depleters can recover large portions of the trends in **FIGURE 4** and **SUPPLEMENTARY FIGURES 4-5.** The striking trend between dMFE and gnomAD frequency in CpG-transitional contexts is largely driven by the proportion of CpGs in the surrounding 120 nucleotides (73% for CpG>TpG sSNVs and 79% for CpG>CpA sSNVs). CpG content is also the most powerful feature when accounting for the behavior of CFEED and dCD in these contexts, with high CpG content consistently correlating with depletion. The natural inference is that an abundance of CpGs signifies important mRNA structure nearby, the disruption of which could be deleterious.

In non-CpG-transitional contexts, the Depleter almost always proves to be a nucleotide upstream or downstream of the sSNV. In the context C>A we can recover 28% of the relationship between dMFE and gnomAD frequency simply by looking at whether the C is followed by a G. The power of CpG dinucleotides in recovering our structural trends in the contexts C>A, C>G, G>C and then G>T, emphasizes the powerful but poorly understood role of CpGs in both mRNA stability and mammalian genomes.

Global quantification of mRNA constraint

Our analysis shows that polymorphisms predicted to influence mRNA secondary structures are constrained in the population. However, due to the multiple facets that need to be considered when studying RNA secondary structure, by focusing on a single RNA-folding metric such as dMFE or CFEED, we run the risk of missing functionally relevant information. To overcome this potential limitation of our RNA folding metrics, we set out to devise a more diversified method for predicting possible pathogenicity due to mRNA structure. Our strategy is to consider the additional statistical power bestowed by mRNA structure. In each of our 14 sequence contexts from **TABLE 1** we build two general logistic models for

predicting MAF >0: a null model that uses the natural variables of sequence context, local nucleotide composition, transcript position and tRNA propensity, but NOT mRNA structure (***n***); and a structural model which also includes the 10 metrics obtained from our Vienna analysis (***s***). These models yield two separate probability-predictions P_n and P_s for the quantity P(MAF >0) (see **METHODS** for details). Then we define the metric:

$$\text{SPI} = \log_{10}\left(\frac{P_s}{P_n}\right)$$

The metric SPI thus measures the additional predictive power bestowed by mRNA-structural variables. When it varies from 0, mRNA structural predictions yield new insight about a SNV's potential to have a functional role in mRNA secondary structure. The power of SPI in each context (given by its area under the curve in predicting whether gnomAD is >0) is supplied in **TABLE 2** and we plot SPI vs. Y in CpG-transitional contexts in **FIGURE 5** (and in all contexts in **SUPPLEMENTARY FIGURE 6**). The classification rules of SPI vary widely by context. We see the most impressive performance in the context of CpG transitions. For both CpG>CpA and CpG>TpG transitions, those sSNVs with low SPI values are clearly under constraint.

The behavior of SPI in non-CpG-transitional contexts is less regular and harder to weave into a coherent story. Every context shows a clear pattern, but this may amount to either enrichment or depletion (or both) as SPI moves in either direction. Given the strong dependence on REF-ALT context, the use of SPI as a deleteriousness score in non-CpG may need further evaluation.

Clinical Examples of Structural Pathogenicity

The literature reveals only a few examples of synonymous sSNVs unequivocally shown to be pathogenic through their effects on mRNA structure. These sSNVs, with accompanying values of our three Vienna metrics and SPI, are listed in **TABLE 3.** The sSNVs show a definite enrichment for our

structural metrics as each shows a value of |dMFE|, CFEED, |dCD| or |SPI| that is in at least the 80th percentile in its context. For example, the pathogenic sSNV in NKX2-5, linked to congenital heart disease, has a dCD score in the 90th percentile (Reamon-Buettner *et al.* 2013). It should be noted that none of these clinical sSNVs qualifies as a truly exceptional outlier for any of our Vienna metrics or SPI. None of the clinical sSNVs rises above the 95th percentile for |dMFE|, CFEED, |dCD| or |SPI|. We address this surprising "moderateness" in known pathogenic sSNVs in the **DISCUSSION**.

DISCUSSION

We have shown that *in silico* mRNA structural predictions can be used to predict and explain the population allele frequency of a synonymous variant. By calculating Vienna RNA folding metrics for nearly 0.5 billion possible SNVs, we demonstrate that there is significant selection against sSNVs that are predicted to either stabilize or de-stabilize the given transcript's local mRNA secondary structure. While the observed trends can be partially explained by sequence-based variables like CpG or GC content or membership in a CpG/AT/TA dinucleotide (as given by the "Depleter" field in **TABLE 1**), we believe our data supports our hypothesis that RNA structure itself plays a critical role in human health and disease. As such, polymorphisms impacting mRNA structure are under negative selection in the population and should be more carefully evaluated in the context of both Mendelian disorders and complex human disease.

When determining if the connection between mRNA disruption and population incidence is direct and causal, we need to consider a number of factors. First, constraint of mRNA structure *must* be exercised through sequence-based variables, since the underlying primary mRNA sequence largely determines the secondary structure. Thus, although the trends we have observed may be influenced by sequence-features, such as CpG content (as illustrated by the "Depleter" variable in **TABLE 1**), it does not necessarily indicate that the trends are spurious. Second, it is important to note that our trends operate in the directions implied by our hypothesis: sSNVs that disrupt mRNA (measured three different ways) are depleted rather than enriched in the population for almost all REF>ALT contexts. Finally, mutations that are predicted to change stronger base pairs to weaker ones are consistently constrained against de-stabilization rather than over-stabilization. If the association were spurious, we would not expect such agreement with prediction.

Our data also include some irregularities which can be elegantly explained through mRNA structure. For example, when considering CFEED in **FIGURE 3B** we observed that sSNVs were enriched

when their CFEED values were multiples of 4. Since a CFEED value being divisible by 4 is a necessary condition to preserve the total number of base-pairs, this observed enrichment suggests changes to base pairing were constrained. We also observed bi-directional constraint for dMFE in the context T>C, visible in **SUPPLEMENTARY FIGURE 5** and also inferable from the low quadratic p-value in **TABLE 1**. We conjecture the dual constraint in this context might be due to guanine's unique ability to wobble base-pair. Wobble base-pairing occurs between two nucleotides such as guanine-uracil (G-U), that are not canonical Watson-Crick base pairs, but have comparable thermodynamic stabilities. Unlike G-U, the three other main examples of wobble-base pairs (hypoxanthine-uracil (I-U), hypoxanthine-adenine (I-A), and hypoxanthine-cytosine (I-C)) all require the non-standard purine derivative hypoxanthine. Thus, the dual constraint from mutations T>C could be related to the transformation of T=G wobble base-pairs into stronger C=G Watson-Crick base pairs.

Finally, in addition to the metrics output by our Vienna analysis, we devised our own metric to measure structural pathogenicity. The Structural Predictivity Index (**SPI**), created specifically to control for all confounding factors, shows that mRNA structure has predictive power all by itself. Also, the clusters of outliers at the extreme values of our structural metrics (see **FIGURES 4** and **SUPPLEMENTARY FIGURES 4,5)** suggest a constraint beyond that explained by a confounding variable.

Taken together, this evidence provides significant support for the hypothesis that disruptions to mRNA structure are directly under constraint. However, we realize that in addition to the structural role that the primary mRNA sequence plays, there are other molecular mechanisms at work in the regulation of the transcriptional and translational processes. For example, while the retention of CpG dinucleotides is certainly connected to mRNA structure, other factors such as tRNA binding, binding of miRNAs and other RNA binding proteins, DNA chromatin structure and epigenetic modifications in the ORF could also be involved. Relatedly, we found a few contexts where sSNVs which disrupt mRNA structure actually

have higher population frequencies than those that do not (**TABLE 1)**, the main example being the enrichment of high CFEED values in the contexts C>A and G>A. In both cases, the Depleter variable is a trailing A, which correlates negatively with gnomAD frequency and CFEED. The presence of an A is likely to have minimal effect on mRNA structure, suggesting that in these contexts the connection is partly spurious.

Regulation of CpG transitions

In the case of CpG transitions, it is difficult to state whether selection for mRNA structure causes the retention of CpGs, or whether the retention of CpGs is regulated by a process independent of mRNA structure. A strong reason for CpGs to operate causally with regard to mRNA structure is that they are the single most important determinant of mRNA structure. Retention of 5' ORF CpG sites occurs at a high frequency in the first exon of coding genes; a stacked C:G + G:C base pairing, has the lowest free energy of the 36 possible stacked base-pair combinations (Mathews *et al.* 1999); and deamination of CpGs can be suppressed and repaired by existing enzymatic mechanisms. Thus, CpG dinucleotides represent the easiest and most natural way to determine mRNA structure.

