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

22 **TABLE OF CONTENTS**

23 ABSTRACT 3

24 INTRODUCTION 4

25 RESULTS 7
26 *Massively parallel generation of RNA stability metrics* *7*
27 *Global constraint to maintain stability* *9*
28 *Variation of constraint with REF>ALT context* *10*
29 *CpG transitions have constraint against de-stabilization of their mRNA structures* *12*
30 *Constraint for mRNA stability in non-CpG-transitional contexts* *14*
31 *Depleter variables* *15*
32 *Global quantification of mRNA constraint* *16*
33 *Clinical Examples of Structural Pathogenicity* *17*

34 DISCUSSION 19
35 *Regulation of CpG transitions* *21*
36 *Importance of CpG and AT dinucleotides* *21*
37 *Successful identification of structurally disruptive sSNVs in known pathogenic synonymous*
38 *variants* *22*
39 *Mitigation of competing constraints* *24*
40 *Molecular mechanisms underlying constraint of variants impacting mRNA secondary structure* *25*

41 CONCLUSION 27

42 METHODS 28
43 *Raw Dataset* *28*
44 *Overview of RNA structure prediction process* *28*
45 *RNA structure prediction methodology* *28*
46 *Construction of final dataset for synonymous SNVs* *29*
47 *Merging of sSNV GRCh38 transcript coordinates with gnomAD GRCh37 coordinates* *30*
48 *Further variant annotations* *30*
49 *Partition of dataset* *31*
50 *Depleter variables* *31*
51 *Construction of SPI* *32*

52 COMPETING INTERESTS 34

53 AUTHOR CONTRIBUTIONS 34

54 ADDITIONAL FILES 34

55 ACKNOWLEDGEMENTS 34

56 FIGURE LEGENDS 35

57 REFERENCES 39

58

59

60 **ABSTRACT**

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

66 ***Results.*** We developed novel software to conduct a global transcriptome study in which RNA
67 folding statistics were computed for 469 million SNVs in 45,800 transcripts using an Apache Spark
68 implementation of the ViennaRNA software package in the cloud. Focusing our analysis on the subset of
69 17.9 million sSNVs we discover that variants predicted to disrupt mRNA structure have lower rates of
70 incidence in the human population. Given that the community lacks tools to evaluate the potential
71 pathogenic impact of sSNVs, we introduce a "Structural Predictivity Index" (SPI) to quantify this
72 constraint due to mRNA structure.

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

80

81 **INTRODUCTION**

82 While next generation sequencing (**NGS**) has accelerated the discovery of new functional variants
83 in syndromic and rare monogenic diseases, many more disease-causing genes and novel genetic etiologies
84 remain to be discovered (Wright *et al.* 2015; Deciphering Developmental Disorders Study 2017). Accurate
85 molecular genetic diagnosis of a rare disease is essential for patient care (Wright *et al.* 2018),yet today's
86 best molecular tests and analysis strategies leave 60-75% of patients undiagnosed (Yang *et al.* 2013; Yang
87 *et al.* 2014b; Ellingford *et al.* 2016; Hegde *et al.* 2017; Worthey 2017). Current clinical practice for
88 sequence variant interpretation focuses primarily on missense, nonsense or canonical splice variants
89 (Richards *et al.* 2015), with numerous bioinformatics prediction algorithms and databases developed for
90 functional prediction and annotation of non-synonymous single-nucleotide variants (nsSNVs) that impact
91 protein function through changes in the underlying coding sequence (Alfares *et al.* 2018). However, these
92 algorithms are inadequate to infer pathogenicity in non-protein-altering variants such as intronic or
93 synonymous variants, which are under different and weaker evolutionary constraints (Gelfman *et al.*
94 2017). While the potentially pathogenic impact of non-synonymous single nucleotide variants (**nsSNVs**)
95 that change the protein sequence are well understood, we have limited knowledge in regard to the role that
96 synonymous SNVs (**sSNVs**) may have in human health and disease.

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

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

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

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

126 2003; Wan *et al.* 2012; Hunt *et al.* 2014; Bevilacqua *et al.* 2016). A structure-altering sSNV in the
127 dopamine receptor DRD2 was shown to inhibit protein synthesis and accelerate mRNA degradation (Duan
128 *et al.* 2003). A sSNV in the *COMT* gene, implicated in cognitive impairment and pain sensitivity, was
129 shown *in vitro* to constrain enzymatic activity and protein expression (Nackley *et al.* 2006). A sSNV
130 discovered in the OPTC gene of a glaucoma patient resulted in decreased protein expression *in vivo*
131 (Acharya et al. 2007). In cystic fibrosis patients, a sSNV in the *CFTR* gene was linked to decreased gene
132 expression (Bartoszewski *et al.* 2010). Additionally, a silent codon change, I507-ATC→ATT, contributes
133 to CFTR dysfunction by a change in mRNA secondary structure that alters the dynamics of translation
134 leading to misfolding of the CFTR protein (Lazrak *et al.* 2013; Shah *et al.* 2015). Two sSNVs in the
135 NKX2-5 gene decreased the mRNA's transactivation potential in a yeast-based assay (Reamon-Buettner
136 et al. 2013). In hemophilia B, the sSNV c.459G>A in factor IX impacts the transcript's secondary structure
137 and reduces extracellular protein levels (Simhadri *et al.* 2017), and both synonymous and nonsynonymous
138 SNVs were shown more likely be deleterious when occurring in a stable region of mRNA in hemophilia
139 associated genes F8 and Duchenne's Muscular Dystrophy (Hamasaki-Katagiri *et al.* 2017).

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

149 **RESULTS**

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

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

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

172 four metrics describing the sSNV's effect on mRNA stability. Next, we fed the predicted structures from
173 RNAfold into the Vienna programs RNApdist and RNAdistance to obtain 6 additional metrics quantifying
174 the change in base-pairing and ensemble diversity due to each SNV. We performed this procedure for all
175 469 million possible SNVs in 45,800 transcripts.

