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

ABSTRACT

Background. In most organisms the structure of an mRNA molecule is crucial in determining speed of translation, half-life, splicing propensities and final protein configuration. Synonymous variants 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 ViennaRNA 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.

Conclusions. Our findings support the hypothesis that sSNVs may play a role in genetic disorders due to their effects on mRNA structure. Our RNA-folding scores provide a means of gauging the structural constraint operating on any sSNV in the human genome. Given that the majority of patients with rare or as yet to be diagnosed disease lack a molecular diagnosis, these scores have the potential to enable discovery of novel genetic etiologies. Our RNA Stability Pipeline as well as ViennaRNA structural metrics and SPI scores for all human synonymous variants can be downloaded from GitHub https://github.com/nch-igm/rna-stability.

42 **INTRODUCTION**

43 While next generation sequencing (**NGS**) has accelerated the discovery of new functional variants
44 in syndromic and rare monogenic diseases, many more disease-causing genes and novel genetic etiologies
45 remain to be discovered [1, 2]. Accurate molecular genetic diagnosis of a rare disease is essential for
46 patient care [3], yet today's best molecular tests and analysis strategies leave 60-75% of patients
47 undiagnosed [4-8]. Current clinical practice for sequence variant interpretation focuses primarily on
48 missense, nonsense or canonical splice variants [9], with numerous bioinformatics prediction algorithms
49 and databases developed for functional prediction and annotation of non-synonymous single-nucleotide
50 variants (nsSNVs) that impact protein function through changes in the underlying coding sequence [10].
51 However, these algorithms are inadequate to infer pathogenicity in non-protein-altering variants such as
52 intronic or synonymous variants, which are under different and weaker evolutionary constraints [11].
53 While the potentially pathogenic impact of non-synonymous single nucleotide variants (**nsSNVs**) that
54 change the protein sequence are well understood, we have limited knowledge in regard to the role that
55 synonymous SNVs (**sSNVs**) may have in human health and disease.

56 Synonymous variants result in codon changes that do not alter the amino acid sequence of the
57 translated protein and as such for decades were referred to as "silent." However, there is a growing body
58 of evidence demonstrating that synonymous codons have vital regulatory roles [12-16] among the most
59 important of which is their contribution to RNA structure.

60 Messenger RNA (**mRNA**) is a single-stranded molecule that adopts three levels of structure: the
61 *primary sequence* forms base pairs among its own nucleotides to build the *secondary structure*, which
62 further folds through covalent attractions to form the *tertiary structure* (**FIGURE 1**) [17]. While the tertiary
63 structure of mRNA is challenging to model and poorly understood, sophisticated tools exist to compute

64 the ensemble of possible secondary structures and determine the optimal structure for a given mRNA
65 strand [18].

66 Studies first published in 1999 indicated that stable mRNA secondary structures are often selected
67 for in key genomic regions across all kingdoms of life [19-22]. Synonymous variants impacting RNA
68 structure can alter global RNA stability, where stable mRNAs tend to have longer half-lives and less stable
69 RNA molecules may be more rapidly degraded resulting in lower protein levels [23-28]. The stability of
70 an mRNA transcript affects translational initiation and can determine how quickly a given protein is
71 translated [19-21, 29-31]. Recent studies strongly linked mRNA structure to protein confirmation and
72 function, with synonymous codons acting as a subliminal code for the protein folding process [16, 30, 32-
73 35]. mRNA structure can also facilitate or prevent miRNAs and RNA-binding proteins from attaching to
74 specific structural motifs [36-39]. Given these multiple mechanisms, when synonymous variants are
75 ignored, we are almost certainly missing novel plausible explanations for genetic disease.

76 The role of mRNA structure in human health and disease, however, is poorly comprehended and
77 relatively few pathogenic variants impacting mRNA folding have been described [23, 24, 26, 28]. A
78 structure-altering sSNV in the dopamine receptor DRD2 was shown to inhibit protein synthesis and
79 accelerate mRNA degradation [40]. A sSNV in the *COMT* gene, implicated in cognitive impairment and
80 pain sensitivity, was shown *in vitro* to constrain enzymatic activity and protein expression [41]. A sSNV
81 discovered in the OPTC gene of a glaucoma patient resulted in decreased protein expression *in vivo* [42].
82 In cystic fibrosis patients, a sSNV in the *CFTR* gene was linked to decreased gene expression [43].
83 Additionally, a silent codon change, I507-ATC→ATT, contributes to CFTR dysfunction by a change in
84 mRNA secondary structure that alters the dynamics of translation leading to misfolding of the CFTR
85 protein [25, 27]. Two sSNVs in the NKX2-5 gene, identified in patients with congenital heart disease,
86 decreased the mRNA's transactivation potential in a yeast-based assay [44]. In hemophilia B, the sSNV

87 c.459G>A in factor IX impacts the transcript's secondary structure and reduces extracellular protein levels
88 [45], and both synonymous and nonsynonymous SNVs were shown more likely be deleterious when
89 occurring in a stable region of mRNA in hemophilia associated genes F8 and Duchenne's Muscular
90 Dystrophy [46].

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

100 **RESULTS**

101 ***Massively parallel generation of RNA stability metrics***

102 Global assessment of sSNVs is truly a big data problem as it requires generation and evaluation of
103 several raw values for each of hundreds of millions of positions within the genome. To address this
104 challenge and successfully predict the mRNA-structural effects of every possible sSNV, we developed
105 novel software built upon the Apache Spark framework (**FIGURE 2**). Apache Spark is a distributed, open
106 source compute engine that drastically reduces the bottleneck of disk I/O by processing its data in memory
107 whenever possible [47]. This leads to a 100x increase in speed and allows for more flexible software
108 design than can be achieved in the traditional Hadoop MapReduce paradigm. Spark is well suited to
109 address many of the challenges faced in analyzing big genomics data in a highly scalable manner and
110 adoption is growing steadily, with applications such as SparkSeq [48] for general processing, SparkBWA
111 [49] for alignment and VariantSpark for variant clustering [50]. By developing a solution within this
112 framework, we eliminate significant computational hurdles standing in the way of large-scale analysis of
113 sSNVs.

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

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

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

141 To test whether certain sSNVs are under constraint due to their effect on mRNA structure, RNA
142 folding metrics from our ViennaRNA pipeline were combined with population frequencies from the
143 Genome Aggregation Database (gnomAD), containing aggregate WGS and WES data from a total of
144 138,632 unrelated human individuals [51]. Our expectation was that SNVs with disruptive structural
145 properties would be found less frequently in human population. Constrained variants were defined as those

146 absent from gnomAD, versus un-constrained variants as being those with exon minor allele frequency
147 (**MAF**) > 0, a strategy similar to that employed by other groups [52, 53].
148
149 ***Global constraint to maintain stability***
150 Our study reveals a striking connection between a given SNV's impact on mRNA structure and its
151 frequency in the gnomAD database. We define the central variable Y to be Y=1 when a SNV is present
152 in gnomAD and Y=0 when the SNV is absent. We hypothesize that variants which disrupt mRNA
153 structure should have Y=0 (i.e. be absent from the gnomAD database), while those with limited impact
154 on structure should tend to have Y=1 (i.e. appear at least once in the gnomAD database).
155 This central hypothesis is validated in **FIGURE 3**, which depicts the proportion of SNVs with Y=1
156 at every value of our stability-metric dMFE. All four variant classifications – synonymous, 5 prime UTR,
157 3 prime UTR and missense – show a bi-directional constraint to maintain the wild-type mRNA structure.
158 In each context the bell-shaped pattern indicates that a SNV is most likely to appear in the population
159 when it maintains the existing level of stability i.e. has dMFE close to 0. When the SNV either over-
160 stabilizes the mRNA (low dMFE) or de-stabilizes it (high dMFE) the SNV is depleted in the population
161 roughly in proportion to the level of disruption. While this pattern of constraint was observed across all
162 four variant classes, **FIGURE 3** indicates that synonymous variants experience the most extreme constraint
163 for mRNA structure. This justifies our decision to regard the synonymous case as prototypical, and
164 henceforth we restrict our attention to it.
165 **FIGURE 4** depicts the effect of mRNA-structural changes for sSNVs. **FIGURE 4A** (also depicted
166 by green circles in **FIGURE 3**) shows the correlation between Y and the stability metric dMFE across all
167 possible sSNVs. The bell-shaped distribution shows that any change to mRNA stability bidirectionally
168 decreases the chance of the SNV's appearing in a human population.

169 **FIGURE 4B** shows an analogous plot for the structural disruption metric CFEED (see
170 **SUPPLEMENTARY FIGURE 2** for an illustration of how CFEED is calculated). This plot appears to depict
171 two separate trends, but actually shows a single pattern that alternates between high and low on successive
172 values: the sSNVs with CFEED=0,4,8,12... are enriched over those with CFEED=2,6,10,14... (CFEED
173 can only take on even values because the destruction/creation of a base pair always requires two edits).
174 One possible explanation for this duality is that when CFEED fails to be divisible by 4, there is necessarily
175 a change in the total number of base-pairings in the mRNA centroid structure. Thus, sSNVs which
176 conserve the total number of base-pairs could be potentially favored. **FIGURE 4B** also supports the
177 hypothesis that structurally disruptive sSNVs should appear less frequently in the population. We see that
178 sSNVs which leave the centroid structure unchanged (i.e. CFEED=0) are roughly 15% more common
179 than those sSNVs predicted to alter it. And within each of the two separate trends (that is, the multiples
180 and non-multiples of 4) the population frequency declines as the number of centroid base-pairing changes
181 grows from small to large.

182 Finally, sSNVs which either diversify the ensemble of mRNA structures (high dCD) or
183 homogenize it (low dCD) are depleted in the population proportionately to their disruptions, as shown in
184 **FIGURE 4C**. The symmetry in depletion between over- and under-diversifying sSNVs is surprisingly
185 regular. (Here and throughout, dCD values are rounded to the nearest integer, a step necessitated by the
186 fact that the dCD metric does not cluster into just a few values like the other two.)