Importance of CpG and AT dinucleotides

Our results in non-CpG-transitional dinucleotide contexts are largely explained by the reference nucleotide's membership in a CpG/AT dinucleotide. These dinucleotides have the apparent effect of mitigating the structural distortion caused by the sSNV, e.g. mutations C>A and C>G are less de-stabilizing if the reference is part of a CpG (see **TABLE 1**, which shows that in these contexts the gnomAD variable Y varies inversely with dMFE but directly with a trailing G). This presents us with the same causal conundrum we have faced throughout our study: do sSNVs in a dinucleotide have higher

frequencies because the dinucleotide mitigates the structural damage, or is it due to some other reason, unrelated to mRNA structure? While this question is difficult to answer definitively, we believe that the data presented in this present study and that the body of mRNA structural literature supports that preservation of mRNA secondary structure is acting as a functional constraint on sSNVs in a dinucleotide. For example, one point in favor of a causal role is that both CpGs and ATs have been specifically implicated as drivers of mRNA structure (Al-Saif and Khabar 2012). Moreover, consistent with our findings, a seminal paper in the field of RNA folding suggested that it is the *dinucleotide* content of an mRNA that contributes most to its stability (Workman and Krogh 1999).

Successful identification of structurally disruptive sSNVs in known pathogenic synonymous variants

Over the last decade numerous studies have demonstrated that synonymous variants play essential molecular roles in regulating both mRNA structure and processing, including regulation of protein expression, folding and function (reviewed in Sauna and Kimchi-Sarfaty 2011; Shabalina *et al.* 2013; Fahraeus *et al.* 2016). However, the potential for pathogenic synonymous variants that impact RNA folding in human genetic disease remains largely unknown. Current American College of Medical Genetics (**ACMG**) guidelines for the assessment of clinically relevant genetic variants focus primarily on missense, nonsense or canonical splice variants (Richards *et al.* 2015). These guidelines suggest that synonymous "silent" variants should be classified as likely benign, if the nucleotide position is not conserved and splicing assessment algorithms predict neither an impact to a splice consensus sequence nor the creation of a new alternate splice consensus sequence. In the absence of functional tools that would aid in the simultaneous assessment of both nsSNVs and sSNVs in a given patients genome, we are almost certainly missing novel disease etiologies that have their molecular underpinnings in pathological alterations to mRNA structure.

Numerous *in silico* tools exist to aid in the prediction of disease-causing missense variants, and have accuracy in the 65%-85% range when evaluating known pathogenic variants (Li *et al.* 2018). Such algorithms infer pathogenicity based on amino-acid substitutions (SIFT (Kumar *et al.* 2009), PolyPhen (Adzhubei *et al.* 2010), FATHMM (Shihab *et al.* 2015)), nucleotide conservation (SiPhy (Garber *et al.* 2009), GERP++ (Davydov *et al.* 2010)) or an ensemble of annotations and scores (CADD (Kircher *et al.* 2014), DANN (Quang *et al.* 2015), REVEL (Ioannidis *et al.* 2016)). These tools predict whether a nsSNV is pathogenic or benign, primarily due to the high conservation of protein sequences. However, these algorithms are not equipped to assess pathogenicity in synonymous variants, which are under different constraints (Gelfman *et al.* 2017). Recognizing that there is a critical need for methods that better predict the potential whether sSNVs have pathogenic impact and function, our goal in this present study was the generation of such metrics. Vienna RNA stability and SPI metrics are available for download for all known sSNVs, to enable researchers and clinicians to evaluate WES and WGS data in combination with tools such as Annovar (Wang *et al.* 2010), SnpEff (Cingolani *et al.* 2012) and VEP (McLaren *et al.* 2016).

At this present time a comprehensive evaluation of our metrics is not possible as there are simply too few known examples of pathogenic synonymous variants in human genetic disease. While we found approximately a dozen examples of sSNVs implicated in human disease, several merely suggested that a sSNV may have a role through modification of mRNA structure, but lacked functional studies to conclusively implicate the given variant in disease. As such we focused on a set of six sSNVs that we believe the authors unequivocally demonstrated to be pathogenic through their effects on mRNA structure (**Table 3**). This dataset included one variant in OPTC associated with glaucoma (Acharya *et al.* 2007), two variants in NKX2-5 associated with congenital heart defects (Reamon-Buettner *et al.* 2013), one variant in DRD2 associated with post-traumatic stress disorder (Duan *et al.* 2003), and two variants in COMT associated with pain sensitivity (Nackley *et al.* 2006).

All six sSNVs demonstrated definite enrichment for our structural metrics, be it stability, edge distance, diversity or SPI, with values in the 80th to 90th percentile range. However, none of these clinically relevant sSNVs qualifies as a truly exceptional outlier for any of our Vienna metrics or SPI with all percentiles being below 90. It is theoretically possible that such extreme outliers are not biologically tenable, making them less likely to appear in the human population. As such, a change in the 80th percentile could represent a cutoff for biological significance. Another possibility (perhaps equally strong) is that these sSNVs occupy important regulatory positions, and that a sSNV deleterious to mRNA secondary structure may exhibit pathogenicity when it distorts structure *in a key region* of the transcript.

The enrichment of our structural metrics, while moderate, is still clear and our hope is that future studies will allow refinement and enhancement of our metrics. As new discoveries of pathogenic sSNVs in human genetic disease occur, a larger data set of known clinically relevant sSNVs will help determine cutoff values. For now, our recommendation is that a conservative 80th percentile cutoff across the four metrics is used initially, but this may need to be lowered to reveal pathogenic sSNVs that have a less extreme change to mRNA structure.

Mitigation of competing constraints

In addition to a potential role in mRNA structure, synonymous codons are likely under selection for purposes other than mRNA structure, which could have confounded our analysis. Synonymous codon utilization (codon bias) is known to direct gene expression and protein synthesis through regulating tRNA recruitment (Rocha 2004; Sabi and Tuller 2014; Quax *et al.* 2015). Synonymous codons may also act as a subliminal code for protein folding, with changes in a preferred locus potentially leading to pathogenicity in synonymous mutations (McCarthy *et al.* 2017; Hanson and Coller 2018). While the stability of an mRNA transcript can determine how quickly it is translated (Seffens and Digby 1999; Yang *et al.* 2014a;

548 Presnyak *et al.* 2015), translation speed is also regulated through codon usage and the abundance of the
549 tRNAs (Dong *et al.* 1996). This may have a confounding impact on our analysis of constraint, but
550 attempted to mitigate this by including the tRNA Adaptivity Index (a measure of tRNA abundance) in our
551 set of confounding variables.

552 While we took care to exclude sSNVs impacting the canonical splice sites from our constraint
553 analysis, exonic variants beyond the canonical splice site can disrupt splice enhancers (Soukarieh *et al.*
554 2016), or they may also activate cryptic splice sites, leading to aberrant pre-mRNA splicing and loss of
555 coding sequence (Molinski *et al.* 2014). Synonymous mutations that affect the kinetics of translation can
556 slow down the rate of protein synthesis or lead to protein misfolding, which in turn can result in
557 proteotoxicity (Chaney and Clark 2015). Synonymous mutations may also result in the formation of
558 translational "pause sites" and alternative conformations during co-translational folding (Hanson and
559 Coller 2018). Recent genome-wide analyses revealed that bicodons (i.e., pairs of consecutive codons)
560 demonstrate biased usage and confer different pause propensities during the translation process (McCarthy
561 *et al.* 2017). Similar to the scores we present here for assessing a variants impact on protein folding, it will
562 be important for future studies to create scores by which all these possible mechanisms of pathogenic
563 sSNVs could occur.

564

565 ***Molecular mechanisms underlying constraint of variants impacting mRNA secondary structure***

566 While our score does not specifically identify the underlying molecular mechanism, it will aid in
567 identification of sSNVs impacting secondary structure which could confer pathogenicity in numerous
568 ways. For example, sSNVs impacting RNA structure can alter global RNA stability, where less stable
569 RNA molecules may be degraded more quickly resulting in lower protein levels (Duan and Antezana
570 2003; Lazrak *et al.* 2013; Shah *et al.* 2015). As local RNA structure is essential for the translation process,

a more stable mRNA may not be able to initiate translation, also resulting in lower protein levels (Katz and Burge 2003; Chamary and Hurst 2005; Presnyak *et al.* 2015; Bazzini *et al.* 2016). Additionally, numerous studies argue that synonymous codons may also act as a subliminal code for protein folding (Plotkin and Kudla 2011; Chaney and Clark 2015; Presnyak *et al.* 2015; McCarthy *et al.* 2017; Hanson and Coller 2018). Structure-deforming sSNVs exert their pathogenicity chiefly by making the mRNA structure too difficult, or too easy, for the ribosome to process, leading to issues with translation elongation and protein misfolding.