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

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

195 absent from gnomAD, versus un-constrained variants as being those with exon minor allele frequency >0,
196 a strategy similar to that employed by other groups (Gronau *et al.* 2013; Huang *et al.* 2017).
197
198 ***Global constraint to maintain stability***
199 Our study reveals a striking connection between a given sSNV's impact on mRNA structure and
200 its frequency in the gnomAD database. We define the central variable Y to be Y=1 when a sSNV is present
201 in gnomAD and Y=0 when the sSNV is absent. Synonymous variants that disrupt structure tend to have
202 Y=0 (i.e. are absent from the gnomAD database), while those with limited impact on structure tend to
203 have Y=1 (i.e. appear at least once in the gnomAD database). This central finding is summarized in
204 **FIGURE 3**, which shows the proportion of synonymous SNVs with Y=1 at every value of the metrics
205 dMFE, CFEED and dCD (note: here and throughout, dCD values are rounded to the nearest integer). The
206 leading **FIGURE 3A** shows the correlation between Y and the stability metric dMFE. The bell-shaped
207 distribution shows that Y=1 occurs most often among those sSNVs that maintain the mRNA's existing
208 level of stability, i.e. those sSNVs with dMFE close to 0. When the sSNV either over-stabilizes the mRNA
209 (low dMFE) or de-stabilizes it (high dMFE) the sSNV is depleted in the population roughly in proportion
210 to the level of disruption.
211 **FIGURE 3B** shows an analogous plot for the structural disruption metric CFEED (see
212 **SUPPLEMENTARY FIGURE 2** for an illustration of how CFEED is calculated). This plot appears to depict
213 two separate trends, but actually shows a single pattern that alternates between high and low on successive
214 values: the SNVs with CFEED=0,4,8,12... are enriched over those with CFEED=2,6,10,14... (CFEED can
215 only take on even values because the destruction/creation of a base pair always requires two edits). One
216 possible explanation for this duality is that when CFEED fails to be divisible by 4, there is necessarily a
217 change in the total number of base-pairings in the mRNA centroid structure. Thus, sSNVs which conserve

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

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

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

236

237 ***Variation of constraint with REF>ALT context***

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

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

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

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

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

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

275

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

277 The data in **TABLE 1** highlight that our observed constraint for mRNA structure is the greatest
278 when considering CpG transitions. Since these variants (and their suppression) are crucial to the story of
279 mRNA stability, it is important to have an appreciation of their role in a biochemical context. The
280 dinucleotide CG (usually denoted CpG to distinguish this linear sequence from the CG base-pairing of
281 cytosine and guanine) is capable of becoming methylated and then mutating by a process called
282 "deamination" into a TG dinucleotide. While studies have demonstrated that methylated CpG residues are
283 up to 40X times more likely to be deaminated than their unmethylated counterparts (Vinson and Chatterjee
284 2012), mechanisms exist to enzymatically repair CpG deaminations (Morgan *et al.* 2007; Bellacosa and
285 Drohat 2015). In mammals 70-80% of CpGs are methylated, which makes a CpG transition almost 5x
286 more common than any other mutation-type among mammals (see **SUPPLEMENTARY DATA TABLE 3**) (Li

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

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

295 In the present study we find strong evidence that CpG transitions are constrained against de-
296 stabilization of their mRNA structures. This striking trend is largely explained (in a statistical sense) by
297 CpG content, i.e. number of CpG dinucleotides in the surrounding 120 nucleotides of the mRNA transcript
298 (see "Depleter R^2 in **TABLE 1B**). We distinguish CpG>CpA versus CpG>TpG transitions (the former of
299 these usually results from a CpG>TpG deamination on an anti-sense DNA strand), as these two mutation-
300 types show a qualitatively different constraint for mRNA structure. **FIGURE 4** shows the performance of
301 our three main metrics in CpG-transitional contexts. Most strikingly, we find that synonymous CpG>CpA
302 and CpG>TpG mutations both show a steady constraint against de-stabilization (high dMFE) (**FIGURES**
303 **4A & 4B**). Fascinatingly, both contexts exhibit a cluster of outliers in the most destructive (i.e. most de-
304 stabilizing region), suggestive of extreme constraint borne of significant structural disruption. Though the
305 two plots exhibit the same basic shape, the context CpG>CpA of **FIGURE 4A** shows higher de-stabilizing
306 tendencies (higher dMFE values) and also a stronger constraint (lower P(Y=1)).

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

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

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

320

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

322 We see the strongest constraint for mRNA structure in CpG transitions, but we observe an
323 analogous pattern in most REF>ALT contexts (as indicated by **TABLE 1**). We can classify these remaining
324 contexts based on whether their slopes in **TABLE 1A** are positive or negative. **SUPPLEMENTARY FIGURE**
325 **4** shows plots of contexts where dMFE and the gnomAD variable Y are negatively correlated. In such
326 contexts the data are consistent with the hypothesis that sSNVs which de-stabilize mRNA are constrained.
327 Notably, all these contexts are strong>weak (or strong>strong in the case of C>G), consistent with the
328 principle that one purpose of such nucleotides is to maintain stability. The coloring by CFEED indicates
329 that a change in either direction is likely to alter the mRNA secondary structures.

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

333 where SNPs that alter stability or diversity are actually enriched). The context T>C exhibits a notable
334 constraint in either direction, an anomaly which we speculate on in the **DISCUSSION.**
335
336 ***Depleter variables***
337 In **TABLE 1** we provide a "Depleter" for the connection between our RNA folding metrics and
338 gnomAD frequencies for each mutational context. The name "Depleter" signifies that each such variable
339 is chosen so as to correlate negatively with gnomAD (which is why these variables are given with +/-
340 signs in **TABLE 1**). For example, the Depleter for dMFE in the context CpG>CpA is +CpG content,
341 meaning that when CpG content increases in this context, the varaible Y is depleted.
342 The Depleter is chosen to be the variable that best explains the connection between the mRNA
343 structural variable and Y in the given context. The proportion of connection explained is given by the field
344 "Depleter R^2". For example, in the context CpG>CpA we can explain 78% of the dMFE-gnomAD
345 connection using a model that relies only CpG content.
346 To determine which variable is most informative (and should therefore be called the Depleter) we
347 compute an associated R^2 for a set of features of the sequence around the sSNV (the upstream/downstream
348 nucleotides and the proportion of A, C, G, T, CpG or ApT [di]nucleotides in the surrounding 120 bases).
349 Each of these features is used to build a simple logistic model to predict Y=1, and the predictions of the
350 model are then compared to the actual proportion P(Y=1) at a value of the metric. For example, building
351 a CpG-context-based model allows us to compute the quantity P(Y=1 | CpG content), and then we consider
352 the difference:

353 $$\mathbf{E}(\,\mathbf{P}(\mathrm{Y}=1 \mid \mathrm{CpG\ content}) \mid \mathrm{dMFE}) \;-\mathbf{P}(\mathrm{Y}=1 \mid \mathrm{dMFE}\,)$$

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

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

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

370

371 ***Global quantification of mRNA constraint***

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

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

384 $$\mathrm{SPI} = \log_{10}\left(\frac{P_s}{P_n}\right)$$

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

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

397

398 ***Clinical Examples of Structural Pathogenicity***

399 The literature reveals only a few examples of synonymous sSNVs unequivocally shown to be
400 pathogenic through their effects on mRNA structure. These sSNVs, with accompanying values of our
401 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**.

410 **DISCUSSION**

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

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

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

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

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

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

456 have higher population frequencies than those that do not (**TABLE 1)**, the main example being the
457 enrichment of high CFEED values in the contexts C>A and G>A. In both cases, the Depleter variable is
458 a trailing A, which correlates negatively with gnomAD frequency and CFEED. The presence of an A is
459 likely to have minimal effect on mRNA structure, suggesting that in these contexts the connection is partly
460 spurious.
461
462 ***Regulation of CpG transitions***
463 In the case of CpG transitions, it is difficult to state whether selection for mRNA structure causes
464 the retention of CpGs, or whether the retention of CpGs is regulated by a process independent of mRNA
465 structure. A strong reason for CpGs to operate causally with regard to mRNA structure is that they are the
466 single most important determinant of mRNA structure. Retention of 5' ORF CpG sites occurs at a high
467 frequency in the first exon of coding genes; a stacked C:G + G:C base pairing, has the lowest free energy
468 of the 36 possible stacked base-pair combinations (Mathews *et al.* 1999); and deamination of CpGs can
469 be suppressed and repaired by existing enzymatic mechanisms. Thus, CpG dinucleotides represent the
470 easiest and most natural way to determine mRNA structure.
471
472 ***Importance of CpG and AT dinucleotides***
473 Our results in non-CpG-transitional dinucleotide contexts are largely explained by the reference
474 nucleotide's membership in a CpG/AT dinucleotide. These dinucleotides have the apparent effect of
475 mitigating the structural distortion caused by the sSNV, e.g. mutations C>A and C>G are less de-
476 stabilizing if the reference is part of a CpG (see **TABLE 1**, which shows that in these contexts the gnomAD
477 variable Y varies inversely with dMFE but directly with a trailing G). This presents us with the same
478 causal conundrum we have faced throughout our study: do sSNVs in a dinucleotide have higher

479 frequencies because the dinucleotide mitigates the structural damage, or is it due to some other reason,
480 unrelated to mRNA structure? While this question is difficult to answer definitively, we believe that the
481 data presented in this present study and that the body of mRNA structural literature supports that
482 preservation of mRNA secondary structure is acting as a functional constraint on sSNVs in a dinucleotide.
483 For example, one point in favor of a causal role is that both CpGs and ATs have been specifically
484 implicated as drivers of mRNA structure (Al-Saif and Khabar 2012). Moreover, consistent with our
485 findings, a seminal paper in the field of RNA folding suggested that it is the *dinucleotide* content of an
486 mRNA that contributes most to its stability (Workman and Krogh 1999).
487
488 ***Successful identification of structurally disruptive sSNVs in known pathogenic synonymous variants***
489 Over the last decade numerous studies have demonstrated that synonymous variants play essential
490 molecular roles in regulating both mRNA structure and processing, including regulation of protein
491 expression, folding and function (reviewed in Sauna and Kimchi-Sarfaty 2011; Shabalina *et al.* 2013;
492 Fahraeus *et al.* 2016). However, the potential for pathogenic synonymous variants that impact RNA
493 folding in human genetic disease remains largely unknown. Current American College of Medical
494 Genetics (**ACMG**) guidelines for the assessment of clinically relevant genetic variants focus primarily on
495 missense, nonsense or canonical splice variants (Richards *et al.* 2015). These guidelines suggest that
496 synonymous "silent" variants should be classified as likely benign, if the nucleotide position is not
497 conserved and splicing assessment algorithms predict neither an impact to a splice consensus sequence
498 nor the creation of a new alternate splice consensus sequence. In the absence of functional tools that would
499 aid in the simultaneous assessment of both nsSNVs and sSNVs in a given patients genome, we are almost
500 certainly missing novel disease etiologies that have their molecular underpinnings in pathological
501 alterations to mRNA structure.

502 Numerous *in silico* tools exist to aid in the prediction of disease-causing missense variants, and
503 have accuracy in the 65%-85% range when evaluating known pathogenic variants (Li *et al.* 2018). Such
504 algorithms infer pathogenicity based on amino-acid substitutions (SIFT (Kumar *et al.* 2009), PolyPhen
505 (Adzhubei *et al.* 2010), FATHMM (Shihab *et al.* 2015)), nucleotide conservation (SiPhy (Garber *et al.*
506 2009), GERP++ (Davydov *et al.* 2010)) or an ensemble of annotations and scores (CADD (Kircher *et al.*
507 2014), DANN (Quang *et al.* 2015), REVEL (Ioannidis *et al.* 2016)). These tools predict whether a nsSNV
508 is pathogenic or benign, primarily due to the high conservation of protein sequences. However, these
509 algorithms are not equipped to assess pathogenicity in synonymous variants, which are under different
510 constraints (Gelfman *et al.* 2017). Recognizing that there is a critical need for methods that better predict
511 the potential whether sSNVs have pathogenic impact and function, our goal in this present study was the
512 generation of such metrics. Vienna RNA stability and SPI metrics are available for download for all known
513 sSNVs, to enable researchers and clinicians to evaluate WES and WGS data in combination with tools
514 such as Annovar (Wang *et al.* 2010), SnpEff (Cingolani *et al.* 2012) and VEP (McLaren *et al.* 2016).