187 The relationship between the three metrics is illuminated by color-coding in **FIGURE 4.** We observe
188 in **FIGURES 4A AND 4B** that disruptions in the magnitude of stability (|dMFE|) and base-pairing (CFEED)
189 of a sSNV are markedly correlated, with the two metrics enriched for each other at extreme values (red
190 coloring). **FIGURE 4C** depicts a clear relationship between diversity and stability, with those sSNVs
191 diversifying the ensemble (high dCD) also tending to de-stabilize it (red). This diversity-instability

192 relationship is intuitive, as a destabilizing mutation "frees up" portions of the mRNA to assume new forms.
193 Together, these observations validate the central hypothesis that sSNVs which disrupt mRNA structure
194 should be constrained in human populations.
195
196 ***Variation of constraint with REF>ALT context***
197 An mRNA's secondary structure is largely determined by its primary structure (i.e. by the
198 sequence of nucleotides). We therefore expect the constraint in **FIGURE 4** to be partially mediated by
199 the sequence features around each sSNV, the most immediate of which are its reference and alternate
200 bases. Accordingly we divide our sSNVs into 14 classes (**TABLE 1):** 12 classes based on their reference
201 and alternate alleles (e.g. A>C, C>G, T>C, etc.) and 2 additional classes based on potential loss of
202 methylated cytosine (CpG>TpG or CpG>CpA, the latter of which results from a deamination on an
203 antisense strand). Then within each REF>ALT context we reconstruct the three plots of **FIGURE 4** and
204 also perform weighted linear and quadratic regressions between the three different stability metrics and
205 Y=1 (see **METHODS** for details). All significant results ($p < 0.005$) of this procedure appear in **TABLE**
206 **1**.
207 Looking at **TABLE 1A** (which shows the results for dMFE) we find that disruptions to mRNA
208 stability are constrained across many of our sSNV classes. The fact that most of linear p-values are much
209 smaller than the quadratic p-values indicates that in most contexts the dMFE-Y relationship is *linear,* in
210 contrast to the bell-shaped relationship we see when considering global dMFE (**FIGURE 4A**). Therefore,
211 the slope of the regression line indicates which direction of dMFE is enriched for Y=1. For example, in
212 the context of G>T the negative normalized slope indicates that lower dMFE values (i.e. stabilizing) are
213 less constrained (i.e. Y=1). The slope of the regression line (and the relationships it models) proves to
214 depend largely on whether a context's REF and ALT nucleotides are "strong" (C,G) or "weak" (A,T)

215 binders. We note from **TABLE 1A** that strong>weak mutations consistently have negative slopes (except
216 in the irregular context G>A; see *Constraint for mRNA stability in non-CpG-transitional contexts*), while
217 the two weak>strong contexts A>G and T>C have positive slopes.

218 In **TABLE 1B** we observe the constraint for the structural disruption metric CFEED. The results
219 here are surprising – the contexts are split between positive and negative slopes. In support of our
220 hypothesis, four of the sequence contexts display a negative slope, implying that sSNVs with high CFEED
221 values are constrained. However, in contrast to our hypothesis, three of the sequence contexts have a
222 positive slope, which implies that sSNVs in these contexts with high CFEED values are enriched. In the
223 case of CpG>TpG mutations the low quadratic p-value indicates that the pattern is actually bell-shaped,
224 with both low and high CFEED values being depleted; but in G>A and C>A contexts the quadratic term
225 is not significant. Actual plots of these patterns reveal that the ones in which CFEED is depleted are more
226 striking (see **FIGURE 5** and **SUPPLEMENTARY FIGURES 4-5** for plots of Y vs. CFEED in all stability-
227 significant contexts), but this unexpected result must still be addressed. We speak more on this topic in
228 the **DISCUSSION.**

229 Finally, **TABLE 1C** shows mutation contexts that are significantly constrained against changes to
230 ensemble diversity. We see that only a few contexts experience this constraint. But when significant, the
231 constraint for diversity appears to inherit the bidirectionality of **FIGURE 4C** (with the quadratic term being
232 the most significant and the linear fit being very poor). In these contexts, decreases and increases to
233 ensemble diversity appear to be equally harmful.

234

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

236 The data in **TABLE 1** highlight that our observed constraint for mRNA structure is the greatest
237 when considering CpG transitions. Since these variants (and their suppression) are crucial to the story of

238 mRNA stability, it is important to have an appreciation of their role in a biochemical context. The
239 dinucleotide CG (usually denoted CpG to distinguish this linear sequence from the CG base-pairing of
240 cytosine and guanine) is capable of becoming methylated and then mutating by a process called
241 "deamination" into a TG dinucleotide. While studies have demonstrated that methylated CpG residues are
242 up to 40X times more likely to be deaminated than their unmethylated counterparts [54], mechanisms
243 exist to enzymatically repair CpG deaminations [55, 56]. In mammals 70-80% of CpGs are methylated,
244 which makes a CpG transition almost 5x more common than any other mutation-type among mammals
245 (see **SUPPLEMENTARY DATA TABLE 3**) [57]. Possible explanations for the distribution and retention of
246 CpGs in mammals have been extensively debated, with some arguing that there is no pattern of selective
247 constraint on CpGs when compared with other dinucleotides within CpG islands [58].

248 The nucleotides C and G also form the foundation of mRNA secondary structures. Most of the
249 energy of an mRNA structure lies in its "stacks" of nucleotides with the average energy of a C-G pair in
250 a stack around 65% stronger than that of any other base-pairing [59]. Moreover, the self-complementarity
251 of CpGs means that upstream and downstream instances can bind together and form a four-base stack
252 which other base-pairs can then build around.

253 In the present study we find strong evidence that CpG transitions are constrained against de-
254 stabilization of their mRNA structures. This striking trend is largely explained (in a statistical sense) by
255 CpG content, i.e. number of CpG dinucleotides in the surrounding 120 nucleotides of the mRNA transcript
256 (see "Depleter" R^2 in **TABLE 1**). We distinguish CpG>CpA versus CpG>TpG transitions (the former of
257 these usually results from a CpG>TpG deamination on an anti-sense DNA strand), as these two mutation-
258 types show a qualitatively different constraint for mRNA structure. **FIGURE 5** shows the performance of
259 our three main metrics in CpG-transitional contexts. Most strikingly, we find that synonymous CpG>CpA
260 and CpG>TpG mutations both show a steady constraint against de-stabilization (high dMFE) (**FIGURES**

261 **5A & 5B**). Fascinatingly, both contexts exhibit a cluster of outliers in the most destructive (i.e. most de-
262 stabilizing region), suggestive of extreme constraint borne of significant structural disruption. Though the
263 two plots exhibit the same basic shape, the context CpG>CpA of **FIGURE 5A** shows higher de-stabilizing
264 tendencies (higher dMFE values) and also a stronger constraint (lower P(Y=1)).

265 The behavior of the edge metric CFEED in these contexts is less clear-cut. In **FIGURE 5**C we see
266 a clear pattern of constraint against mutations with high CFEED values; and the red coloring shows that
267 such changes are, on average, de-stabilizing. But the constraint in the context CpG>TpG (**FIGURE 5D)** is
268 much less forceful (in fact, its quadratic p-value is much smaller than its linear) and the blue coloring by
269 dMFE shows such mutations are on average neutral or even de-stabilizing. Finally, **FIGURES 5E & 5F**
270 show that the basic pattern of constraint for diversity in **FIGURE 4C** is reproduced and is essentially
271 unchanged for both types of CpG transition. The coloring again indicates that mutations CpG>CpA are
272 much more destabilizing than their CpG>TpG counterparts.

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

278

279 ***Constraint for mRNA stability in non-CpG-transitional contexts***

280 We see the strongest constraint for mRNA structure in CpG transitions, but we observe an
281 analogous pattern in most REF>ALT contexts (as indicated by **TABLE 1**). We can classify these remaining
282 contexts based on whether their slopes in **TABLE 1A** are positive or negative. **SUPPLEMENTARY FIGURE**
283 **4** shows plots of contexts where dMFE and the gnomAD variable Y are negatively correlated. In such

284 contexts the data are consistent with the hypothesis that sSNVs which de-stabilize mRNA are constrained.
285 Notably, all these contexts are strong>weak (or strong>strong in the case of C>G), consistent with the
286 principle that one purpose of such nucleotides is to maintain stability. The coloring by CFEED indicates
287 that a change in either direction is likely to alter the mRNA secondary structures.

288 In **SUPPLEMENTARY FIGURE 5** we show the contexts where dMFE and Y vary positively, which
289 amounts to the claim that stabilizing mutations are constrained in these contexts. Correspondingly, we
290 note that two out of three of these contexts are weak>strong (and the third is the unusual context G>A
291 where SNPs that alter stability or diversity are actually enriched). The context T>C exhibits a notable
292 constraint in either direction, an anomaly which we speculate on in the **DISCUSSION.**

293

294 ***Depleter variables***

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

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

304 To determine which variable is most informative (and should therefore be called the Depleter) we
305 compute an associated R^2 for a set of features of the sequence around the sSNV (the upstream/downstream
306 nucleotides and the proportion of A, C, G, T, CpG or ApT [di]nucleotides in the surrounding 120 bases).

307 Each of these features is used to build a simple logistic model to predict Y=1, and the predictions of the
308 model are then compared to the actual proportion P(Y=1) at a value of the metric. For example, building
309 a CpG-context-based model allows us to compute the quantity P(Y=1 | CpG content), and then we consider
310 the difference:

311 $$\mathbf{E}(\ \mathbf{P}(\mathrm{Y} = 1 \mid \mathrm{CpG\ content}) \mid \mathrm{dMFE}) \ - \mathbf{P}(\mathrm{Y} = 1 \mid \mathrm{dMFE}\ )$$

312 Squaring this difference and taking a weighted sum over all values of dMFE in a context, we recover the
313 variance left unexplained by a particular non-structural variable. We obtain an R^2 by comparing
314 unexplained variance to that obtained using a null model, and the Depleter is then the variable with the
315 largest R^2 (see **METHODS** for more details). **TABLE 1** shows that Depleters can recover large portions of
316 the trends in **FIGURE 5** and **SUPPLEMENTARY FIGURES 4-5.** The striking trend between dMFE and
317 gnomAD frequency in CpG-transitional contexts is largely driven by the proportion of CpGs in the
318 surrounding 120 nucleotides (73% for CpG>TpG sSNVs and 79% for CpG>CpA sSNVs). CpG content
319 is also the most powerful feature when accounting for the behavior of CFEED and dCD in these contexts,
320 with high CpG content consistently correlating with depletion. The natural inference is that an abundance
321 of CpGs signifies important mRNA structure nearby, the disruption of which could be deleterious.