Structural elements within the first 5 to 16 codons of mRNA have been shown to significantly regulate protein expression levels in *E. coli* (Sato *et al.* 2001; Kudla *et al.* 2009). It is likely that both the stability of mRNA folding near the ribosomal binding site and the reduced abundance of tRNAs coding for N-terminal amino acids play crucial roles in slowing down initial stages of translation elongation prevent subsequent ribosomal traffic jams (Tuller *et al.* 2010; Li and Qu 2013). More recently, it was been shown that sequence motifs and mRNA structure within the first five codons are key in dictating the efficiency of protein synthesis (Verma *et al.* 2019). By assessing over 250,000 reporter sequences in *E. coli*, Verma and colleagues demonstrated that differences in this short ramp lead to striking changes in protein abundance, of up to 3 to 4 orders of magnitude. Our own data show marked preservation of CpG dinucleotides, which are crucial for mRNA structure, that appear to be independent of tRNA abundance.

CONCLUSION

We have shown that sSNVs which stabilize or destabilize mRNA are significantly constrained in the human population, thereby supporting a growing body of evidence that previously assumed "silent" polymorphisms, actually play crucial roles in regulation of gene expression and protein function. We have demonstrated that this connection is rich, complex, and biologically intuitive. Given that there are multiple mechanisms by which sSNVs influence biological function, we are almost certainly missing undiscovered disease etiologies when these variants are ignored. In addition to providing the community with a dataset of ten Vienna RNA structural metrics for every known synonymous variant, our Structural Predictivity Index is the first metric of its kind to enable global assessment of sSNVs in human genetic studies. We hope that these metrics will be utilized to accurately assess and prioritize an underrepresented class of genetic variation that may be playing significant and as yet to be realized role in human health and disease.

METHODS

Raw Dataset

To obtain all human mRNA transcripts we downloaded the NCBI RefSeq Release 81 from an online repository (ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/). Transcript sequences corresponded to human reference genome build GRCh38.

Overview of RNA structure prediction process

To estimate the structural properties of a sSNVs we used the ViennaRNA software package, a secondary structure prediction package that has been extensively utilized and continuously developed for nearly twenty-five years. ViennaRNA uses the standard partition-function paradigm of RNA structural prediction (McCaskill 1990). We utilize version 2.0 of ViennaRNA (Lorenz *et al.* 2011). Applying Vienna to every possible SNV in the human genome (about 500,000,000 calculations) was a computationally challenging task which we carried out using an Apache Spark framework powered by Amazon Web Services (AWS). We built a pipeline which read in and analyzed a SNV and stored the results in AWS Simple Storage Service (S3) in Parquet columnar file format (**FIGURE 2**). The ease and capacity of AWS greatly facilitated the project, and the affordability of S3 storage means our data can easily be shared with others. The software we developed is available on GitHub: https://github.com/nch-igm/rna-stability.

RNA structure prediction methodology

To analyze a given SNV we built a 101-base sequence consisting of a central nucleotide at the 51st position (which we set to either the reference or the three alternates) along with the 50 flanking bases on either side. If the nucleotide lay 50 bases from the transcript boundary, the window was simply taken to be the first or last 101 bases in the transcript. We processed these sequences in fasta format with

ViennaRNA's flagship module RNAfold, which yielded three predicted mRNA secondary structures – the minimum free energy, centroid, and maximum expected accuracy structure – as well as numeric values for the free energy of each structure, and a fourth metric measuring the energy of the whole ensemble (see the documentation of (Lorenz *et al.* 2011) for detailed descriptions of these concepts). Comparing the free energies between the wildtype and mutant for each type of structure gave us the four stability metrics delta-MFE (dMFE), dCFE, dMEAFE and dEFE. Next, the predicted structures were processed by the Vienna module RNApdist, which counted the edge-differences to produce the four edge-metrics MFEED (minimum free energy edit distance), CFEED, MEAED and EFEED. As a final step, the predicted structures were further processed by the Vienna program RNAdistance to obtain the diversity metrics dCD and dEND (change in distance from centroid and change in ensemble diversity, respectively).

This whole procedure was carried out using custom developed Spark wrappers of RNAfold, RNApdist and RNAdistance, with slight modifications to the source code to suppress the creation of graphics files. After building our fasta files, we were able to compute all 10 Vienna metrics for over a half billion sequences in less than 24 hours using 51 c4.8xlarge AWS EMR computing nodes.

Construction of final dataset for synonymous SNVs

The next step was to extract the sSNVs. This task was complicated by the fact that a SNV might have appeared in several different transcripts, and could be synonymous in some and non-synonymous in others. To address this challenge, we first annotated every SNV using the program snpEff (Cingolani *et al.* 2012), whose source code was modified to allow record-by-record calling via Spark. This snpEff analysis produced annotations of predicted biotype, e.g. missense, synonymous, canonical splice site, etc. To validate these snpEff predictions we manually predicted the biotype of each SNV using start and stop codon information from RefSeq

(ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/RefSeqGene/refseqgene.*.genomic.gbff.gz). The small number of sSNVs where our predicted biotype disagreed with snpEff's were discarded. We then defined a "synonymous SNV" to be one that was (A) synonymous in at least one transcript, (B) synonymous or within the UTRs in all transcripts, and (C) not implicated in splicing by snpEff. Each sSNV identified as "synonymous" by this scheme was assigned a "home transcript," chosen based on proximity to the start codon, then on maximal transcript length, and then arbitrarily.

This filtration and duplicate-removal process yielded a final set of 17.9 million sSNVs in 34,000 transcripts. See **SUPPLEMENTARY TABLE 2** for a table giving the landscape of our final dataset and the number of sSNVs filtered at each stage.

Merging of sSNV GRCh38 transcript coordinates with gnomAD GRCh37 coordinates

To measure constraint operating on a sSNV we used population frequencies obtained from the gnomAD database. Since this resource only exists for the GRCh37 reference build, we lifted our entire dataset from GRCH38 to GRCh37. The lifting procedure was carried out using the Picard Tools program liftOver , which was executed using a custom Spark wrapper. The joining of the gnomAD frequencies to our main dataset was a task greatly facilitated by Spark's parallel processing and native Parquet support.

Since the great majority (approximately 90%) of sSNVs were marked with gnomAD frequency 0, it is important to identify sSNVs marked zero purely through a lack of coverage. To achieve this, we flagged and removed all sSNVs where fewer than 70% of samples had at least 20X coverage.

Further variant annotations

Next, we estimated the local nucleotide content around each sSNV. We divided each transcript into windows of 40 bases and in each window computed the proportion of A's, C's, G's, T's, CpG's and

AT's in the surrounding three windows. Finally we joined multiple additional annotations (including conservation metrics such as PhyloP) from the dbNSFP dataset (Liu *et al.* 2016). Again, this heavy task was greatly facilitated by our Spark framework.

Partition of dataset

We carried out most of the analysis separately on subsets of data defined by a common mRNA reference and alternate allele, e.g those sSNVs of form C>A. The reference and alternate alleles exert such a huge influence on gnomAD that best solution seemed to be to control for them explicitly. Dividing our dataset based on mRNA alleles (as opposed to DNA alleles, which do not depend on transcript sense) is a step justified in **SUPPLEMENTARY TABLE 3**.

Depleter variables

Depleter variables (so called because they explain some of the gnomAD depletion at values of a structural variable) are given in **TABLE 1**. They are chosen to be the sequence feature that explains the greatest portion of the connection between a structural metric (e.g. dMFE) and Y in a context. Possible Depleter variables are local nucleotide content and the specific nucleotides up/downstream of the sSNV.

To compute the correlation between a structural metric (e.g. dMFE) and Y that is left unexplained by a sequence feature (e.g. CpG content) in a particular REF-ALT context, we first build a simple logistic regression model between CpG content and Y, which gives us an estimate $\mathbf{P}(\mathrm{Y} = 1 \mid \mathrm{CpG\ content})$ for every sSNV in the context (based on the proportion of CpGs in the surrounding 120 nucleotides.) We then plug this "structure-less" estimate into the expression

$$\mathrm{V}_{\mathrm{CpG\ content}} = \sum_{x} \mathbf{n_x} * (\mathbf{E}(\mathbf{P}(\mathrm{Y} = 1 \mid \mathrm{CpG\ content}) \mid \mathrm{dMFE} = \mathrm{x}) - \mathbf{P}(\mathrm{Y} = 1 \mid \mathrm{dMFE} = \mathrm{x}))^2$$

where we sum over all values x of dMFE and let $\mathbf{n_x}$ denote the number of sSNVs in the context with dMFE= x. Comparing this quantity to the null variance

$$V_{null} = \sum_{x} \mathbf{n_x} * (\mathbf{P}(Y = 1) - \mathbf{P}(Y = 1 \mid dMFE = x))^2$$

allows us to compute the proportion of the variation explained by CpG content:

$$R^2_{\text{CpG content}} = 1 - \frac{V_{\text{CpG content}}}{V_{null}}$$

The "Depleter" for a given structural metric in a given context is chosen as the variable with the highest R^2. Finally the correlation between the Depleter and Y was checked, and the Depleter given a sign (+/-) so that the signed Depleter correlated negatively with Y.