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

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

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

539

540 ***Mitigation of competing constraints***

541 In addition to a potential role in mRNA structure, synonymous codons are likely under selection
542 for purposes other than mRNA structure, which could have confounded our analysis. Synonymous codon
543 utilization (codon bias) is known to direct gene expression and protein synthesis through regulating tRNA
544 recruitment (Rocha 2004; Sabi and Tuller 2014; Quax *et al.* 2015). Synonymous codons may also act as
545 a subliminal code for protein folding, with changes in a preferred locus potentially leading to pathogenicity
546 in synonymous mutations (McCarthy *et al.* 2017; Hanson and Coller 2018). While the stability of an
547 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,

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

578 Structural elements within the first 5 to 16 codons of mRNA have been shown to significantly
579 regulate protein expression levels in *E. coli* (Sato *et al.* 2001; Kudla *et al.* 2009). It is likely that both the
580 stability of mRNA folding near the ribosomal binding site and the reduced abundance of tRNAs coding
581 for N-terminal amino acids play crucial roles in slowing down initial stages of translation elongation
582 prevent subsequent ribosomal traffic jams (Tuller *et al.* 2010; Li and Qu 2013). More recently, it was been
583 shown that sequence motifs and mRNA structure within the first five codons are key in dictating the
584 efficiency of protein synthesis (Verma *et al.* 2019). By assessing over 250,000 reporter sequences in *E.*
585 *coli*, Verma and colleagues demonstrated that differences in this short ramp lead to striking changes in
586 protein abundance, of up to 3 to 4 orders of magnitude. Our own data show marked preservation of CpG
587 dinucleotides, which are crucial for mRNA structure, that appear to be independent of tRNA abundance.
588

589 **CONCLUSION**

590 We have shown that sSNVs which stabilize or destabilize mRNA are significantly constrained in
591 the human population, thereby supporting a growing body of evidence that previously assumed "silent"
592 polymorphisms, actually play crucial roles in regulation of gene expression and protein function. We have
593 demonstrated that this connection is rich, complex, and biologically intuitive. Given that there are multiple
594 mechanisms by which sSNVs influence biological function, we are almost certainly missing undiscovered
595 disease etiologies when these variants are ignored. In addition to providing the community with a dataset
596 of ten Vienna RNA structural metrics for every known synonymous variant, our Structural Predictivity
597 Index is the first metric of its kind to enable global assessment of sSNVs in human genetic studies. We
598 hope that these metrics will be utilized to accurately assess and prioritize an underrepresented class of
599 genetic variation that may be playing significant and as yet to be realized role in human health and disease.
600
601

602 **METHODS**

603 ***Raw Dataset***

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

607

608 ***Overview of RNA structure prediction process***

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

619

620 ***RNA structure prediction methodology***

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

625 ViennaRNA's flagship module RNAfold, which yielded three predicted mRNA secondary structures –
626 the minimum free energy, centroid, and maximum expected accuracy structure – as well as numeric values
627 for the free energy of each structure, and a fourth metric measuring the energy of the whole ensemble (see
628 the documentation of (Lorenz *et al.* 2011) for detailed descriptions of these concepts). Comparing the free
629 energies between the wildtype and mutant for each type of structure gave us the four stability metrics
630 delta-MFE (dMFE), dCFE, dMEAFE and dEFE. Next, the predicted structures were processed by the
631 Vienna module RNApdist, which counted the edge-differences to produce the four edge-metrics MFEED
632 (minimum free energy edit distance), CFEED, MEAED and EFEED. As a final step, the predicted
633 structures were further processed by the Vienna program RNAdistance to obtain the diversity metrics dCD
634 and dEND (change in distance from centroid and change in ensemble diversity, respectively).
635 This whole procedure was carried out using custom developed Spark wrappers of RNAfold,
636 RNApdist and RNAdistance, with slight modifications to the source code to suppress the creation of
637 graphics files. After building our fasta files, we were able to compute all 10 Vienna metrics for over a half
638 billion sequences in less than 24 hours using 51 c4.8xlarge AWS EMR computing nodes.
639
640 ***Construction of final dataset for synonymous SNVs***
641 The next step was to extract the sSNVs. This task was complicated by the fact that a SNV might
642 have appeared in several different transcripts, and could be synonymous in some and non-synonymous in
643 others. To address this challenge, we first annotated every SNV using the program snpEff (Cingolani *et*
644 *al.* 2012), whose source code was modified to allow record-by-record calling via Spark. This snpEff
645 analysis produced annotations of predicted biotype, e.g. missense, synonymous, canonical splice site, etc.
646 To validate these snpEff predictions we manually predicted the biotype of each SNV using start and stop
647 codon information from RefSeq

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

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

657

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

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

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

667

668 ***Further variant annotations***

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

671 AT's in the surrounding three windows. Finally we joined multiple additional annotations (including
672 conservation metrics such as PhyloP) from the dbNSFP dataset (Liu *et al.* 2016). Again, this heavy task
673 was greatly facilitated by our Spark framework.
674
675 ***Partition of dataset***
676 We carried out most of the analysis separately on subsets of data defined by a common mRNA
677 reference and alternate allele, e.g those sSNVs of form C>A. The reference and alternate alleles exert such
678 a huge influence on gnomAD that best solution seemed to be to control for them explicitly. Dividing our
679 dataset based on mRNA alleles (as opposed to DNA alleles, which do not depend on transcript sense) is a
680 step justified in **SUPPLEMENTARY TABLE 3**.
681
682 ***Depleter variables***
683 Depleter variables (so called because they explain some of the gnomAD depletion at values of a
684 structural variable) are given in **TABLE 1**. They are chosen to be the sequence feature that explains the
685 greatest portion of the connection between a structural metric (e.g. dMFE) and Y in a context. Possible
686 Depleter variables are local nucleotide content and the specific nucleotides up/downstream of the sSNV.
687 To compute the correlation between a structural metric (e.g. dMFE) and Y that is left unexplained
688 by a sequence feature (e.g. CpG content) in a particular REF-ALT context, we first build a simple logistic
689 regression model between CpG content and Y, which gives us an estimate $\mathbf{P}(\mathrm{Y} = 1 \mid \mathrm{CpG\ content})$ for
690 every sSNV in the context (based on the proportion of CpGs in the surrounding 120 nucleotides.) We then
691 plug this "structure-less" estimate into the expression