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

328

329 ***Global quantification of mRNA constraint***

330 Our analysis shows that variants predicted to influence mRNA secondary structures are
331 constrained in the population. However, due to the multiple facets that need to be considered when
332 studying RNA secondary structure, by focusing on a single RNA-folding metric such as dMFE or CFEED,
333 we run the risk of missing functionally relevant information. To overcome this potential limitation of our
334 RNA folding metrics, we set out to devise a more diversified method for predicting possible pathogenicity
335 due to mRNA structure. Our strategy is to consider the additional statistical power bestowed by mRNA
336 structure. In each of our 14 sequence contexts from **TABLE 1** we build two general logistic models for
337 predicting MAF >0: a null model that uses the natural variables of sequence context, local nucleotide
338 composition, transcript position and tRNA propensity, but NOT mRNA structure (***n***); and a structural
339 model which also includes the 10 metrics obtained from our ViennaRNA analysis (***s***). These models yield
340 two separate probability-predictions P_n and P_s for the quantity P(MAF >0) (see **METHODS** for details).
341 Then we define the metric:

342 $$\mathrm{SPI} = \log_{10}\left(\frac{P_s}{P_n}\right)$$

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

351 The behavior of SPI in non-CpG-transitional contexts is less regular and harder to weave into a
352 coherent story. Every context shows a clear pattern, but this may amount to either enrichment or depletion

353 (or both) as SPI moves in either direction. Given the strong dependence on REF-ALT context, the use of
354 SPI as a deleteriousness score in non-CpG contexts may need further evaluation.
355
356 ***Clinical Examples of Structural Pathogenicity***
357 The literature reveals only a few examples of synonymous sSNVs unequivocally shown to be
358 pathogenic through their effects on mRNA structure. These sSNVs, with accompanying values of our
359 three ViennaRNA metrics and SPI, are listed in **TABLE 3.** The sSNVs show a definite enrichment for our
360 structural metrics as each shows a value of |dMFE|, CFEED, |dCD| or |SPI| that is in at least the 80th
361 percentile in its context. For example, the pathogenic sSNV in NKX2-5, linked to congenital heart disease,
362 has a dCD score in the 90th percentile [44]. It should be noted that none of these clinical sSNVs qualify as
363 truly exceptional outliers for any of our ViennaRNA metrics or SPI; all have scores below the 95th
364 percentile for |dMFE|, CFEED, |dCD| or |SPI| (see **DISCUSSION** for suggested score cutoff values).
365

366 **DISCUSSION**

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

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

386 Our data also include some patterns which can be elegantly explained through mRNA structure.
387 For example, when considering CFEED in **FIGURE 4B** we observed that sSNVs were enriched when their
388 CFEED values were multiples of 4. Since a CFEED value being divisible by 4 is a necessary condition to

389 preserve the total number of base-pairs, this observed enrichment suggests changes to base pairing are
390 constrained. We also observed bi-directional constraint for dMFE in the context T>C, visible in
391 **SUPPLEMENTARY FIGURE 5** and also inferable from the low quadratic p-value in **TABLE 1**. We conjecture
392 the dual constraint in this context might be due to guanine's unique ability to wobble base-pair. Wobble
393 base-pairing occurs between two nucleotides such as guanine-uracil (G-U), that are not canonical Watson-
394 Crick base pairs, but have comparable thermodynamic stabilities. Thus, the dual constraint from mutations
395 T>C could be related to the transformation of T=G wobble base-pairs into stronger C=G Watson-Crick
396 base pairs.

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

402 Taken together, this evidence provides significant support for the hypothesis that disruptions to
403 mRNA structure are directly under constraint. However, we also realize there are other molecular
404 mechanisms at work in the regulation of the transcriptional and translational processes. For example, while
405 the retention of CpG dinucleotides is certainly connected to mRNA structure, other factors such as
406 regulatory methylation, tRNA binding, binding of miRNAs and other RNA binding proteins, DNA
407 chromatin structure and epigenetic modifications in the ORF could also be involved. Relatedly, we found
408 a few contexts where sSNVs which disrupt mRNA structure actually have higher population frequencies
409 than those that do not (**TABLE 1)**, the main example being the enrichment of high CFEED values in the
410 contexts C>A and G>A. In both cases, the Depleter variable is a trailing A, which correlates negatively

411 with gnomAD frequency and CFEED. The presence of an A is likely to have minimal effect on mRNA
412 structure, suggesting that in these contexts the connection is partly spurious.
413
414 ***Successful identification of structurally disruptive sSNVs in known pathogenic synonymous variants***
415 Over the last decade numerous studies have demonstrated that synonymous variants play essential
416 molecular roles in regulating both mRNA structure and processing, including regulation of protein
417 expression, folding and function [reviewed in 12, 60, 61]. However, the potential for pathogenic
418 synonymous variants that impact RNA folding in human genetic disease remains largely unknown.
419 Current American College of Medical Genetics (**ACMG**) guidelines for the assessment of clinically
420 relevant genetic variants focus primarily on missense, nonsense or canonical splice variants [9]. These
421 guidelines suggest that synonymous "silent" variants should be classified as likely benign, if the nucleotide
422 position is not conserved and splicing assessment algorithms predict neither an impact to a splice
423 consensus sequence nor the creation of a new alternate splice consensus sequence. In the absence of
424 functional tools that would aid in the simultaneous assessment of both nsSNVs and sSNVs in a given
425 patient's genome, we are almost certainly missing novel disease etiologies that have their molecular
426 underpinnings in pathological alterations to mRNA structure.
427 Numerous *in silico* tools exist to aid in the prediction of disease-causing missense variants, and
428 have accuracy in the 65%-85% range when evaluating known pathogenic variants [62]. Such algorithms
429 infer pathogenicity based on amino-acid substitutions (SIFT [63], PolyPhen [64], FATHMM [65]),
430 nucleotide conservation (SiPhy [66], GERP++ [67]) or an ensemble of annotations and scores (CADD
431 [68], DANN [69], REVEL [70]). These tools predict whether a nsSNV is pathogenic or benign, primarily
432 due to the high conservation of protein sequences. However, these algorithms are not equipped to assess
433 pathogenicity in synonymous variants, which are under different constraints [11]. Recognizing that there
434 is a critical need for methods that better predict the potential whether sSNVs have pathogenic impact and

435 function, our goal in this present study was the generation of such metrics. ViennaRNA stability and SPI
436 metrics are available for download for all known sSNVs, to enable researchers and clinicians to evaluate
437 WES and WGS data in combination with tools such as Annovar [71], SnpEff [72] and VEP [73].

438 The role of synonymous mutations in human disease is gaining recognition in the community. The
439 RNAsnp Web Server predicts the change in optimal mRNA structure and base-pairing probabilities due
440 to a SNV [74]. A command line tool called remuRNA calculates the relative entropy between the mutant
441 and wildtype mRNA structural ensembles [75]. While these tools predict disruptions to mRNA structure,
442 they do not attempt to predict pathogenicity and must be executed manually on each variant of interest.
443 Both RNAsnp and remuRNA were recently utilized to create a database of synonymous mutations in
444 cancer (SynMICdb), using data from COSMIC across 88 tumor types [76]. For constitutional genetic
445 disease, a related resource is the databases of Deleterious Synonymous (dbDSM), which manually curates
446 sSNVs reported to pathogenic in the literature and in databases like ClinVar [77]. These resources
447 represent an important step towards evaluating sSNVs in disease. However, relatively few sSNVs have
448 well supported evidence of their pathogenicity, due in part to the complexity of demonstrating the
449 molecular impact of these variants through functional studies. While in the above databases and elsewhere
450 we found dozens of examples of sSNVs suggested to be pathogenic due to their effects on mRNA
451 structure, most of these cases lacked functional studies to conclusively implicate the given variant in
452 disease. As such we focused on a set of six sSNVs that we believe the authors unequivocally demonstrated
453 to be pathogenic through their effects on mRNA structure (**TABLE 3**). This dataset included one variant in
454 OPTC associated with glaucoma [42], two variants in NKX2-5 associated with congenital heart defects
455 [44], one variant in DRD2 associated with post-traumatic stress disorder [40], and two variants in COMT
456 associated with pain sensitivity [41].

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

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

471

472 ***Molecular mechanisms underlying constraint of sSNVs***

473 Synonymous variants that impact mRNA secondary structure could confer pathogenicity in
474 numerous ways. For example, sSNVs can alter global RNA stability, where less stable RNA molecules
475 may be degraded more quickly resulting in lower protein levels [23, 25, 27]. As local RNA structure is
476 essential for the translation process, a more stable mRNA may not be able to initiate translation, also
477 resulting in lower protein levels [20, 21, 30, 31]. Additionally, numerous studies argue that by making the
478 mRNA structure too difficult, or too easy, for the ribosome to process, synonymous codons may act as a
479 subliminal code for protein folding [16, 30, 32, 33, 35].

480 While there is a large body of literature supporting the essential role of sSNVs in preservation of
481 mRNA secondary, sSNVs also play roles in other molecular processes that could impact our observations.
482 While the stability of an mRNA transcript can determine how quickly it is translated [19, 29, 30], protein
483 synthesis is also regulated by both the abundance [78] and recruitment of tRNAs through synonymous
484 codon utilization (codon bias) [79-81]. Changes in the bicodon bias has the potential to lead to
485 pathogenicity of synonymous mutations in human disease [16, 35]. As such, this may have had a
486 confounding impact on our analysis of constraint, but we attempted to mitigate this by including the tRNA
487 Adaptivity Index (a measure of tRNA abundance) in our set of confounding variables.

488 Some of our constraint might have its roots in the tertiary DNA structure, where DNA bending
489 plays roles in DNA packaging, transcriptional regulation, and specificity in DNA–protein recognition
490 [82]. While it has been suggested that synonymous codon choice may play a role in DNA bending [83],
491 the process is largely governed by intergenic and intronic DNA, and as such is unlikely to be directly
492 responsible for the constraint shown in **FIGURE 2**. We attempted to mitigate the role of local DNA
493 sequence context and CpG content, by including both in our set of confounding variables when generating
494 SPI. However, methylation of CpG dinucleotides can regulate gene expression, potentially making it
495 difficult to state whether selection for mRNA structure causes the retention of CpGs, or whether the
496 retention of CpGs is regulated by a process independent of mRNA structure. A strong reason for CpGs to
497 operate causally with regard to mRNA structure is that they are the most important determinant of mRNA
498 structure (along with G+C content, with which CpG content correlates closely). Retention of 5' ORF CpG
499 sites occurs at a high frequency in the first exon of coding genes; a stacked C:G + G:C base pairing has
500 the lowest free energy of the 36 possible stacked base-pair combinations [84]; and deamination of CpGs
501 can be suppressed and repaired by existing enzymatic mechanisms. Thus, CpG dinucleotides represent the
502 easiest and most natural way to determine mRNA structure.