Construction of SPI

To construct our final SPI scores we built two separate models over each of our 14 contexts to predict the event MAF > 0. The "null" model used all natural features - the nine nucleotides in the SNV's home and adjacent codons, the proportion of A/C/G/T/CpG/AT's in the surrounding 120 nucleotides, the sSNV's position in its transcript and the transcript's length, and the tAI (tRNA Adapation Index obtained from a supplement of (Tuller *et al.* 2010) from https://ars.els-cdn.com/content/image/1-s2.0-S0092867410003193-mmc2.xls) of the wildtype and mutant codons. The second, "active" model used all these features plus our 10 Vienna metrics. Both sets of variables were then used to predict MAF > 0. We then defined the SPI score for a sSNV to be the base-10 logarithm of the active model's predicted P(Y=1) probability divided by the null model's predicted P(Y=1). Context wise plots and statistics for SPI are given in the **SUPPLEMENTARY FIGURE 6**.

We tried three different model-styles for computing the raw predictions that comprise SPI – general logistic as implemented in python's sklearn LogisticRegression module, random forest as

715 implemented in sklearn's RandomForestClassifier, and gradient-boosted trees as implemented in the
716 python package xgboost. Performance of each SPI "flavor" is given in **SUPPLEMENTARY TABLE 4**. We
717 eventually settled on the general logistic model, as it out-performs the gradient-boosted tree model and
718 does not overtrain as the random forest mode does.
719

COMPETING INTERESTS

The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

J.B.S.G., J.L.L and P.W. developed methodology, performed data analysis and results interpretation. G.E.L. developed AWS Spark Vienna RNA pipeline and developed variant annotation tools. G.E.L. generated folding metrics. J.B.S.G. developed Structural Predictivity Index (SPI). D.M.G., H.C.K., B.J.K, and J.R.F assisted with data analysis, interpretation of results and development of variant annotation tools. J.B.S.G, G.E.L and P.W. prepared figures. All authors contributed to the preparation and editing of the final manuscript.

ADDITIONAL FILES

SUPPLEMENTARY DATA FILE 1: This file contains four supplementary data tables and six supplementary figures further detailing the methodology and results presented in this manuscript.

ACKNOWLEDGEMENTS

We thank the Nationwide Foundation Pediatric Innovation Fund for generously supporting this body of work. James L. Li was supported by the Pelotonia Fellowship for Undergraduate Research through The Ohio State University Comprehensive Cancer Society.

FIGURE LEGENDS

FIGURE 1. A synonymous variant introduces a marked change in local minimum free energy of the mRNA secondary structures in the *DRD2* gene. Using a known synonymous variant of pharmacogenomic significance in the dopamine receptor, DRD2 (NM_000795.4:c.957C>T (p.Pro319=)), this figure demonstrates how the 101-bp window used in our analysis captures the variants impact on RNA secondary structure. Wildtype (**A**) and mutant (**B** and **C**) sequences (RefSeq transcript NM_000795.4, coding positions 907-1008) are identical except for a synonymous C->T mutation at position 51 (major "C" allele is indicated by the black arrow, minor "T" allele is indicated by the red arrow). **(A)** Wildtype optimal and centroid structures (which coincide) demonstrate a relatively stable secondary structure with a minimum free energy of -12.5 kcal/mol. Of the ensemble of possible structures arising from the sSNV a position 51, there is a significant reduction in stability of the molecule in terms of both the (**B**) mutant optimal structure (-11.5 kcal/mol) and (**C**) mutant centroid structure (-5.1 kcal/mol). The synonymous variant results in a less stable mRNA molecule which laboratory studies demonstrate reduces the half-life of the transcript, ultimately reducing protein expression of the dopamine receptor, DRD2. Nucleotides are colored according to the type of structure that they are in: Green: Stems (canonical helices); Red: Multiloops (junctions); Yellow: Interior Loops; Blue: Hairpin loops; Orange: 5' and 3' unpaired region.

FIGURE 2. Graphical depiction of computational workflow used to generate ViennaRNA folding metrics for the entire transcriptome. The entire analysis workflow was parallelized using Apache Spark and the Amazon Elastic Map Reduce (EMR) service, generating 5 billion Vienna RNA metrics over the course of 2 days. Using a custom pipeline developed for the process that was executed across 47 Amazon Elastic Cloud Compute (EC2) spot instances, input data was retrieved from an Amazon

763 Simple Storage Solution (S3) bucket and processed through the pipeline consisting of 8 steps. We first
764 obtained the 101-base sequence centered around a SNV in a transcript and generated three alternate
765 sequences (with the ALT rather than the REF at position 51) (step 1). We next applied Vienna modules to
766 sequence to obtain structural metrics (step 2). Results were then mapped to chromosomal coordinates (step
767 3) and annotated with SnpEff to identify splice variants (step 4), lifted to the hg19 build (step 5), annotated
768 with gnomAD population frequencies (step 6) and coverage information (step 7), and finally annotated
769 with metrics from dbNSFP (step 8). Final dataset was written to Amazon S3 in Parquet columnar file
770 format for further analysis and interpretation.
771
772 **FIGURE 3. Synonymous variants predicted to impact mRNA structure are constrained in the**
773 **human population.** Population frequency of sSNVs were plotted against the predicted impact on mRNA
774 structure. Synonymous variants that disrupt structure tend to be absent from the gnomAD database, while
775 those with limited impact on structure appear at least once in the gnomAD database. **(A)** Proportion of
776 sSNVs with nonzero gnomAD frequency at each value of the RNA stability metric dMFE. Points with
777 fewer than 2000 positive-MAF sSNVs excluded. Color represents average CFEED value, to highlight the
778 relationship between minimum free energy and edit distance. **(B)** Analogous plot for metric CFEED
779 measuring edge differences between mutant/wildtype centroid structures. Color represents |dMFE|,
780 measuring absolute change in stability. **(C)** Analogous plot for diversity-metric dCD measuring change in
781 structural ensemble diversity due to sSNV. Color is by dMFE measuring change in stability.
782
783 **FIGURE 4. Synonymous CpG transitions are markedly constrained against destabilization of their**
784 **mRNA structures.** Population frequency of sSNV vs. effect on mRNA structure in synonymous CpG
785 transitions was examined. Proportion of synonymous CpG transitions with nonzero MAF at each value of

dMFE were determined for (**A**) CpG>CpA and (**B**) CpG>TpG synonymous mutations. dMFE values with fewer than 75 nonzero-MAF sSNVs are excluded. Color gives average CFEED in each context, ranging from 15 (blue) to 50 (red). Similarly, proportion of synonymous CpG transitions with nonzero MAF at each value of CFEED were determined for (**C**) CpG>CpA sSNVs and (**D**) CpG>TpG sSNVs. Color represents average dMFE and ranges from -0.8 (blue) to 1.85 (red). CFEED values with fewer than 75 nonzero-MAF sSNVs are excluded). Finally, proportion of synonymous CpG transitions with nonzero MAF at each value of dCD (after rounding to nearest integer) were determined for (**E**) CpG>CpA and (**F**) CpG>TpG sSNVs sSNVS. Color represents average dMFE and ranges from -3 (blue) to 4 (red). Rounded dCD values with fewer than 75 nonzero-MAF sSNVs are excluded.

FIGURE 5. SPI score correlates with constraint in synonymous CpG transitions. Variants in the contexts **(A)** CpG>CpA and **(B)** CpG>TpG are divided by SPI score into 20 equal bins and the value P(MAF>0) plotted against the mean of each bin. We also colored by the mean dMFE over each bin. In both contexts the constraint is highest towards negative SPI, i.e. sSNVs for which structural information decreases the predicted probability that MAF > 0.

TABLE 1. **Structural metrics correlate with gnomAD frequency in most REF>ALT contexts.** Correlation between structural metrics (**A**) dMFE, (**B**) CFEED and integer-rounded (**C**) dCD on the one hand, and the quantity P(MAF>0) on the other, over all sSNVs in a given context. The R^2 and p-values are obtained from a weighted least-squares linear regression, with the p-value corresponding to the linear coefficient; a quadratic regression was also performed, but only the p-value was retained. Only context-metric pairs with p-value < 0.005 are included. "Normalized slope" was obtained by dividing slope of regression line by average P(MAF>0) in the context and then multiplying by range covered by metric in

its central 90% of sSNVs. "Depleter" is raw sequence variable that explains largest proportion of structural trend *in this context*, with sign adjusted to correlate negatively with gnomAD frequency. "Depleter R^2" gives proportion of variance explained by Depleter (see *Depleter variables* in **RESULTS** for details).

TABLE 2. **Area under curve for SPI score.** SPI was used to discriminate MAF > 0 using a simple logistic model with 5-fold cross-validation. Table shows area under curve for model, averaged over the 5 training and testing sets.