692 $$\mathrm{V}_{\mathrm{CpG\ content}} = \sum_{x} \mathbf{n_x} * (\mathbf{E}(\ \mathbf{P}(\mathrm{Y} = 1 \mid \mathrm{CpG\ content}) \mid \mathrm{dMFE} = \mathrm{x}) - \mathbf{P}(\mathrm{Y} = 1 \mid \mathrm{dMFE} = \mathrm{x}))^2$$

693 where we sum over all values x of dMFE and let $\mathbf{n_x}$ denote the number of sSNVs in the context with

694 dMFE= x. Comparing this quantity to the null variance

695 $$V_{null} = \sum_{x} \mathbf{n_x} * (\mathbf{P}(Y = 1) - \mathbf{P}(Y = 1 \mid dMFE = x))^2$$

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

697 $$R^2_{CpG\ content} = 1 - \frac{V_{CpG\ content}}{V_{null}}$$

698 The "Depleter" for a given structural metric in a given context is chosen as the variable with the

699 highest R^2. Finally the correlation between the Depleter and Y was checked, and the Depleter given a sign

700 (+/-) so that the signed Depleter correlated negatively with Y.

701

702 ***Construction of SPI***

703 To construct our final SPI scores we built two separate models over each of our 14 contexts to

704 predict the event MAF > 0. The "null" model used all natural features - the nine nucleotides in the SNV's

705 home and adjacent codons, the proportion of A/C/G/T/CpG/AT's in the surrounding 120 nucleotides, the

706 sSNV's position in its transcript and the transcript's length, and the tAI (tRNA Adapation Index obtained

707 from a supplement of (Tuller *et al.* 2010) from https://ars.els-cdn.com/content/image/1-s2.0-

708 S0092867410003193-mmc2.xls) of the wildtype and mutant codons. The second, "active" model used all

709 these features plus our 10 Vienna metrics. Both sets of variables were then used to predict MAF > 0. We

710 then defined the SPI score for a sSNV to be the base-10 logarithm of the active model's predicted P(Y=1)

711 probability divided by the null model's predicted P(Y=1). Context wise plots and statistics for SPI are

712 given in the **SUPPLEMENTARY FIGURE 6**.

713 We tried three different model-styles for computing the raw predictions that comprise SPI –

714 general logistic as implemented in python's sklearn LogisticRegression module, random forest as

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

720 **COMPETING INTERESTS**

721 The authors declare no competing interests.

722

723 **AUTHOR CONTRIBUTIONS**

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

730

731 **ADDITIONAL FILES**

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

734

735 **ACKNOWLEDGEMENTS**

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

739

740 **FIGURE LEGENDS**

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

757

758 **FIGURE 2. Graphical depiction of computational workflow used to generate ViennaRNA**
759 **folding metrics for the entire transcriptome.** The entire analysis workflow was parallelized using
760 Apache Spark and the Amazon Elastic Map Reduce (EMR) service, generating 5 billion Vienna RNA
761 metrics over the course of 2 days. Using a custom pipeline developed for the process that was executed
762 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

786 dMFE were determined for (**A**) CpG>CpA and (**B**) CpG>TpG synonymous mutations. dMFE values with
787 fewer than 75 nonzero-MAF sSNVs are excluded. Color gives average CFEED in each context, ranging
788 from 15 (blue) to 50 (red). Similarly, proportion of synonymous CpG transitions with nonzero MAF at
789 each value of CFEED were determined for (**C**) CpG>CpA sSNVs and (**D**) CpG>TpG sSNVs. Color
790 represents average dMFE and ranges from -0.8 (blue) to 1.85 (red). CFEED values with fewer than 75
791 nonzero-MAF sSNVs are excluded). Finally, proportion of synonymous CpG transitions with nonzero
792 MAF at each value of dCD (after rounding to nearest integer) were determined for (**E**) CpG>CpA and (**F**)
793 CpG>TpG sSNVs sSNVS. Color represents average dMFE and ranges from -3 (blue) to 4 (red). Rounded
794 dCD values with fewer than 75 nonzero-MAF sSNVs are excluded.
795
796 **FIGURE 5. SPI score correlates with constraint in synonymous CpG transitions.** Variants in
797 the contexts **(A)** CpG>CpA and **(B)** CpG>TpG are divided by SPI score into 20 equal bins and the value
798 P(MAF>0) plotted against the mean of each bin. We also colored by the mean dMFE over each bin. In
799 both contexts the constraint is highest towards negative SPI, i.e. sSNVs for which structural information
800 decreases the predicted probability that MAF > 0.
801
802 **TABLE 1. Structural metrics correlate with gnomAD frequency in most REF>ALT contexts.**
803 Correlation between structural metrics (**A**) dMFE, (**B**) CFEED and integer-rounded (**C**) dCD on the one
804 hand, and the quantity P(MAF>0) on the other, over all sSNVs in a given context. The R^2 and p-values
805 are obtained from a weighted least-squares linear regression, with the p-value corresponding to the linear
806 coefficient; a quadratic regression was also performed, but only the p-value was retained. Only context-
807 metric pairs with p-value < 0.005 are included. "Normalized slope" was obtained by dividing slope of
808 regression line by average P(MAF>0) in the context and then multiplying by range covered by metric in

809 its central 90% of sSNVs. "Depleter" is raw sequence variable that explains largest proportion of structural
810 trend *in this context*, with sign adjusted to correlate negatively with gnomAD frequency. "Depleter R^2"
811 gives proportion of variance explained by Depleter (see *Depleter variables* in **RESULTS** for details).
812
813 **TABLE 2**. **Area under curve for SPI score.** SPI was used to discriminate MAF > 0 using a simple
814 logistic model with 5-fold cross-validation. Table shows area under curve for model, averaged over the 5 training
815 and testing sets.
816
817 **TABLE 3: Known sSNVs clinically implicated for structural pathogenicity are successfully**
818 **predicted to be pathogenic by our structural metrics.** dbSNP RS number and standardized SNP
819 annotations are provided, along with the genes official gene symbol and disease the sSNV has been
820 associated with. The absolute value of dMFE, CFEED, dCD and SPI are provided, along with the
821 percentile value of that score, computed over each context, in parentheses.