503 Finally, it is important to also consider the essential role that synonymous variants play in RNA
504 splicing. While we took care to exclude sSNVs impacting the canonical splice sites from our constraint
505 analysis, exonic variants beyond the canonical splice site can disrupt splice enhancers [85], or they may
506 also activate cryptic splice sites, leading to aberrant pre-mRNA splicing and loss of coding sequence [86].
507 Given the diversity of molecular roles that synonymous codons have, it will be important for future studies
508 to create scores that would allow assessment of sSNV pathogenicity through any these possible
509 mechanisms.
510

511 **CONCLUSIONS**

512 We have shown that sSNVs which stabilize or destabilize mRNA are significantly constrained in
513 the human population, thereby supporting a growing body of evidence that previously assumed "silent"
514 polymorphisms actually play crucial roles in regulation of gene expression and protein function. We have
515 demonstrated that this connection is rich, complex, and biologically intuitive. Given that there are multiple
516 mechanisms by which sSNVs influence biological function, we are almost certainly missing undiscovered
517 disease etiologies when these variants are ignored. In addition to providing the community with a dataset
518 of ten ViennaRNA structural metrics for every known synonymous variant, our Structural Predictivity
519 Index is the first metric of its kind to enable global assessment of sSNVs in human genetic studies. We
520 hope that these metrics will be utilized to accurately assess and prioritize an underrepresented class of
521 genetic variation that may be playing significant and as yet to be realized role in human health and disease.
522

523 **METHODS**

524 ***Raw Dataset***

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

528

529 ***Overview of RNA structure prediction process***

530 To estimate the structural properties of a sSNVs we used the ViennaRNA software package, a
531 secondary structure prediction package that has been extensively utilized and continuously developed for
532 nearly twenty-five years. ViennaRNA uses the standard partition-function paradigm of RNA structural
533 prediction [87]. We utilize version 2.0 of ViennaRNA [18]. Applying ViennaRNA to every possible SNV
534 in the human genome (about 500,000,000 calculations) was a computationally challenging task which we
535 carried out using an Apache Spark framework powered by Amazon Web Services (AWS). We built a
536 pipeline which read in and analyzed a SNV and stored the results in AWS Simple Storage Service (S3) in
537 Parquet columnar file format (**FIGURE 2**). The ease and capacity of AWS greatly facilitated the project,
538 and the affordability of S3 storage means our data can easily be shared with others. The software we
539 developed is available on GitHub: https://github.com/nch-igm/rna-stability.

540

541 ***RNA structure prediction methodology***

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

546 ViennaRNA's flagship module RNAfold, which yielded three predicted mRNA secondary structures –
547 the minimum free energy, centroid, and maximum expected accuracy structure – as well as numeric values
548 for the free energy of each structure, and a fourth metric measuring the energy of the whole ensemble (see
549 the documentation of [18] for detailed descriptions of these concepts). Comparing the free energies
550 between the wildtype and mutant for each type of structure gave us the four stability metrics delta-MFE
551 (dMFE), dCFE, dMEAFE and dEFE. Next, the predicted structures were processed by the ViennaRNA
552 module RNApdist, which counted the edge-differences to produce the four edge-metrics MFEED
553 (minimum free energy edit distance), CFEED, MEAED and EFEED. As a final step, the predicted
554 structures were further processed by the ViennaRNA program RNAdistance to obtain the diversity metrics
555 dCD and dEND (change in distance from centroid and change in ensemble diversity, respectively).
556 This whole procedure was carried out using custom developed Spark wrappers of RNAfold,
557 RNApdist and RNAdistance, with slight modifications to the source code to suppress the creation of
558 graphics files. After building our fasta files, we were able to compute all 10 ViennaRNA metrics for over
559 a half billion sequences in less than 24 hours using 51 c4.8xlarge AWS EMR computing nodes.
560
561 ***Construction of final dataset for synonymous SNVs***
562 The next step was to extract the sSNVs. This task was complicated by the fact that a SNV might
563 have appeared in several different transcripts, and could be synonymous in some and non-synonymous in
564 others. To address this challenge, we first annotated every SNV using the program snpEff [72], whose
565 source code was modified to allow record-by-record calling via Spark. This snpEff analysis produced
566 annotations of predicted biotype, e.g. missense, synonymous, canonical splice site, etc. To validate these
567 snpEff predictions we manually predicted the biotype of each SNV using start and stop codon information
568 from RefSeq (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/RefSeqGene/refseqgene.*.genomic.gbff.gz). The

569 small number of sSNVs where our predicted biotype disagreed with snpEff's were discarded. We then
570 defined a "synonymous SNV" to be one that was (A) synonymous in at least one transcript, (B)
571 synonymous or within the UTRs in all transcripts, and (C) not implicated in splicing by snpEff. Each
572 sSNV identified as "synonymous" by this scheme was assigned a "home transcript," chosen based on
573 proximity to the start codon, then on maximal transcript coding sequence length, and then arbitrarily.

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

577

578 ***Merging of sSNV GRCh38 transcript coordinates with gnomAD GRCh37 coordinates***

579 To measure constraint operating on a sSNV we used population frequencies obtained from the
580 gnomAD database. Since this resource only existed for the GRCh37 reference build, we lifted our entire
581 dataset from GRCH38 to GRCh37. The lifting procedure was carried out using the Picard Tools program
582 liftOver [88], which was executed using a custom Spark wrapper. The joining of the gnomAD frequencies
583 to our main dataset was a task greatly facilitated by Spark's parallel processing and native Parquet support.
584 Since the great majority (approximately 90%) of sSNVs were marked with gnomAD frequency 0, it is
585 important to identify sSNVs marked zero purely through a lack of coverage. To achieve this, we flagged
586 and removed all sSNVs where fewer than 70% of samples had at least 20X coverage.

587

588 ***Further variant annotations***

589 Next, we estimated the local nucleotide content around each sSNV. We divided each transcript
590 into windows of 40 bases and in each window computed the proportion of A's, C's, G's, T's, CpG's and
591 AT's in the surrounding three windows. Finally we joined multiple additional annotations (including

592 conservation metrics such as PhyloP) from the dbNSFP dataset [89]. Again, this heavy task was greatly
593 facilitated by our Spark framework.
594
595 ***Partition of dataset***
596 We carried out most of the analysis separately on subsets of data defined by a common mRNA
597 reference and alternate allele, e.g those sSNVs of form C>A. The reference and alternate alleles exert such
598 a huge influence on gnomAD frequency that the best solution seemed to be to control for them explicitly.
599 Dividing our dataset based on mRNA alleles (as opposed to DNA alleles, which do not depend on
600 transcript sense) is a step justified in **SUPPLEMENTARY TABLE 3**.
601
602 ***Depleter variables***
603 Depleter variables (so called because they explain some of the gnomAD depletion at values of a
604 structural variable) are given in **TABLE 1**. They are chosen to be the sequence feature that explains the
605 greatest portion of the connection between a structural metric (e.g. dMFE) and Y in a context. Possible
606 Depleter variables are local nucleotide content and the specific nucleotides up/downstream of the sSNV.
607 To compute the correlation between a structural metric (e.g. dMFE) and Y that is left unexplained
608 by a sequence feature (e.g. CpG content) in a particular REF-ALT context, we first build a simple logistic
609 regression model between CpG content and Y, which gives us an estimate $\mathbf{P}(\mathrm{Y} = 1 \mid \mathrm{CpG\ content})$ for
610 every sSNV in the context (based on the proportion of CpGs in the surrounding 120 nucleotides). We then
611 plug this "structure-less" estimate into the expression

612 $$\mathrm{V}_{\mathrm{CpG\ content}} = \sum_{x} \mathbf{n_x} * (\mathbf{E}(\,\mathbf{P}(\mathrm{Y} = 1 \mid \mathrm{CpG\ content}) \mid \mathrm{dMFE} = \mathrm{x}) - \mathbf{P}(\mathrm{Y} = 1 \mid \mathrm{dMFE} = \mathrm{x}))^2$$

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

615 $$V_{null} = \sum_{x} \mathbf{n_x} * (\mathbf{P}(Y = 1) - \mathbf{P}(Y = 1 \mid dMFE = x))^2$$

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

617 $$R^2_{CpG\ content} = 1 - \frac{V_{CpG\ content}}{V_{null}}$$

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

621

622 ***Construction of SPI***

623 To construct our final SPI scores we built two separate models over each of our 14 contexts to
624 predict the event MAF > 0. The "null" model used all natural features - the nine nucleotides in the SNV's
625 home and adjacent codons, the proportion of A/C/G/T/CpG/AT's in the surrounding 120 nucleotides, the
626 sSNV's position in its transcript and the transcript's length, and the tAI (tRNA Adapation Index obtained
627 from a supplement of [90] from https://ars.els-cdn.com/content/image/1-s2.0-S0092867410003193-
628 mmc2.xls) of the wildtype and mutant codons. The second, "active" model used all these features plus our
629 10 ViennaRNA metrics. Both sets of variables were then used to predict MAF > 0. We then defined the
630 SPI score for a sSNV to be the base-10 logarithm of the active model's predicted P(Y=1) probability
631 divided by the null model's predicted P(Y=1). Context wise plots for SPI are given in the
632 **SUPPLEMENTARY FIGURE 6**.

633 We tried three different model-styles for computing the raw predictions that comprise SPI –
634 general logistic as implemented in python's sklearn LogisticRegression module, random forest as
635 implemented in sklearn's RandomForestClassifier, and gradient-boosted trees as implemented in the
636 python package xgboost. Performance of each SPI "flavor" is given in **SUPPLEMENTARY TABLE 4**. We

637 eventually settled on the general logistic model, as it out-performs the gradient-boosted tree model and
638 does not over-fit as the random forest mode does.
639
640 **DECLARATIONS**
641 ***Ethics approval and consent to participate***
642 Not applicable.
643
644 ***Consent for publication***
645 Not applicable.
646
647 ***Availability of data and materials:***
648 The software we developed and structural scores are available on GitHub: https://github.com/nch-
649 igm/rna-stability.
650
651 ***Competing interests***
652 The authors declare no competing interests.
653
654 ***Funding***
655 We thank the Nationwide Children's Foundation and The Abigail Wexner Research Institute at
656 Nationwide Children's Hospital for generously supporting this body of work. James L. Li was supported
657 by the Pelotonia Fellowship for Undergraduate Research through The Ohio State University
658 Comprehensive Cancer Society. These funding bodies had no role in the design of the study, no role in
659 the collection, analysis, and interpretation of data and no role in writing the manuscript.
660

661 ***Authors' contributions***

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

668

669 ***Acknowledgements***

670 This team works in the Steve and Cindy Rasmussen Institute for Genomic Medicine at Nationwide
671 Children's Hospital. The Institute is generously supported by the Nationwide Foundation Pediatric
672 Innovation Fund.