TABLE 3: Known sSNVs clinically implicated for structural pathogenicity are successfully predicted to be pathogenic by our structural metrics. dbSNP RS number and standardized SNP annotations are provided, along with the genes official gene symbol and disease the sSNV has been associated with. The absolute value of dMFE, CFEED, dCD and SPI are provided, along with the percentile value of that score, computed over each context, in parentheses.

REFERENCES

Picard: a set of tools (in Java) for working with next generation sequencing data in the BAM format.

Abuin JM, Pichel JC, Pena TF, Amigo J. 2016. SparkBWA: Speeding Up the Alignment of High-Throughput DNA Sequencing Data. *PLoS One* **11**: e0155461.

Acharya M, Mookherjee S, Bhattacharjee A, Thakur SK, Bandyopadhyay AK, Sen A, Chakrabarti S, Ray K. 2007. Evaluation of the OPTC gene in primary open angle glaucoma: functional significance of a silent change. *BMC Mol Biol* **8**: 21.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7**: 248-249.

Al-Saif M, Khabar KS. 2012. UU/UA dinucleotide frequency reduction in coding regions results in increased mRNA stability and protein expression. *Mol Ther* **20**: 954-959.

Alfares A, Aloraini T, Subaie LA, Alissa A, Qudsi AA, Alahmad A, Mutairi FA, Alswaid A, Alothaim A, Eyaid W et al. 2018. Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing. *Genet Med* **20**: 1328-1333.

Bartoszewski RA, Jablonsky M, Bartoszewska S, Stevenson L, Dai Q, Kappes J, Collawn JF, Bebok Z. 2010. A synonymous single nucleotide polymorphism in DeltaF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein. *J Biol Chem* **285**: 28741-28748.

Bazzini AA, Del Viso F, Moreno-Mateos MA, Johnstone TG, Vejnar CE, Qin Y, Yao J, Khokha MK, Giraldez AJ. 2016. Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *EMBO J* **35**: 2087-2103.

Bellacosa A, Drohat AC. 2015. Role of base excision repair in maintaining the genetic and epigenetic integrity of CpG sites. *DNA Repair (Amst)* **32**: 33-42.

Bevilacqua PC, Ritchey LE, Su Z, Assmann SM. 2016. Genome-Wide Analysis of RNA Secondary Structure. *Annu Rev Genet* **50**: 235-266.

Brummer A, Hausser J. 2014. MicroRNA binding sites in the coding region of mRNAs: extending the repertoire of post-transcriptional gene regulation. *Bioessays* **36**: 617-626.

Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol* **6**: R75.

Chaney JL, Clark PL. 2015. Roles for Synonymous Codon Usage in Protein Biogenesis. *Annu Rev Biophys* **44**: 143-166.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**: 80-92.

Cohen NM, Kenigsberg E, Tanay A. 2011. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* **145**: 773-786.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**: e1001025.

Deciphering Developmental Disorders Study. 2017. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**: 433-438.

Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, Bazile C, Lambert NJ, Van Nostrand EL, Pratt GA et al. 2018. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell* **70**: 854-867 e859.

Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J Mol Biol* **260**: 649-663.

Duan J, Antezana MA. 2003. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. *J Mol Evol* **57**: 694-701.

868 Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J, Gejman PV. 2003. Synonymous mutations
869 in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum Mol*
870 *Genet* **12**: 205-216.
871 Ellingford JM, Barton S, Bhaskar S, Williams SG, Sergouniotis PI, O'Sullivan J, Lamb JA, Perveen R, Hall G, Newman
872 WG et al. 2016. Whole Genome Sequencing Increases Molecular Diagnostic Yield Compared with Current
873 Diagnostic Testing for Inherited Retinal Disease. *Ophthalmology* **123**: 1143-1150.
874 Fahraeus R, Marin M, Olivares-Illana V. 2016. Whisper mutations: cryptic messages within the genetic code.
875 *Oncogene* **35**: 3753-3759.
876 Faure G, Ogurtsov AY, Shabalina SA, Koonin EV. 2016. Role of mRNA structure in the control of protein folding.
877 *Nucleic Acids Res* **44**: 10898-10911.
878 Fernandez M, Kumagai Y, Standley DM, Sarai A, Mizuguchi K, Ahmad S. 2011. Prediction of dinucleotide-specific
879 RNA-binding sites in proteins. *BMC Bioinformatics* **12 Suppl 13**: S5.
880 Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by
881 exploiting biased substitution patterns. *Bioinformatics* **25**: i54-62.
882 Gelfman S, Wang Q, McSweeney KM, Ren Z, La Carpia F, Halvorsen M, Schoch K, Ratzon F, Heinzen EL, Boland MJ
883 et al. 2017. Annotating pathogenic non-coding variants in genic regions. *Nat Commun* **8**: 236.
884 Gronau I, Arbiza L, Mohammed J, Siepel A. 2013. Inference of natural selection from interspersed genomic
885 elements based on polymorphism and divergence. *Mol Biol Evol* **30**: 1159-1171.
886 Gu W, Zhou T, Wilke CO. 2010. A universal trend of reduced mRNA stability near the translation-initiation site in
887 prokaryotes and eukaryotes. *PLoS Comput Biol* **6**: e1000664.
888 Hamasaki-Katagiri N, Lin BC, Simon J, Hunt RC, Schiller T, Russek-Cohen E, Komar AA, Bar H, Kimchi-Sarfaty C. 2017.
889 The importance of mRNA structure in determining the pathogenicity of synonymous and non-synonymous
890 mutations in haemophilia. *Haemophilia* **23**: e8-e17.

Hanson G, Coller J. 2018. Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol* **19**: 20-30.

Hegde M, Santani A, Mao R, Ferreira-Gonzalez A, Weck KE, Voelkerding KV. 2017. Development and Validation of Clinical Whole-Exome and Whole-Genome Sequencing for Detection of Germline Variants in Inherited Disease. *Arch Pathol Lab Med* **141**: 798-805.

Huang YF, Gulko B, Siepel A. 2017. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* **49**: 618-624.

Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. 2014. Exposing synonymous mutations. *Trends Genet* **30**: 308-321.

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D et al. 2016. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**: 877-885.

Katz L, Burge CB. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res* **13**: 2042-2051.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**: 310-315.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science (80-)* **324**: 255-258.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**: 1073-1081.

Lazrak A, Fu L, Bali V, Bartoszewski R, Rab A, Havasi V, Keiles S, Kappes J, Kumar R, Lefkowitz E et al. 2013. The silent codon change I507-ATC->ATT contributes to the severity of the DeltaF508 CFTR channel dysfunction. *FASEB J* **27**: 4630-4645.

914 Lee M, Roos P, Sharma N, Atalar M, Evans TA, Pellicore MJ, Davis E, Lam AN, Stanley SE, Khalil SE et al. 2017.
915 Systematic Computational Identification of Variants That Activate Exonic and Intronic Cryptic Splice Sites.
916 *Am J Hum Genet* **100**: 751-765.
917 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings
918 BB et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285-291.
919 Li E, Zhang Y. 2014. DNA methylation in mammals. *Cold Spring Harb Perspect Biol* **6**: a019133.
920 Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, Wang X, Sun Z. 2018. Performance evaluation of pathogenicity-
921 computation methods for missense variants. *Nucleic Acids Res* **46**: 7793-7804.
922 Li Q, Qu HQ. 2013. Human coding synonymous single nucleotide polymorphisms at ramp regions of mRNA
923 translation. *PLoS One* **8**: e59706.
924 Liu X, Wu C, Li C, Boerwinkle E. 2016. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations
925 for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**: 235-241.
926 Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA
927 Package 2.0. *Algorithms Mol Biol* **6**: 26.
928 Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters
929 improves prediction of RNA secondary structure. *J Mol Biol* **288**: 911-940.
930 McCarthy C, Carrea A, Diambra L. 2017. Bicodon bias can determine the role of synonymous SNPs in human
931 diseases. *BMC Genomics* **18**: 227.
932 McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary
933 structure. *Biopolymers* **29**: 1105-1119.
934 McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant
935 Effect Predictor. *Genome Biol* **17**: 122.