822 **REFERENCES**

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

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

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

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

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

833 Alfares A, Aloraini T, Subaie LA, Alissa A, Qudsi AA, Alahmad A, Mutairi FA, Alswaid A, Alothaim A, Eyaid W et al.
834 2018. Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of
835 whole-exome sequencing. *Genet Med* **20**: 1328-1333.

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

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

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

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

846 Brummer A, Hausser J. 2014. MicroRNA binding sites in the coding region of mRNAs: extending the repertoire of
847 post-transcriptional gene regulation. *Bioessays* **36**: 617-626.
848 Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA
849 secondary structure in mammals. *Genome Biol* **6**: R75.
850 Chaney JL, Clark PL. 2015. Roles for Synonymous Codon Usage in Protein Biogenesis. *Annu Rev Biophys* **44**: 143-
851 166.
852 Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for
853 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of
854 Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**: 80-92.
855 Cohen NM, Kenigsberg E, Tanay A. 2011. Primate CpG islands are maintained by heterogeneous evolutionary
856 regimes involving minimal selection. *Cell* **145**: 773-786.
857 Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human
858 genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**: e1001025.
859 Deciphering Developmental Disorders Study. 2017. Prevalence and architecture of de novo mutations in
860 developmental disorders. *Nature* **542**: 433-438.
861 Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, Bazile C, Lambert NJ, Van Nostrand EL, Pratt GA
862 et al. 2018. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell* **70**:
863 854-867 e859.
864 Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in Escherichia coli at
865 different growth rates. *J Mol Biol* **260**: 649-663.
866 Duan J, Antezana MA. 2003. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation.
867 *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.

891 Hanson G, Coller J. 2018. Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol*
892 **19**: 20-30.
893 Hegde M, Santani A, Mao R, Ferreira-Gonzalez A, Weck KE, Voelkerding KV. 2017. Development and Validation of
894 Clinical Whole-Exome and Whole-Genome Sequencing for Detection of Germline Variants in Inherited
895 Disease. *Arch Pathol Lab Med* **141**: 798-805.
896 Huang YF, Gulko B, Siepel A. 2017. Fast, scalable prediction of deleterious noncoding variants from functional and
897 population genomic data. *Nat Genet* **49**: 618-624.
898 Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. 2014. Exposing synonymous mutations. *Trends Genet*
899 **30**: 308-321.
900 Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D et
901 al. 2016. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J*
902 *Hum Genet* **99**: 877-885.
903 Katz L, Burge CB. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial
904 genes. *Genome Res* **13**: 2042-2051.
905 Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the
906 relative pathogenicity of human genetic variants. *Nat Genet* **46**: 310-315.
907 Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in
908 Escherichia coli. *Science (80-)* **324**: 255-258.
909 Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function
910 using the SIFT algorithm. *Nat Protoc* **4**: 1073-1081.
911 Lazrak A, Fu L, Bali V, Bartoszewski R, Rab A, Havasi V, Keiles S, Kappes J, Kumar R, Lefkowitz E et al. 2013. The
912 silent codon change I507-ATC->ATT contributes to the severity of the DeltaF508 CFTR channel
913 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.

1005 Wiewiorka MS, Messina A, Pacholewska A, Maffioletti S, Gawrysiak P, Okoniewski MJ. 2014. SparkSeq: fast,
1006 scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision.
1007 *Bioinformatics* **30**: 2652-2653.
1008 Workman C, Krogh A. 1999. No evidence that mRNAs have lower folding free energies than random sequences
1009 with the same dinucleotide distribution. *Nucleic Acids Res* **27**: 4816-4822.
1010 Worthey EA. 2017. Analysis and Annotation of Whole-Genome or Whole-Exome Sequencing Derived Variants for
1011 Clinical Diagnosis. *Curr Protoc Hum Genet* **95**: 9 24 21-29 24 28.
1012 Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, King DA, Ambridge K, Barrett DM,
1013 Bayzetinova T et al. 2015. Genetic diagnosis of developmental disorders in the DDD study: a scalable
1014 analysis of genome-wide research data. *Lancet* **385**: 1305-1314.
1015 Wright CF, FitzPatrick DR, Firth HV. 2018. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet*
1016 **19**: 253-268.
1017 Yang JR, Chen X, Zhang J. 2014a. Codon-by-codon modulation of translational speed and accuracy via mRNA
1018 folding. *PLoS Biol* **12**: e1001910.
1019 Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z et al. 2013. Clinical
1020 whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* **369**: 1502-1511.
1021 Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, Ward P, Braxton A, Wang M, Buhay C et al. 2014b. Molecular
1022 findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**: 1870-1879.
1023 Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M, Franklin MJ, Shenker S, Stoica I. 2012. Resilient
1024 distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the*
1025 *9th USENIX conference on Networked Systems Design and Implementation*, pp. 2-2. USENIX Association,
1026 San Jose, CA.
1027