673

674 **ADDITIONAL FILES**

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

677

678

679 **FIGURE LEGENDS**

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

696

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

702 Simple Storage Solution (S3) bucket and processed through the pipeline consisting of 8 steps. We first
703 obtained the 101-base sequence centered around a SNV in a transcript and generated three alternate
704 sequences (with the ALT rather than the REF at position 51) (step 1). We next applied ViennaRNA
705 modules to sequence to obtain structural metrics (step 2). Results were then mapped to chromosomal
706 coordinates (step 3) and annotated with SnpEff to identify splice variants (step 4), lifted to the hg19 build
707 (step 5), annotated with gnomAD population frequencies (step 6) and coverage information (step 7), and
708 finally annotated with metrics from dbNSFP (step 8). Final dataset was written to Amazon S3 in Parquet
709 columnar file format for further analysis and interpretation.
710
711 **FIGURE 3. Exonic SNVs predicted to impact mRNA structure are constrained in the human**
712 **population.** Population frequency of SNVs was plotted against predicted impact on mRNA structure.
713 Circles show proportion of SNVs with nonzero gnomAD exonic frequency at each value of the RNA
714 stability metric dMFE. The bell-shaped pattern of constraint was observed across all classes of SNVs,
715 with constraint appearing to be greatest in sSNVs (green), followed by SNVs in the 5 prime UTR (orange),
716 then SNVs in the 3 prime UTR (blue), and finally nsSNVs (red). Values of dMFE with fewer than 1000
717 (synonymous), 200 (UTRs) or 2000 (missense) positive-MAF sSNVs are excluded. Axis limits are
718 restricted to show main pattern more clearly, resulting in removal of a few high-P(MAF>0) synonymous
719 outliers. Only SNVs with high exonic coverage are represented (see **METHODS** for details.)
720
721 **FIGURE 4. Synonymous variants predicted to impact mRNA structure are constrained in the**
722 **human population.** Population frequency of sSNVs were plotted against the predicted impact on mRNA
723 structure. Synonymous variants that disrupt structure tend to be absent from the gnomAD database, while
724 those with limited impact on structure appear at least once in the gnomAD database. **(A)** Proportion of

725 sSNVs with nonzero gnomAD frequency at each value of the RNA stability metric dMFE. Points with
726 fewer than 2000 positive-MAF sSNVs excluded. Color represents average CFEED value, to highlight the
727 relationship between minimum free energy and edit distance. **(B)** Analogous plot for metric CFEED
728 measuring edge differences between mutant/wildtype centroid structures. Color represents |dMFE|,
729 measuring absolute change in stability. **(C)** Analogous plot for diversity-metric dCD measuring change in
730 structural ensemble diversity due to sSNV. Color is by dMFE measuring change in stability.
731
732 **FIGURE 5. Synonymous CpG transitions are markedly constrained against destabilization of their**
733 **mRNA structures.** Population frequency of sSNV vs. effect on mRNA structure in synonymous CpG
734 transitions was examined. Proportion of synonymous CpG transitions with nonzero MAF at each value of
735 dMFE were determined for (**A**) CpG>CpA and (**B**) CpG>TpG synonymous mutations. dMFE values with
736 fewer than 75 nonzero-MAF sSNVs are excluded. Color gives average CFEED in each context, ranging
737 from 15 (blue) to 50 (red). Similarly, proportion of synonymous CpG transitions with nonzero MAF at
738 each value of CFEED were determined for (**C**) CpG>CpA sSNVs and (**D**) CpG>TpG sSNVs. Color
739 represents average dMFE and ranges from -0.8 (blue) to 1.85 (red). CFEED values with fewer than 75
740 nonzero-MAF sSNVs are excluded). Finally, proportion of synonymous CpG transitions with nonzero
741 MAF at each value of dCD (after rounding to nearest integer) were determined for (**E**) CpG>CpA and (**F**)
742 CpG>TpG sSNVs sSNVS. Color represents average dMFE and ranges from -3 (blue) to 4 (red). Rounded
743 dCD values with fewer than 75 nonzero-MAF sSNVs are excluded.
744
745 **FIGURE 6. SPI score correlates with constraint in synonymous CpG transitions.** Variants in
746 the contexts **(A)** CpG>CpA and **(B)** CpG>TpG are divided by SPI score into 20 equal bins and the value
747 P(MAF>0) plotted against the mean of each bin. We also colored by the mean dMFE over each bin. In

748 both contexts the constraint is highest towards negative SPI, i.e. sSNVs for which structural information
749 decreases the predicted probability that MAF > 0.
750
751 **TABLE 1. Structural metrics correlate with gnomAD frequency in most REF>ALT contexts.**
752 Correlation between structural metrics (**A**) dMFE, (**B**) CFEED and integer-rounded (**C**) dCD on the one
753 hand, and the quantity P(MAF>0) on the other, over all sSNVs in a given context. The R^2 and p-values
754 are obtained from a weighted least-squares linear regression, with the p-value corresponding to the linear
755 coefficient; a quadratic regression was also performed, but only the p-value was retained. Only context-
756 metric pairs with p-value < 0.005 are included. "Normalized slope" was obtained by dividing slope of
757 regression line by average P(MAF>0) in the context and then multiplying by range covered by metric in
758 its central 90% of sSNVs. "Depleter" is raw sequence variable that explains largest proportion of structural
759 trend *in this context*, with sign adjusted to correlate negatively with gnomAD frequency. "Depleter R^2"
760 gives proportion of variance explained by Depleter (see *Depleter variables* in **RESULTS** for details).
761
762 **TABLE 2. Area under curve for SPI score.** SPI was used to discriminate MAF > 0 using a simple
763 logistic model with 5-fold cross-validation. Table shows area under curve for model, averaged over the 5 training
764 and testing sets.
765
766 **TABLE 3: Known sSNVs clinically implicated for structural pathogenicity are successfully**
767 **predicted to be pathogenic by our structural metrics.** dbSNP RS number and standardized SNP
768 annotations are provided, along with the genes official gene symbol and disease the sSNV has been
769 associated with. The absolute value of dMFE, CFEED, dCD and SPI are provided, along with the
770 percentile value of that score, computed over each context, in parentheses.

References

1. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, King DA, Ambridge K, Barrett DM, Bayzetinova T, et al: **Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data.** *Lancet* 2015, **385:**1305-1314.
2. Deciphering Developmental Disorders Study: **Prevalence and architecture of de novo mutations in developmental disorders.** *Nature* 2017, **542:**433-438.
3. Wright CF, FitzPatrick DR, Firth HV: **Paediatric genomics: diagnosing rare disease in children.** *Nat Rev Genet* 2018, **19:**253-268.
4. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z, et al: **Clinical whole-exome sequencing for the diagnosis of mendelian disorders.** *N Engl J Med* 2013, **369:**1502-1511.
5. Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, Ward P, Braxton A, Wang M, Buhay C, et al: **Molecular findings among patients referred for clinical whole-exome sequencing.** *JAMA* 2014, **312:**1870-1879.
6. Ellingford JM, Barton S, Bhaskar S, Williams SG, Sergouniotis PI, O'Sullivan J, Lamb JA, Perveen R, Hall G, Newman WG, et al: **Whole Genome Sequencing Increases Molecular Diagnostic Yield Compared with Current Diagnostic Testing for Inherited Retinal Disease.** *Ophthalmology* 2016, **123:**1143-1150.
7. Hegde M, Santani A, Mao R, Ferreira-Gonzalez A, Weck KE, Voelkerding KV: **Development and Validation of Clinical Whole-Exome and Whole-Genome Sequencing for Detection of Germline Variants in Inherited Disease.** *Arch Pathol Lab Med* 2017, **141:**798-805.
8. Worthey EA: **Analysis and Annotation of Whole-Genome or Whole-Exome Sequencing Derived Variants for Clinical Diagnosis.** *Curr Protoc Hum Genet* 2017, **95:**9 24 21-29 24 28.
9. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al: **Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.** *Genet Med* 2015, **17:**405-424.
10. Alfares A, Aloraini T, Subaie LA, Alissa A, Qudsi AA, Alahmad A, Mutairi FA, Alswaid A, Alothaim A, Eyaid W, et al: **Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing.** *Genet Med* 2018, **20:**1328-1333.
11. Gelfman S, Wang Q, McSweeney KM, Ren Z, La Carpia F, Halvorsen M, Schoch K, Ratzon F, Heinzen EL, Boland MJ, et al: **Annotating pathogenic non-coding variants in genic regions.** *Nat Commun* 2017, **8:**236.
12. Fahraeus R, Marin M, Olivares-Illana V: **Whisper mutations: cryptic messages within the genetic code.** *Oncogene* 2016, **35:**3753-3759.
13. Lee M, Roos P, Sharma N, Atalar M, Evans TA, Pellicore MJ, Davis E, Lam AN, Stanley SE, Khalil SE, et al: **Systematic Computational Identification of Variants That Activate Exonic and Intronic Cryptic Splice Sites.** *Am J Hum Genet* 2017, **100:**751-765.
14. Ramanouskaya TV, Grinev VV: **The determinants of alternative RNA splicing in human cells.** *Mol Genet Genomics* 2017, **292:**1175-1195.
15. Vaz-Drago R, Custodio N, Carmo-Fonseca M: **Deep intronic mutations and human disease.** *Hum Genet* 2017, **136:**1093-1111.
16. Hanson G, Coller J: **Codon optimality, bias and usage in translation and mRNA decay.** *Nat Rev Mol Cell Biol* 2018, **19:**20-30.
17. Silverman SK: **A forced march across an RNA folding landscape.** *Chem Biol* 2008, **15:**211-213.
18. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL: **ViennaRNA Package 2.0.** *Algorithms Mol Biol* 2011, **6:**26.
19. Seffens W, Digby D: **mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences.** *Nucleic Acids Res* 1999, **27:**1578-1584.