936 Molinski SV, Gonska T, Huan LJ, Baskin B, Janahi IA, Ray PN, Bear CE. 2014. Genetic, cell biological, and clinical
937 interrogation of the CFTR mutation c.3700 A>G (p.Ile1234Val) informs strategies for future medical
938 intervention. *Genet Med* **16**: 625-632.
939 Morgan MT, Bennett MT, Drohat AC. 2007. Excision of 5-halogenated uracils by human thymine DNA glycosylase.
940 Robust activity for DNA contexts other than CpG. *J Biol Chem* **282**: 27578-27586.
941 Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko L. 2006.
942 Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA
943 secondary structure. *Science (80-)* **314**: 1930-1933.
944 O'Brien AR, Saunders NF, Guo Y, Buske FA, Scott RJ, Bauer DC. 2015. VariantSpark: population scale clustering of
945 genotype information. *BMC Genomics* **16**: 1052.
946 Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev*
947 *Genet* **12**: 32-42.
948 Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR et al.
949 2015. Codon optimality is a major determinant of mRNA stability. *Cell* **160**: 1111-1124.
950 Quang D, Chen Y, Xie X. 2015. DANN: a deep learning approach for annotating the pathogenicity of genetic
951 variants. *Bioinformatics* **31**: 761-763.
952 Quax TE, Claassens NJ, Soll D, van der Oost J. 2015. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell*
953 **59**: 149-161.
954 Ramanouskaya TV, Grinev VV. 2017. The determinants of alternative RNA splicing in human cells. *Mol Genet*
955 *Genomics* **292**: 1175-1195.
956 Reamon-Buettner SM, Sattlegger E, Ciribilli Y, Inga A, Wessel A, Borlak J. 2013. Transcriptional defect of an
957 inherited NKX2-5 haplotype comprising a SNP, a nonsynonymous and a synonymous mutation, associated
958 with human congenital heart disease. *PLoS One* **8**: e83295.

959 Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E et al. 2015.
960 Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation
961 of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.
962 *Genet Med* **17**: 405-424.
963 Rocha EP. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding
964 for translation optimization. *Genome Res* **14**: 2279-2286.
965 Sabi R, Tuller T. 2014. Modelling the efficiency of codon-tRNA interactions based on codon usage bias. *DNA Res*
966 **21**: 511-526.
967 Sato T, Terabe M, Watanabe H, Gojobori T, Hori-Takemoto C, Miura K. 2001. Codon and base biases after the
968 initiation codon of the open reading frames in the Escherichia coli genome and their influence on the
969 translation efficiency. *J Biochem* **129**: 851-860.
970 Sauna ZE, Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease.
971 *Nat Rev Genet* **12**: 683-691.
972 Savisaar R, Hurst LD. 2017. Both Maintenance and Avoidance of RNA-Binding Protein Interactions Constrain
973 Coding Sequence Evolution. *Mol Biol Evol* **34**: 1110-1126.
974 Seffens W, Digby D. 1999. mRNAs have greater negative folding free energies than shuffled or codon choice
975 randomized sequences. *Nucleic Acids Res* **27**: 1578-1584.
976 Shabalina SA, Spiridonov NA, Kashina A. 2013. Sounds of silence: synonymous nucleotides as a key to biological
977 regulation and complexity. *Nucleic Acids Res* **41**: 2073-2094.
978 Shah K, Cheng Y, Hahn B, Bridges R, Bradbury NA, Mueller DM. 2015. Synonymous codon usage affects the
979 expression of wild type and F508del CFTR. *J Mol Biol* **427**: 1464-1479.
980 Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR, Campbell C. 2015. An integrative approach
981 to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**: 1536-
982 1543.

983 Silverman SK. 2008. A forced march across an RNA folding landscape. *Chem Biol* **15**: 211-213.

984 Simhadri VL, Hamasaki-Katagiri N, Lin BC, Hunt R, Jha S, Tseng SC, Wu A, Bentley AA, Zichel R, Lu Q et al. 2017.
985 Single synonymous mutation in factor IX alters protein properties and underlies haemophilia B. *J Med*
986 *Genet* **54**: 338-345.

987 Soukarieh O, Gaildrat P, Hamieh M, Drouet A, Baert-Desurmont S, Frebourg T, Tosi M, Martins A. 2016. Exonic
988 Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico
989 Tools. *PLoS Genet* **12**: e1005756.

990 Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An
991 evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**: 344-
992 354.

993 Turner DH, Mathews DH. 2010. NNDB: the nearest neighbor parameter database for predicting stability of nucleic
994 acid secondary structure. *Nucleic Acids Res* **38**: D280-282.

995 Vaz-Drago R, Custodio N, Carmo-Fonseca M. 2017. Deep intronic mutations and human disease. *Hum Genet* **136**:
996 1093-1111.

997 Verma M, Choi J, Cottrell KA, Lavagnino Z, Thomas EN, Pavlovic-Djuranovic S, Szczesny P, Piston DW, Zaher H,
998 Puglisi JD et al. 2019. Short translational ramp determines efficiency of protein synthesis. *bioRxiv*
999 doi:10.1101/571059: 571059.

1000 Vinson C, Chatterjee R. 2012. CG methylation. *Epigenomics* **4**: 655-663.

1001 Wan Y, Qu K, Ouyang Z, Kertesz M, Li J, Tibshirani R, Makino DL, Nutter RC, Segal E, Chang HY. 2012. Genome-wide
1002 measurement of RNA folding energies. *Mol Cell* **48**: 169-181.

1003 Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput
1004 sequencing data. *Nucleic Acids Res* **38**: e164.

Wiewiorka MS, Messina A, Pacholewska A, Maffioletti S, Gawrysiak P, Okoniewski MJ. 2014. SparkSeq: fast, scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision. *Bioinformatics* **30**: 2652-2653.

Workman C, Krogh A. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res* **27**: 4816-4822.

Worthey EA. 2017. Analysis and Annotation of Whole-Genome or Whole-Exome Sequencing Derived Variants for Clinical Diagnosis. *Curr Protoc Hum Genet* **95**: 9 24 21-29 24 28.

Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, King DA, Ambridge K, Barrett DM, Bayzetinova T et al. 2015. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**: 1305-1314.

Wright CF, FitzPatrick DR, Firth HV. 2018. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet* **19**: 253-268.

Yang JR, Chen X, Zhang J. 2014a. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol* **12**: e1001910.

Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z et al. 2013. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* **369**: 1502-1511.

Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, Ward P, Braxton A, Wang M, Buhay C et al. 2014b. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**: 1870-1879.

Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M, Franklin MJ, Shenker S, Stoica I. 2012. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pp. 2-2. USENIX Association, San Jose, CA.