Figure 1

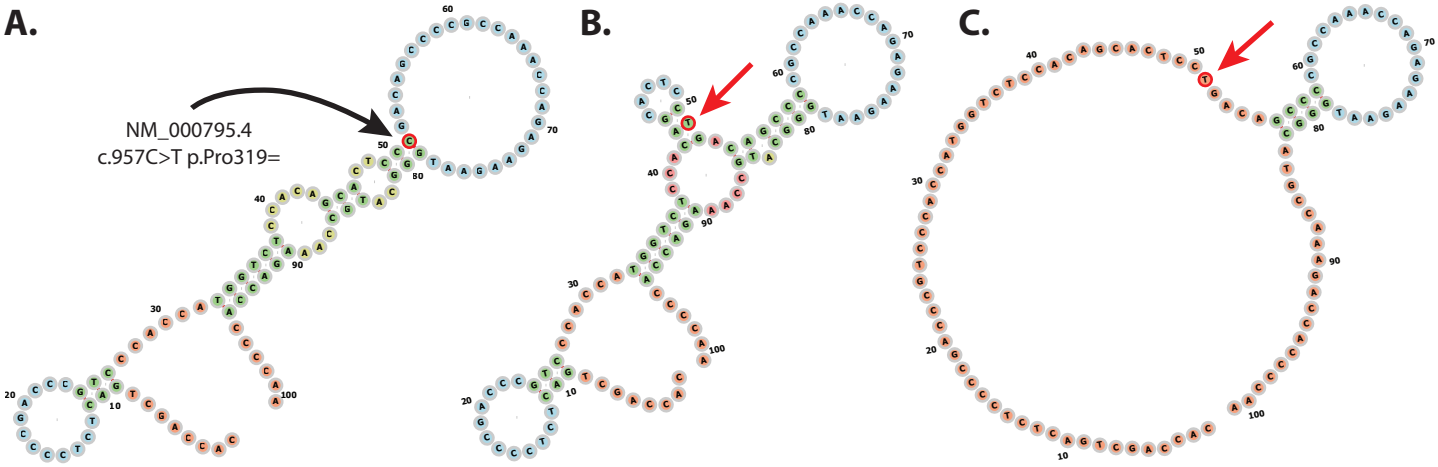

Figure 2

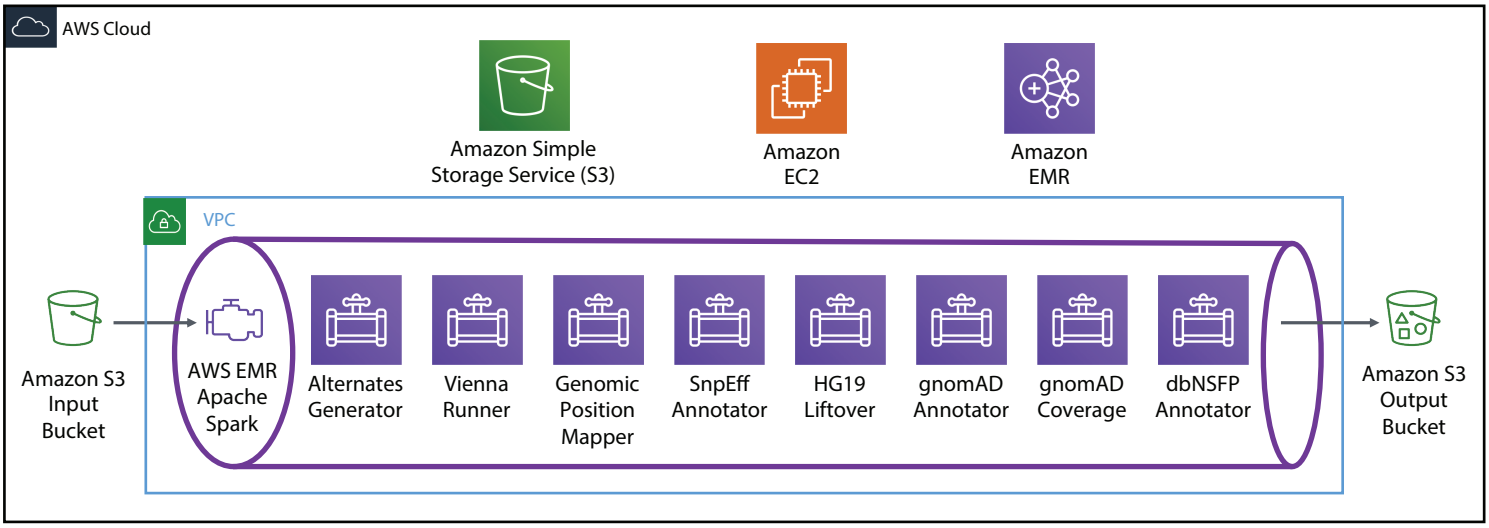
AWS Cloud
Amazon Simple Storage Service (S3)
Amazon EC2
Amazon EMR
VPC
Amazon S3 Input Bucket
AWS EMR Apache Spark
Alternates Generator
Vienna Runner
Genomic Position Mapper
SnpEff Annotator
HG19 Liftover
gnomAD Annotator
gnomAD Coverage
dbNSFP Annotator
Amazon S3 Output Bucket

Figure 3

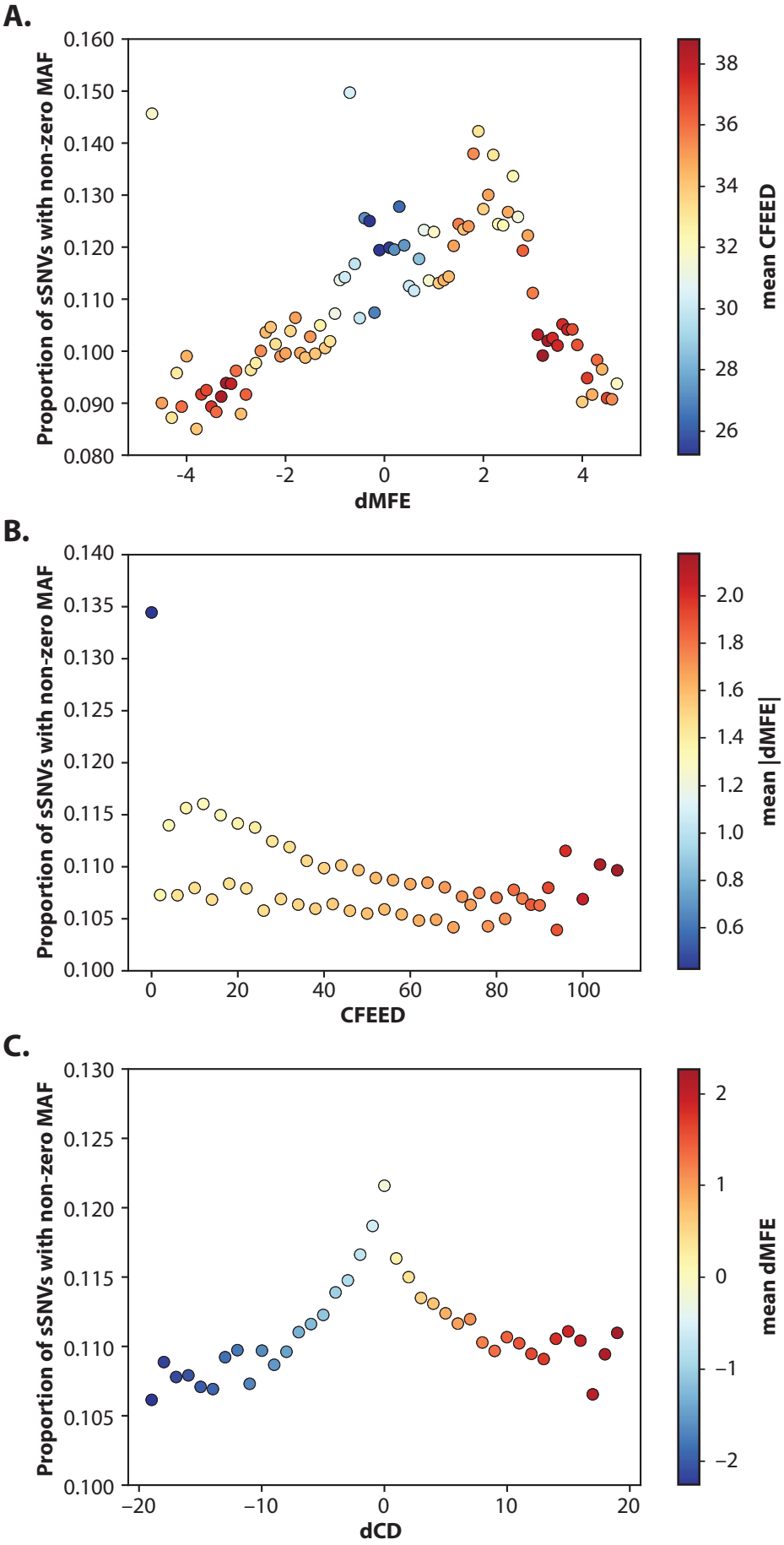
A.
Proportion of sSNVs with non-zero MAF
0.160
0.150
0.140
0.130
0.120
0.110
0.100
0.090
0.080
-4
-2
0
2
4
dMFE
38
36
34
32
30
28
26
mean CFEED
B.
Proportion of sSNVs with non-zero MAF
0.140
0.135
0.130
0.125
0.120
0.115
0.110
0.105
0.100
0
20
40
60
80
100
CFEED
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
mean |dMFE|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