817 20. Katz L, Burge CB: **Widespread selection for local RNA secondary structure in coding regions of bacterial**
818 **genes.** *Genome Res* 2003, **13:**2042-2051.
819 21. Chamary JV, Hurst LD: **Evidence for selection on synonymous mutations affecting stability of mRNA**
820 **secondary structure in mammals.** *Genome Biol* 2005, **6:**R75.
821 22. Gu W, Zhou T, Wilke CO: **A universal trend of reduced mRNA stability near the translation-initiation site**
822 **in prokaryotes and eukaryotes.** *PLoS Comp Biol* 2010, **6:**e1000664.
823 23. Duan J, Antezana MA: **Mammalian mutation pressure, synonymous codon choice, and mRNA**
824 **degradation.** *J Mol Evol* 2003, **57:**694-701.
825 24. Wan Y, Qu K, Ouyang Z, Kertesz M, Li J, Tibshirani R, Makino DL, Nutter RC, Segal E, Chang HY: **Genome-**
826 **wide measurement of RNA folding energies.** *Mol Cell* 2012, **48:**169-181.
827 25. Lazrak A, Fu L, Bali V, Bartoszewski R, Rab A, Havasi V, Keiles S, Kappes J, Kumar R, Lefkowitz E, et al: **The**
828 **silent codon change I507-ATC->ATT contributes to the severity of the DeltaF508 CFTR channel**
829 **dysfunction.** *FASEB J* 2013, **27:**4630-4645.
830 26. Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C: **Exposing synonymous mutations.** *Trends*
831 *Genet* 2014, **30:**308-321.
832 27. Shah K, Cheng Y, Hahn B, Bridges R, Bradbury NA, Mueller DM: **Synonymous codon usage affects the**
833 **expression of wild type and F508del CFTR.** *J Mol Biol* 2015, **427:**1464-1479.
834 28. Bevilacqua PC, Ritchey LE, Su Z, Assmann SM: **Genome-Wide Analysis of RNA Secondary Structure.** *Annu*
835 *Rev Genet* 2016, **50:**235-266.
836 29. Yang JR, Chen X, Zhang J: **Codon-by-codon modulation of translational speed and accuracy via mRNA**
837 **folding.** *PLoS Biol* 2014, **12:**e1001910.
838 30. Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley
839 BR, Coller J: **Codon optimality is a major determinant of mRNA stability.** *Cell* 2015, **160:**1111-1124.
840 31. Bazzini AA, Del Viso F, Moreno-Mateos MA, Johnstone TG, Vejnar CE, Qin Y, Yao J, Khokha MK, Giraldez
841 AJ: **Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic**
842 **transition.** *EMBO J* 2016, **35:**2087-2103.
843 32. Plotkin JB, Kudla G: **Synonymous but not the same: the causes and consequences of codon bias.** *Nat Rev*
844 *Genet* 2011, **12:**32-42.
845 33. Chaney JL, Clark PL: **Roles for Synonymous Codon Usage in Protein Biogenesis.** *Annu Rev Biophys* 2015,
846 **44:**143-166.
847 34. Faure G, Ogurtsov AY, Shabalina SA, Koonin EV: **Role of mRNA structure in the control of protein folding.**
848 *Nucleic Acids Res* 2016, **44:**10898-10911.
849 35. McCarthy C, Carrea A, Diambra L: **Bicodon bias can determine the role of synonymous SNPs in human**
850 **diseases.** *BMC Genomics* 2017, **18:**227.
851 36. Fernandez M, Kumagai Y, Standley DM, Sarai A, Mizuguchi K, Ahmad S: **Prediction of dinucleotide-specific**
852 **RNA-binding sites in proteins.** *BMC Bioinformatics* 2011, **12 Suppl 13:**S5.
853 37. Brummer A, Hausser J: **MicroRNA binding sites in the coding region of mRNAs: extending the repertoire**
854 **of post-transcriptional gene regulation.** *Bioessays* 2014, **36:**617-626.
855 38. Savisaar R, Hurst LD: **Both Maintenance and Avoidance of RNA-Binding Protein Interactions Constrain**
856 **Coding Sequence Evolution.** *Mol Biol Evol* 2017, **34:**1110-1126.
857 39. Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, Bazile C, Lambert NJ, Van Nostrand EL,
858 Pratt GA, et al: **Sequence, Structure, and Context Preferences of Human RNA Binding Proteins.** *Mol Cell*
859 2018, **70:**854-867 e859.
860 40. Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J, Gejman PV: **Synonymous**
861 **mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the**
862 **receptor.** *Hum Mol Genet* 2003, **12:**205-216.

863 41. Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko
864 L: **Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA**
865 **secondary structure.** *Science* 2006, **314:**1930-1933.
866 42. Acharya M, Mookherjee S, Bhattacharjee A, Thakur SK, Bandyopadhyay AK, Sen A, Chakrabarti S, Ray K:
867 **Evaluation of the OPTC gene in primary open angle glaucoma: functional significance of a silent change.**
868 *BMC Mol Biol* 2007, **8:**21.
869 43. Bartoszewski RA, Jablonsky M, Bartoszewska S, Stevenson L, Dai Q, Kappes J, Collawn JF, Bebok Z: **A**
870 **synonymous single nucleotide polymorphism in DeltaF508 CFTR alters the secondary structure of the**
871 **mRNA and the expression of the mutant protein.** *J Biol Chem* 2010, **285:**28741-28748.
872 44. Reamon-Buettner SM, Sattlegger E, Ciribilli Y, Inga A, Wessel A, Borlak J: **Transcriptional defect of an**
873 **inherited NKX2-5 haplotype comprising a SNP, a nonsynonymous and a synonymous mutation,**
874 **associated with human congenital heart disease.** *PLoS One* 2013, **8:**e83295.
875 45. Simhadri VL, Hamasaki-Katagiri N, Lin BC, Hunt R, Jha S, Tseng SC, Wu A, Bentley AA, Zichel R, Lu Q, et al:
876 **Single synonymous mutation in factor IX alters protein properties and underlies haemophilia B.** *J Med*
877 *Genet* 2017, **54:**338-345.
878 46. Hamasaki-Katagiri N, Lin BC, Simon J, Hunt RC, Schiller T, Russek-Cohen E, Komar AA, Bar H, Kimchi-Sarfaty
879 C: **The importance of mRNA structure in determining the pathogenicity of synonymous and non-**
880 **synonymous mutations in haemophilia.** *Haemophilia* 2017, **23:**e8-e17.
881 47. Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M, Franklin MJ, Shenker S, Stoica I: **Resilient**
882 **distributed datasets: a fault-tolerant abstraction for in-memory cluster computing.** In *Proceedings of the*
883 *9th USENIX conference on Networked Systems Design and Implementation*. pp. 2-2. San Jose, CA: USENIX
884 Association; 2012:2-2.
885 48. Wiewiorka MS, Messina A, Pacholewska A, Maffioletti S, Gawrysiak P, Okoniewski MJ: **SparkSeq: fast,**
886 **scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision.**
887 *Bioinformatics* 2014, **30:**2652-2653.
888 49. Abuin JM, Pichel JC, Pena TF, Amigo J: **SparkBWA: Speeding Up the Alignment of High-Throughput DNA**
889 **Sequencing Data.** *PLoS One* 2016, **11:**e0155461.
890 50. O'Brien AR, Saunders NF, Guo Y, Buske FA, Scott RJ, Bauer DC: **VariantSpark: population scale clustering**
891 **of genotype information.** *BMC Genomics* 2015, **16:**1052.
892 51. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ,
893 Cummings BB, et al: **Analysis of protein-coding genetic variation in 60,706 humans.** *Nature* 2016,
894 **536:**285-291.
895 52. Gronau I, Arbiza L, Mohammed J, Siepel A: **Inference of natural selection from interspersed genomic**
896 **elements based on polymorphism and divergence.** *Mol Biol Evol* 2013, **30:**1159-1171.
897 53. Huang YF, Gulko B, Siepel A: **Fast, scalable prediction of deleterious noncoding variants from functional**
898 **and population genomic data.** *Nat Genet* 2017, **49:**618-624.
899 54. Vinson C, Chatterjee R: **CG methylation.** *Epigenomics* 2012, **4:**655-663.
900 55. Bellacosa A, Drohat AC: **Role of base excision repair in maintaining the genetic and epigenetic integrity**
901 **of CpG sites.** *DNA Repair* 2015, **32:**33-42.
902 56. Morgan MT, Bennett MT, Drohat AC: **Excision of 5-halogenated uracils by human thymine DNA**
903 **glycosylase. Robust activity for DNA contexts other than CpG.** *J Biol Chem* 2007, **282:**27578-27586.
904 57. Li E, Zhang Y: **DNA methylation in mammals.** *Cold Spring Harb Perspect Biol* 2014, **6:**a019133.
905 58. Cohen NM, Kenigsberg E, Tanay A: **Primate CpG islands are maintained by heterogeneous evolutionary**
906 **regimes involving minimal selection.** *Cell* 2011, **145:**773-786.
907 59. Turner DH, Mathews DH: **NNDB: the nearest neighbor parameter database for predicting stability of**
908 **nucleic acid secondary structure.** *Nucleic Acids Res* 2010, **38:**D280-282.
909 60. Sauna ZE, Kimchi-Sarfaty C: **Understanding the contribution of synonymous mutations to human**
910 **disease.** *Nat Rev Genet* 2011, **12:**683-691.

911 61. Shabalina SA, Spiridonov NA, Kashina A: **Sounds of silence: synonymous nucleotides as a key to biological**
912 **regulation and complexity.** *Nucleic Acids Res* 2013, **41:**2073-2094.
913 62. Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, Wang X, Sun Z: **Performance evaluation of pathogenicity-**
914 **computation methods for missense variants.** *Nucleic Acids Res* 2018, **46:**7793-7804.
915 63. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein**
916 **function using the SIFT algorithm.** *Nat Protoc* 2009, **4:**1073-1081.
917 64. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A**
918 **method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7:**248-249.
919 65. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR, Campbell C: **An integrative**
920 **approach to predicting the functional effects of non-coding and coding sequence variation.**
921 *Bioinformatics* 2015, **31:**1536-1543.
922 66. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X: **Identifying novel constrained elements by**
923 **exploiting biased substitution patterns.** *Bioinformatics* 2009, **25:**i54-62.
924 67. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S: **Identifying a high fraction of the**
925 **human genome to be under selective constraint using GERP++.** *PLoS Comp Biol* 2010, **6:**e1001025.
926 68. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: **A general framework for estimating**
927 **the relative pathogenicity of human genetic variants.** *Nat Genet* 2014, **46:**310-315.
928 69. Quang D, Chen Y, Xie X: **DANN: a deep learning approach for annotating the pathogenicity of genetic**
929 **variants.** *Bioinformatics* 2015, **31:**761-763.
930 70. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E,
931 Karyadi D, et al: **REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants.**
932 *Am J Hum Genet* 2016, **99:**877-885.
933 71. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput**
934 **sequencing data.** *Nucleic Acids Res* 2010, **38:**e164.
935 72. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for**
936 **annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome**
937 **of Drosophila melanogaster strain w1118; iso-2; iso-3.** *Fly (Austin)* 2012, **6:**80-92.
938 73. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F: **The Ensembl Variant**
939 **Effect Predictor.** *Genome Biol* 2016, **17:**122.
940 74. Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, Gorodkin J: **The RNAsnp web server:**
941 **predicting SNP effects on local RNA secondary structure.** *Nucleic Acids Res* 2013, **41:**W475-479.
942 75. Salari R, Kimchi-Sarfaty C, Gottesman MM, Przytycka TM: **Sensitive measurement of single-nucleotide**
943 **polymorphism-induced changes of RNA conformation: application to disease studies.** *Nucleic Acids Res*
944 2013, **41:**44-53.
945 76. Sharma Y, Miladi M, Dukare S, Boulay K, Caudron-Herger M, Gross M, Backofen R, Diederichs S: **A pan-**
946 **cancer analysis of synonymous mutations.** *Nat Commun* 2019, **10:**2569.
947 77. Wen P, Xiao P, Xia J: **dbDSM: a manually curated database for deleterious synonymous mutations.**
948 *Bioinformatics* 2016, **32:**1914-1916.
949 78. Dong H, Nilsson L, Kurland CG: **Co-variation of tRNA abundance and codon usage in Escherichia coli at**
950 **different growth rates.** *J Mol Biol* 1996, **260:**649-663.
951 79. Sabi R, Tuller T: **Modelling the efficiency of codon-tRNA interactions based on codon usage bias.** *DNA*
952 *Res* 2014, **21:**511-526.
953 80. Quax TE, Claassens NJ, Soll D, van der Oost J: **Codon Bias as a Means to Fine-Tune Gene Expression.** *Mol*
954 *Cell* 2015, **59:**149-161.
955 81. Rocha EP: **Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient**
956 **decoding for translation optimization.** *Genome Res* 2004, **14:**2279-2286.
957 82. Harteis S, Schneider S: **Making the bend: DNA tertiary structure and protein-DNA interactions.** *Int J Mol*
958 *Sci* 2014, **15:**12335-12363.