FIGURE 1

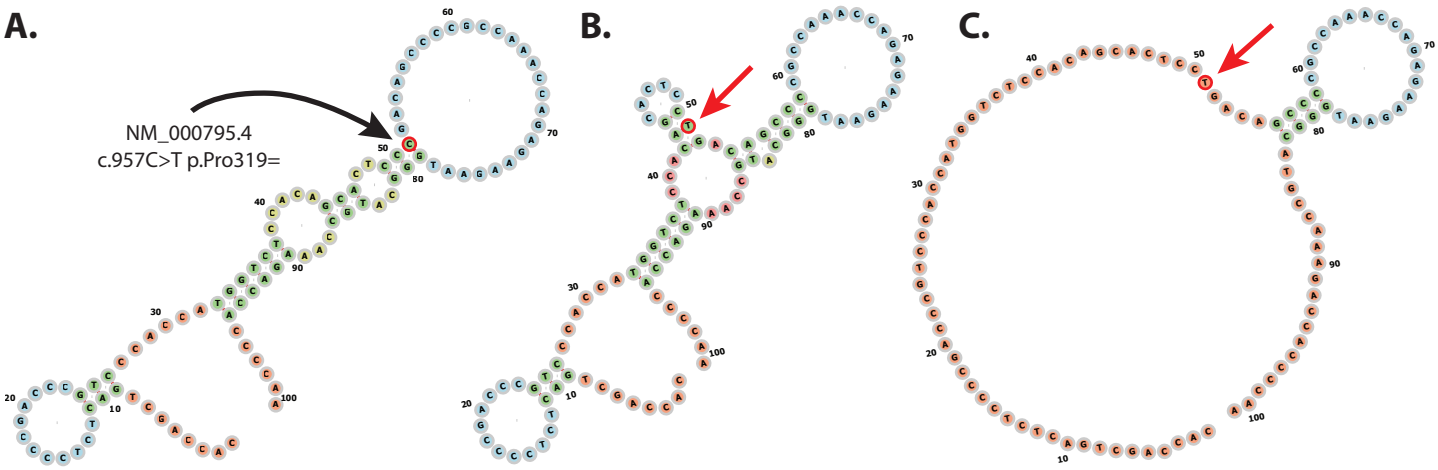

FIGURE 2

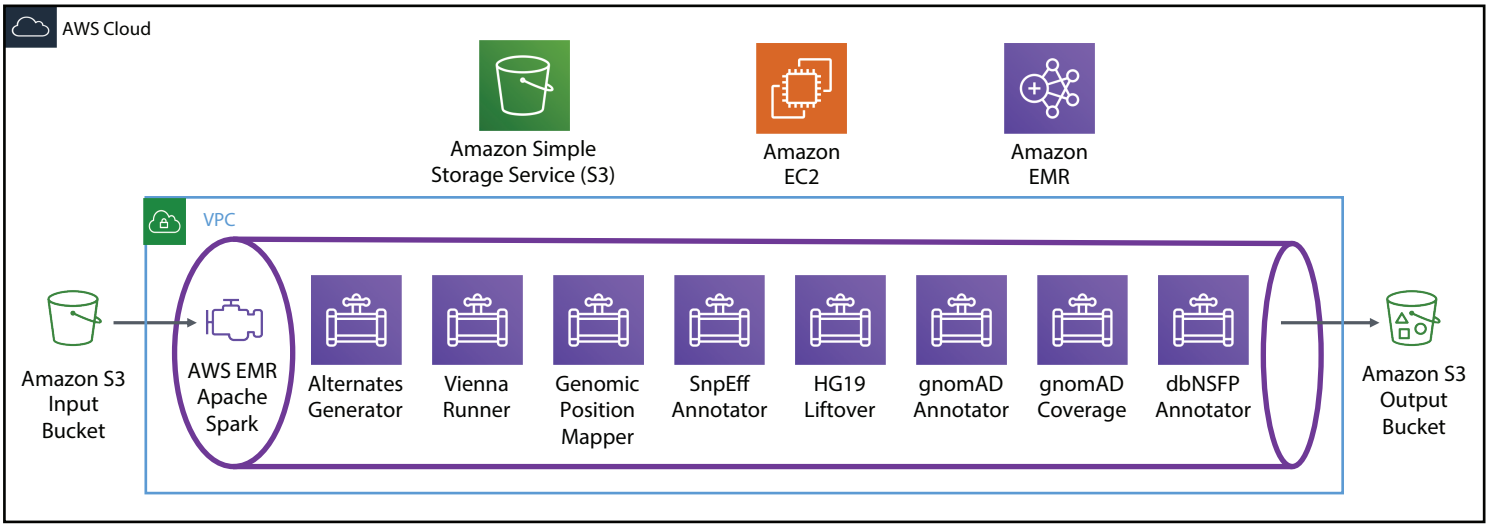

FIGURE 3

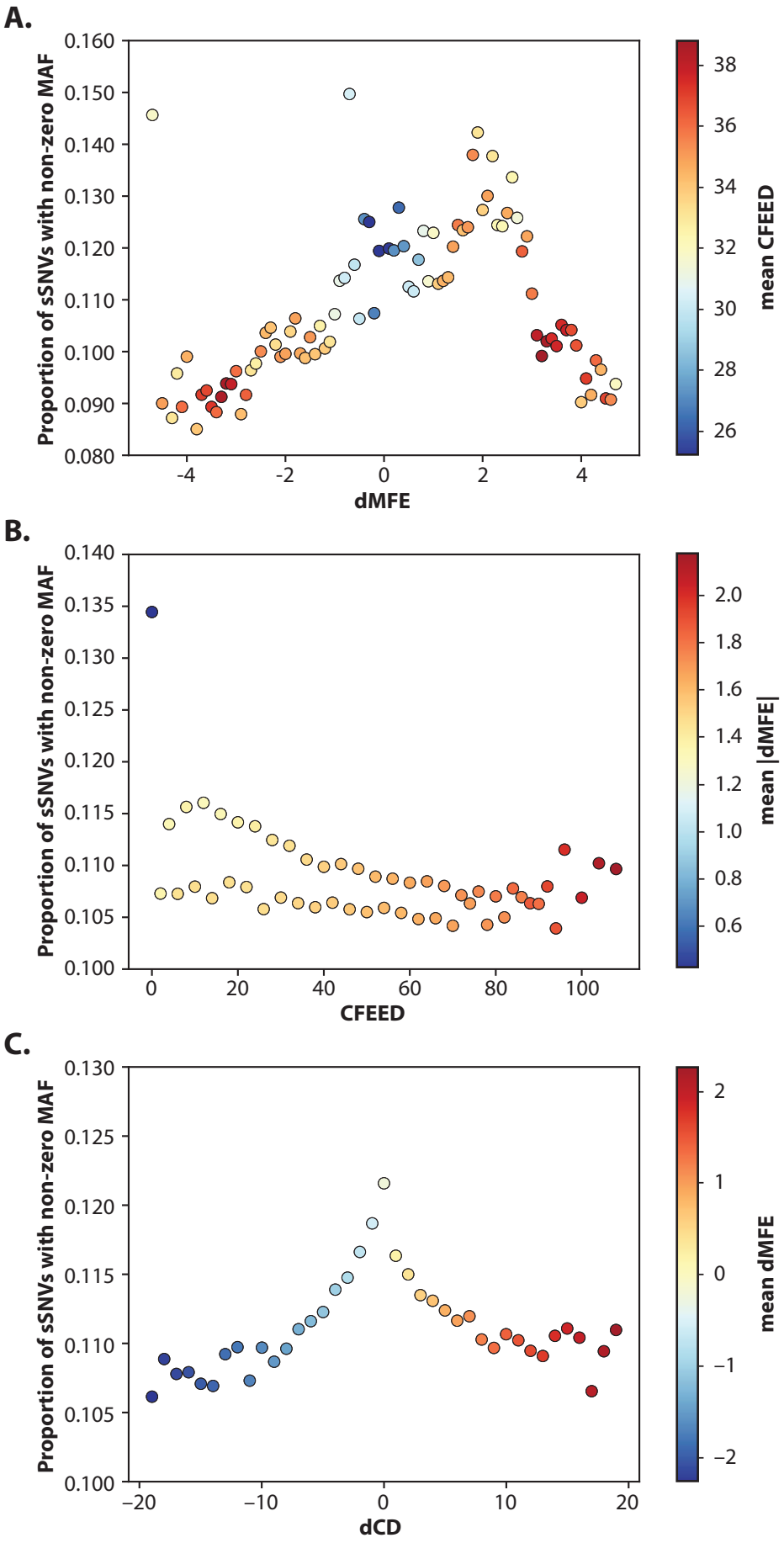

FIGURE 4

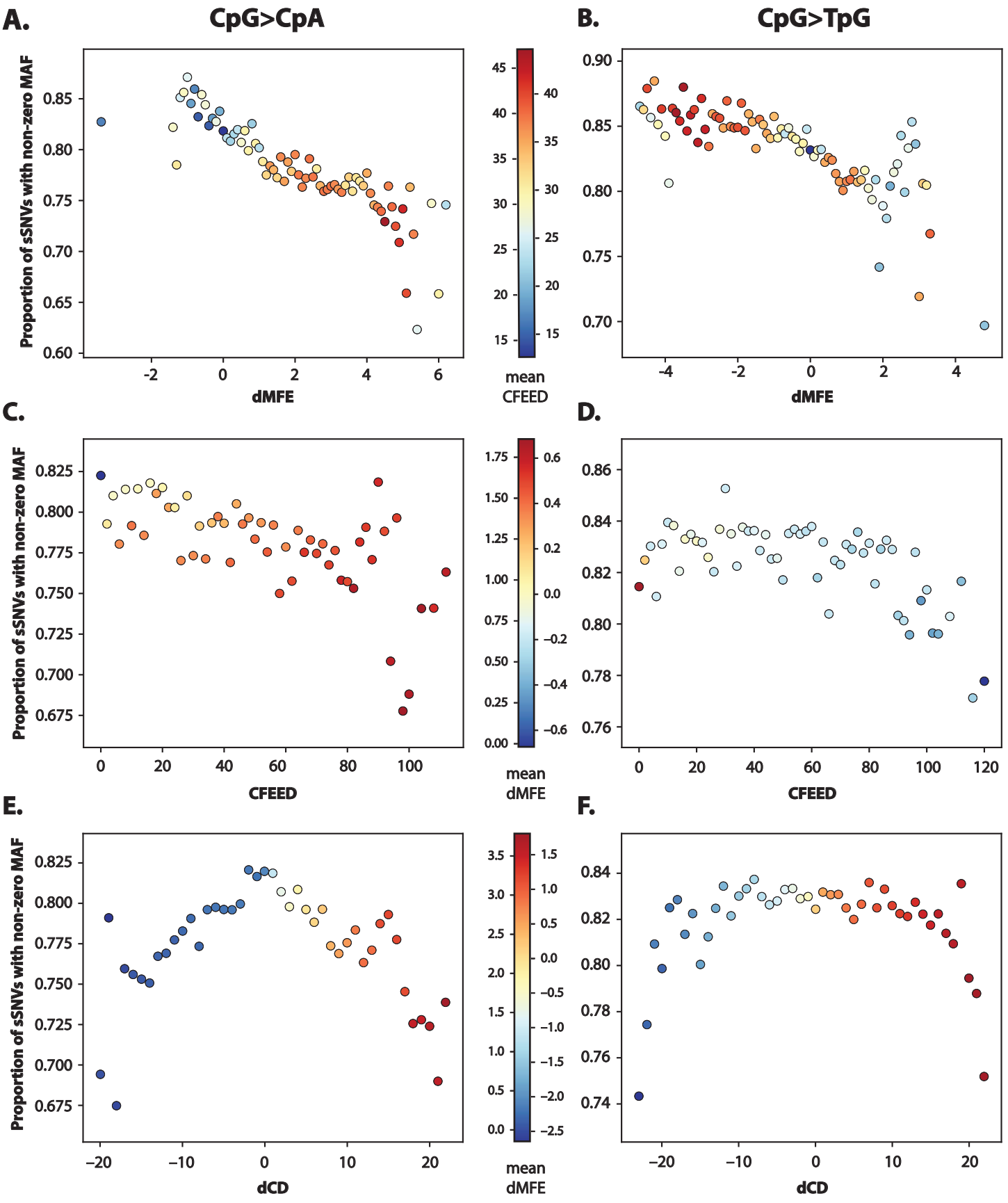

FIGURE 5

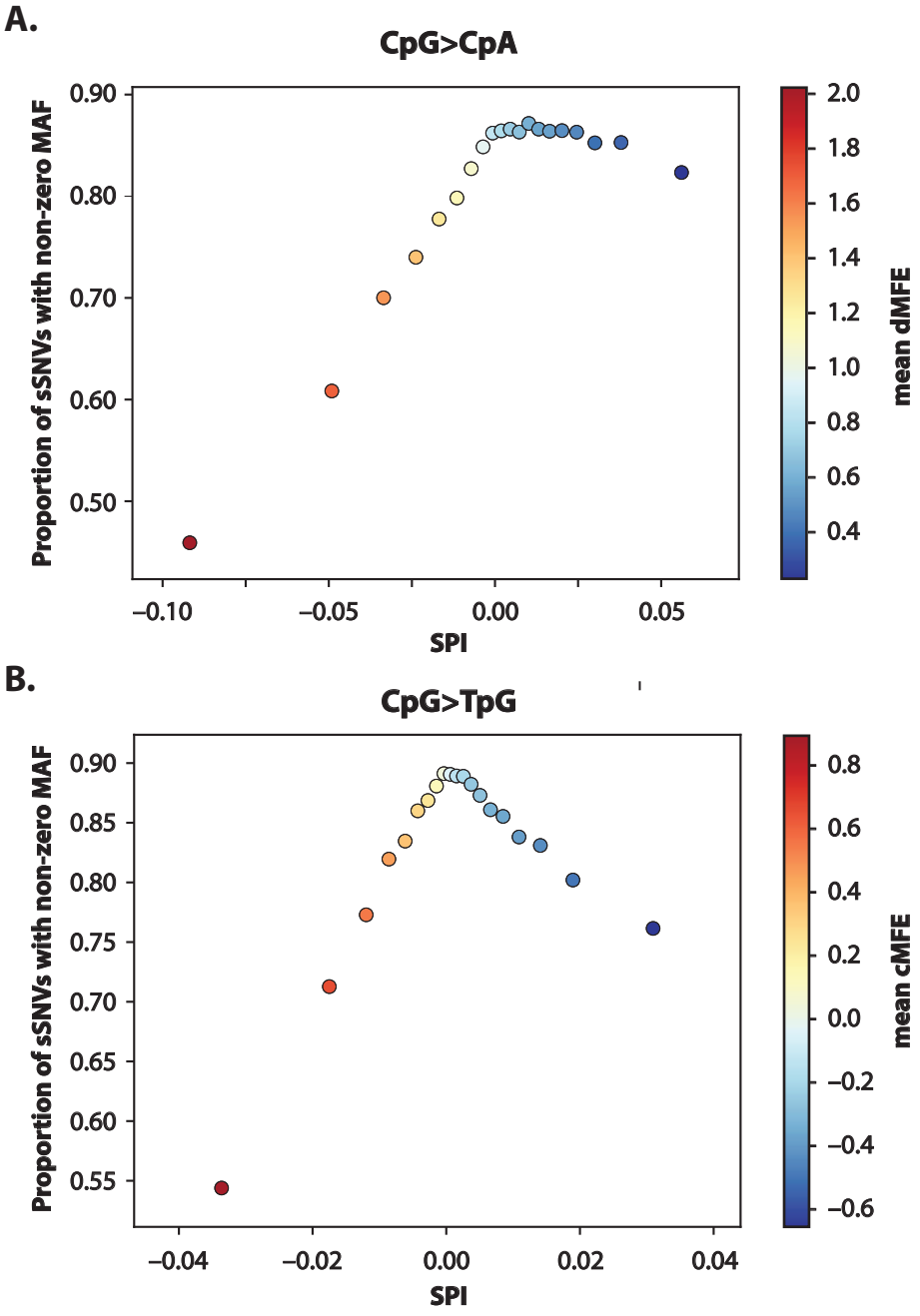

TABLE 1

Context	Normalized slope	R^2	Linear p-value	Quadratic p-value	Depleter	Depleter R^2
(A). dMFE						
CpG>CpA	-0.0861	0.690	1.49e-72	0.26	+CpG content	0.785
CpG>TpG	-0.0553	0.439	5.02e-42	8.39e-06	+CpG content	0.727
G>T	-0.1250	0.165	1.13e-27	0.00168	- leading C	0.317
C>G	-0.1170	0.139	1.35e-27	0.0659	- trailing G	0.141
C>T	-0.0394	0.125	9.61e-21	0.000494	+ leading G	0.159
C>A	-0.0845	0.099	2.9e-19	1.03e-12	- trailing G	0.283
G>A	0.0229	0.034	1.47e-06	0.000138	- leading G	0.221
A>G	0.0377	0.029	7.69e-06	0.34	- trailing T	0.354
T>C	0.0268	0.018	0.000217	1.08e-20	- leading A	0.351
G>C	-0.0435	0.018	0.000169	0.0547	- leading C	0.182
(B). CFEED						
CpG>CpA	-0.0469	0.655	2.28e-17	0.751	+CpG content	0.822
G>A	0.0302	0.338	4.18e-08	0.152	+trailing A	0.557
T>C	-0.0345	0.275	8.2e-07	5.69e-05	- leading A	0.642
C>T	-0.0207	0.244	6.13e-06	1.46e-08	- C content	0.382
C>A	0.0365	0.217	1.96e-05	0.0828	+trailing A	0.335
CpG>TpG	0.0044	0.009	0.207	2.37e-08	+CpG content	0.604
T>A	-0.0020	-0.014	0.887	0.000419	+leading T	0.027
(C). dCD						
CpG>TpG	-0.0012	-0.013	0.638	3.89e-05	+CpG content	0.438
CpG>CpA	-0.0020	-0.016	0.751	1.42e-14	+CpG content	0.825