959 83. Vinogradov AE: **DNA helix: the importance of being GC-rich.** *Nucleic Acids Res* 2003, **31:**1838-1844.
960 84. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded sequence dependence of thermodynamic**
961 **parameters improves prediction of RNA secondary structure.** *J Mol Biol* 1999, **288:**911-940.
962 85. Soukarieh O, Gaildrat P, Hamieh M, Drouet A, Baert-Desurmont S, Frebourg T, Tosi M, Martins A: **Exonic**
963 **Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico**
964 **Tools.** *PLoS Genet* 2016, **12:**e1005756.
965 86. Molinski SV, Gonska T, Huan LJ, Baskin B, Janahi IA, Ray PN, Bear CE: **Genetic, cell biological, and clinical**
966 **interrogation of the CFTR mutation c.3700 A>G (p.Ile1234Val) informs strategies for future medical**
967 **intervention.** *Genet Med* 2014, **16:**625-632.
968 87. McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary**
969 **structure.** *Biopolymers* 1990, **29:**1105-1119.
970 88. **Picard: a set of tools (in Java) for working with next generation sequencing data in the BAM format**
971 [http://broadinstitute.github.io/picard/]
972 89. Liu X, Wu C, Li C, Boerwinkle E: **dbNSFP v3.0: A One-Stop Database of Functional Predictions and**
973 **Annotations for Human Nonsynonymous and Splice-Site SNVs.** *Hum Mutat* 2016, **37:**235-241.
974 90. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y: **An**
975 **evolutionarily conserved mechanism for controlling the efficiency of protein translation.** *Cell* 2010,
976 **141:**344-354.
977

Figure 1

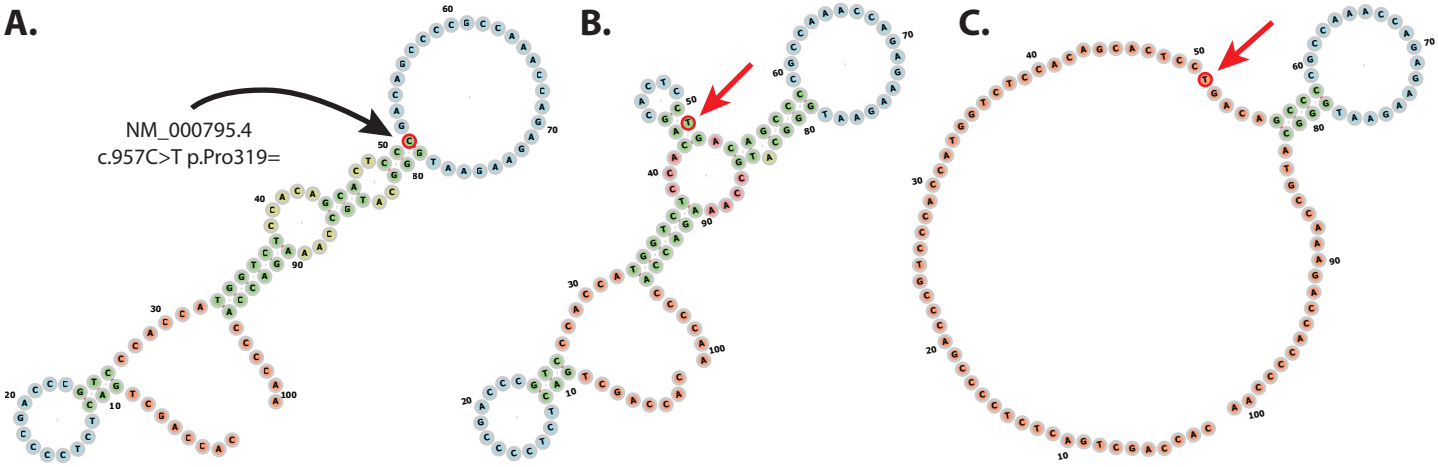


Figure 2

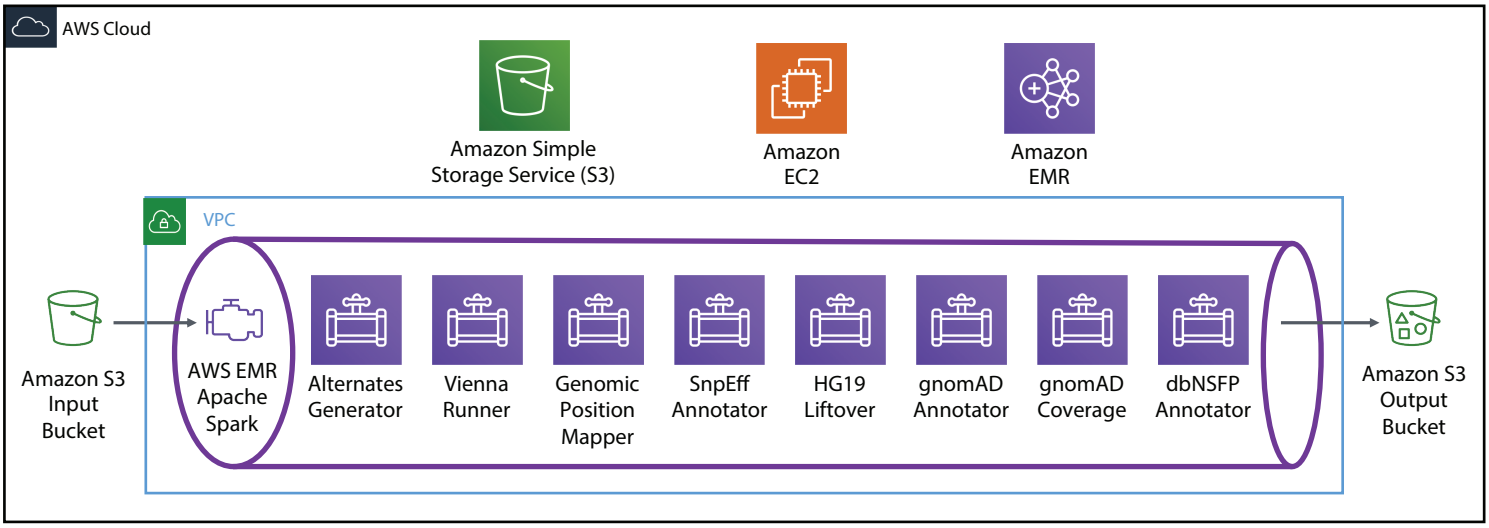

AWS Cloud
Amazon Simple Storage Service (S3)
Amazon EC2
Amazon EMR
VPC
Amazon S3 Input Bucket
AWS EMR Apache Spark
Alternates Generator
Vienna Runner
Genomic Position Mapper
SnpEff Annotator
HG19 Liftover
gnomAD Annotator
gnomAD Coverage
dbNSFP Annotator
Amazon S3 Output Bucket

Figure 3

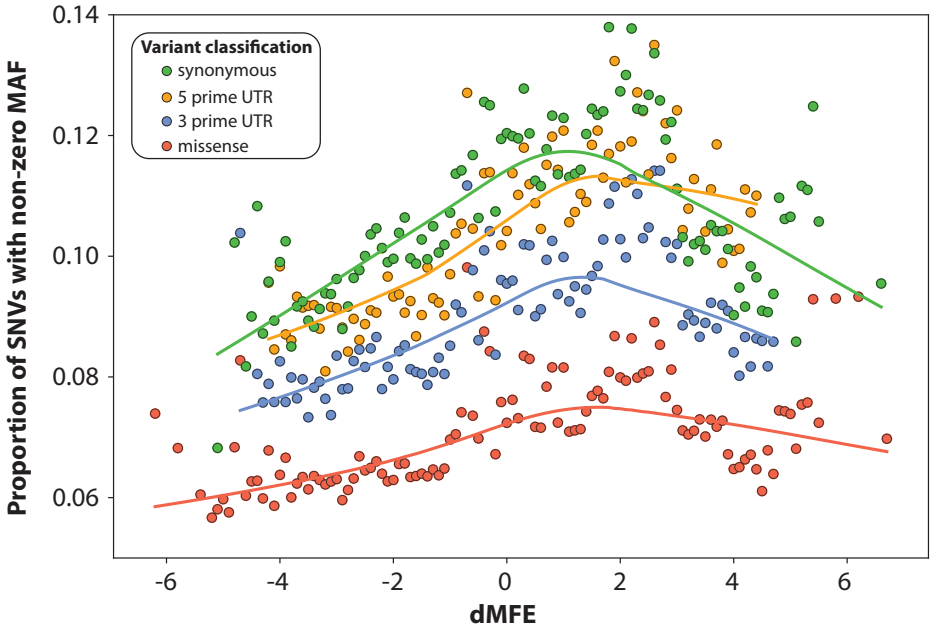

Variant classification
synonymous
5 prime UTR
3 prime UTR
missense
Proportion of SNVs with non-zero MAF
0.14
0.12
0.10
0.08
0.06
-6
-4
-2
0
2
4
6
dMFE