G>A	0.0003	-0.017	0.947	2.08e-08	+trailing A	0.645

TABLE 2

Context	Mean AUC: Training dataset	Mean AUC: Test dataset
CpG>TpG	0.656	0.656
CpG>CpA	0.616	0.618
T>C	0.565	0.564
C>A	0.550	0.548
A>G	0.539	0.539
G>C	0.536	0.538
G>T	0.534	0.532
T>A	0.540	0.532
G>A	0.522	0.524
T>G	0.529	0.522
A>C	0.521	0.516
C>G	0.513	0.509
C>T	0.506	0.506
A>T	0.512	0.505

TABLE 3.

Gene	Condition	SNP (GRCh37)	Context	|dMFE|	CFEED	|dCD|	|SPI|
OPTC	Primary open angle glaucoma	rs559635109 NC_000001.10:g.203467924C>T NM_014359.3:c.486C>T NP_055174.1:p.Phe162=	C>T	0.0 (26.6)	32 (69.8)	2.53 (52.5)	**0.0261** **(87.2)**
NKX2-5	Congenital heart disease	rs72554028 NC_000005.9:g.172660004C>T NM_004387.4:c.543G>A NP_004378.1:p.Gln181=	G>A	**3.5** **(89.1)**	4 (30.0)	0.24 (10.6)	0.0098 (52.9)
NKX2-5	Congenital heart disease	rs2277923 NC_000005.9:g.172662024T>C NM_004387.3:c.63A>G NP_004378.1:p.Glu21=	A>G	0.0 (22.2)	20 (57.8)	**8.85** **(89.9)**	0.0153 (47.3)
DRD2	Schizophrenia, substance abuse	rs6277 NC_000011.9:g.113283459G>A NM_000795.4:c.957C>T NP_000786.1:p.Pro319=	CpG>TpG	1.0 (52.4)	**60** **(86.4)**	**9.42** **(87.4)**	0.0010 (65.7)
COMT	Pain sensitivity	rs4633 NC_000022.10:g.19950235C>T NM_000754.3:c.186C>T NP_000745.1:p.His62=	CpG>TpG	0.5 (35.9)	**66** **(88.9)**	3.80 (58.6)	0.0043 (31.5)
COMT	Pain sensitivity	rs4818 NC_000022.10:g.19951207C>G NM_000754.3:c.408C>G NP_000745.1:p.Leu136=	C>G	**3.0** **(82.4)**	38 (60.3)	6.83 (72.2)	0.0027 (10.4)