Figure 4

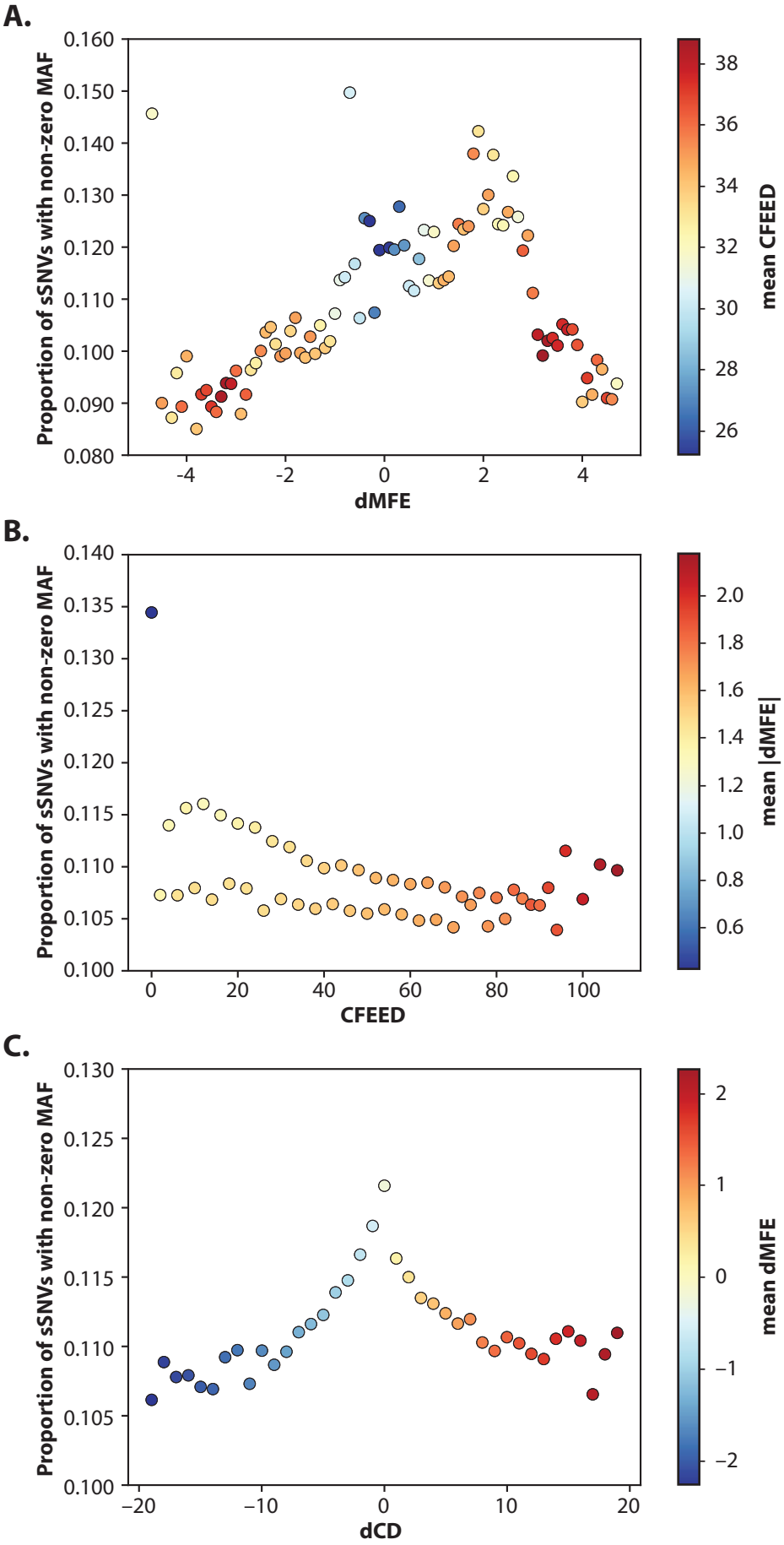

A.
Proportion of sSNVs with non-zero MAF
0.160
0.150
0.140
0.130
0.120
0.110
0.100
0.090
0.080
-4
-2
0
2
4
dMFE
38
36
34
32
30
28
26
mean CFEED
B.
Proportion of sSNVs with non-zero MAF
0.140
0.135
0.130
0.125
0.120
0.115
0.110
0.105
0.100
0
20
40
60
80
100
CFEED
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
mean |dMFE|
C.
Proportion of sSNVs with non-zero MAF
0.130
0.125
0.120
0.115
0.110
0.105
0.100
−20
−10
0
10
20
dCD
2
1
0
−1
−2
mean dMFE

Figure 5

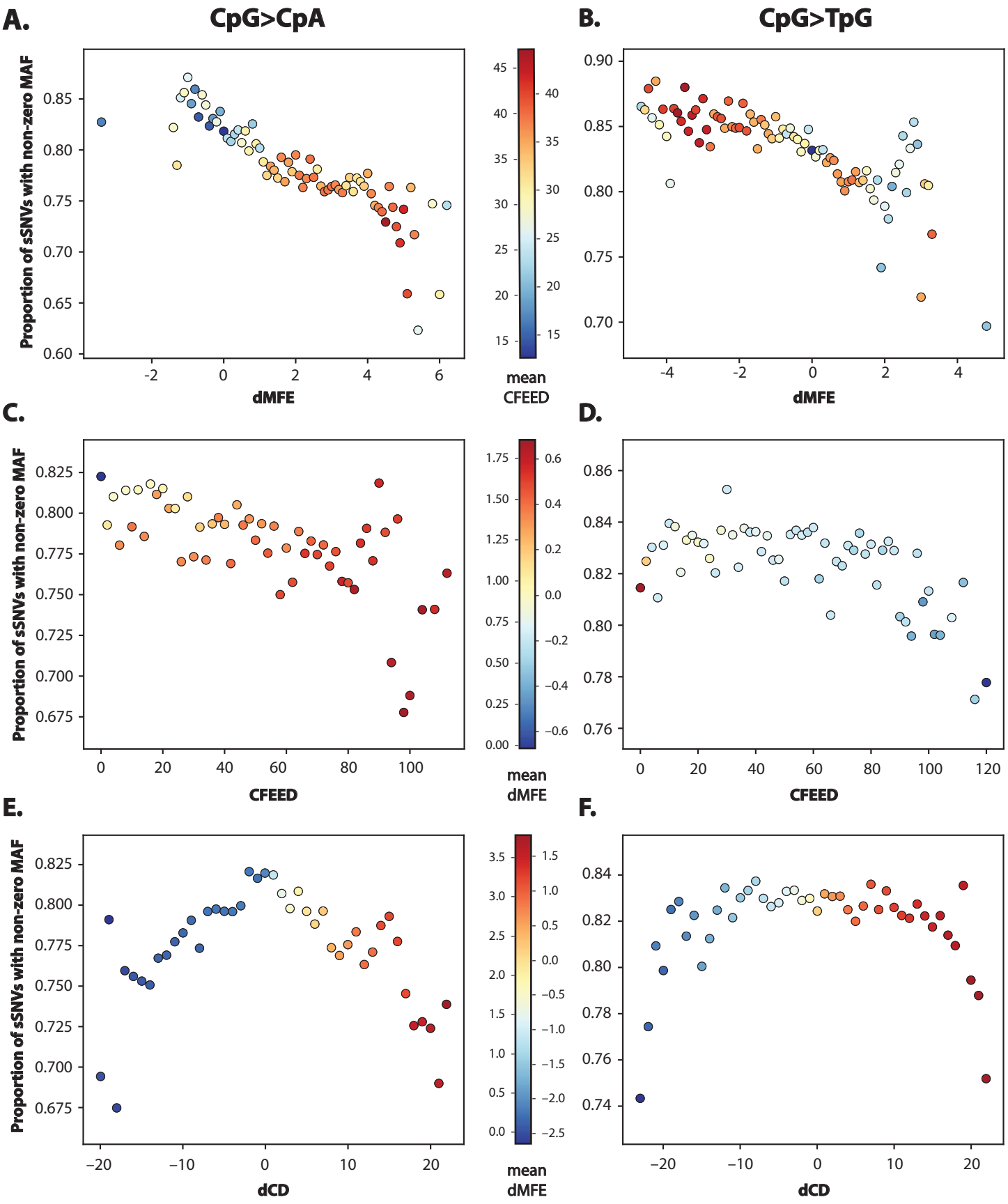
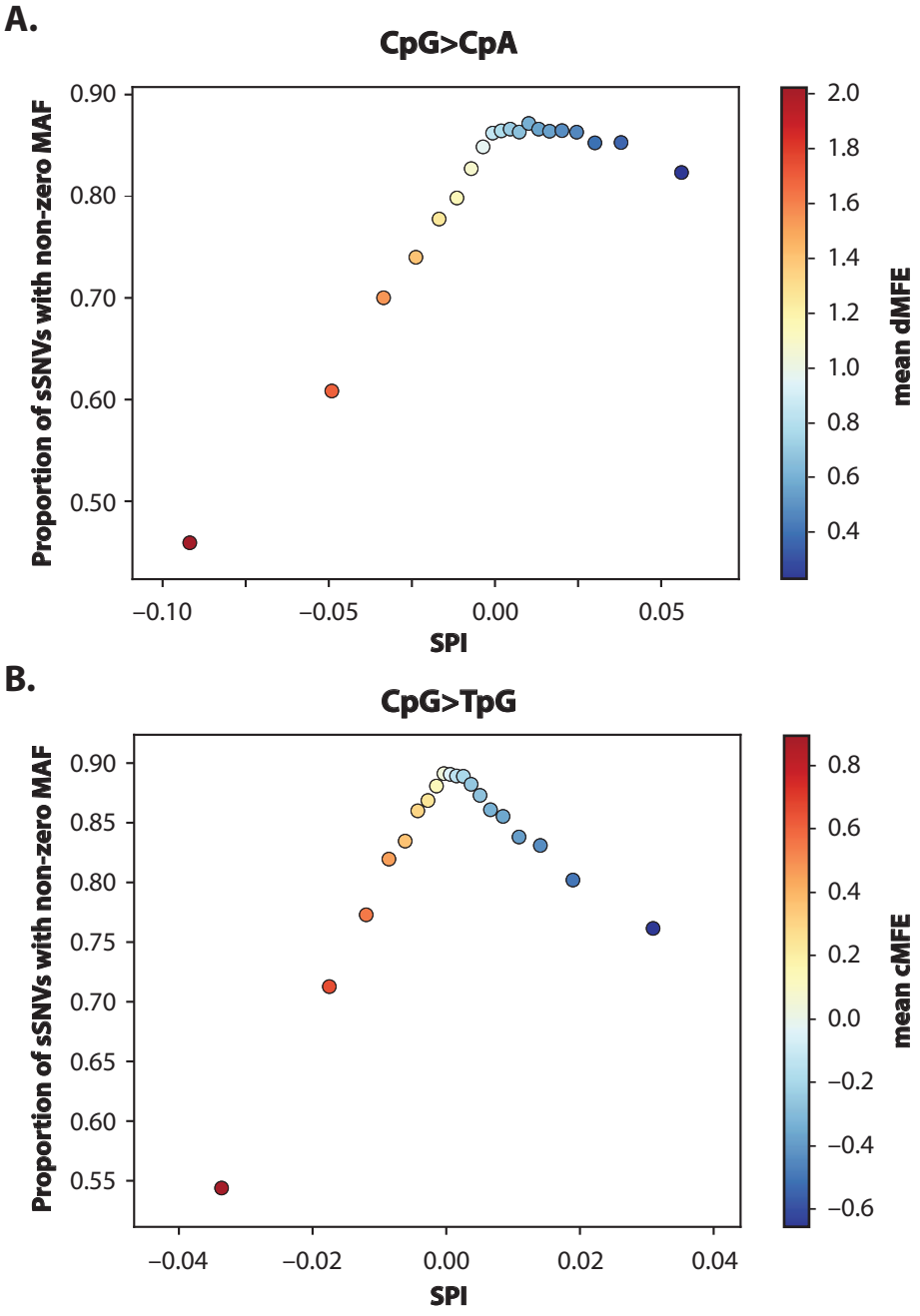

A.
CpG>CpA
B.
CpG>TpG
Proportion of sSNVs with non-zero MAF
dMFE
mean CFEED
C.
D.
CFEED
mean dMFE
E.
F.
dCD

Figure 6


A.
CpG>CpA
Proportion of sSNVs with non-zero MAF
0.90
0.80
0.70
0.60
0.50
−0.10
−0.05
0.00
0.05
SPI
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
mean dMFE
B.
CpG>TpG
Proportion of sSNVs with non-zero MAF
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
−0.04
−0.02
0.00
0.02
0.04
SPI
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
mean cMFE