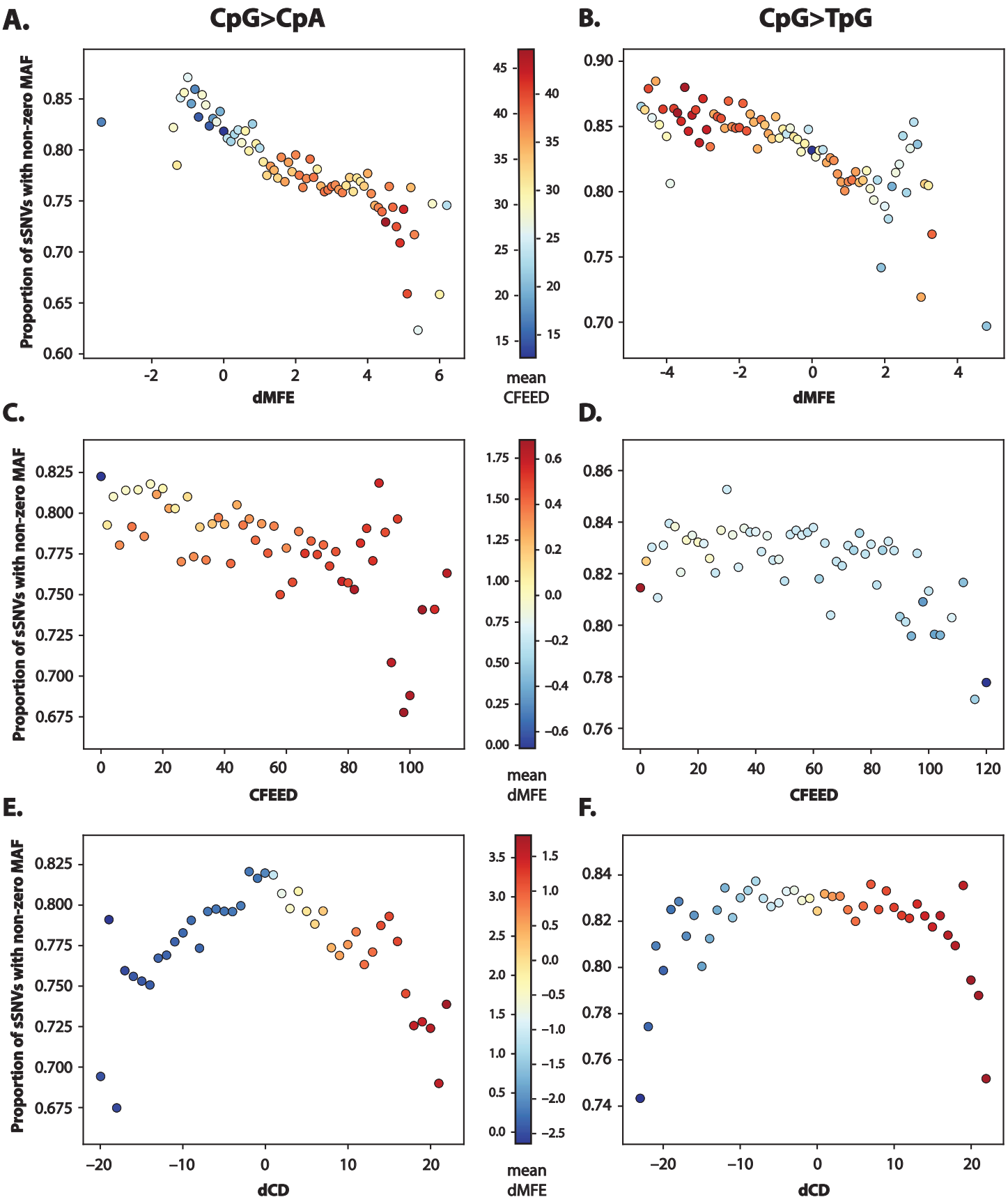
A.
CpG>CpA
B.
CpG>TpG
Proportion of sSNVs with non-zero MAF
dMFE
mean CFEED
C.
D.
CFEED
mean dMFE
E.
F.
dCD

Figure 5

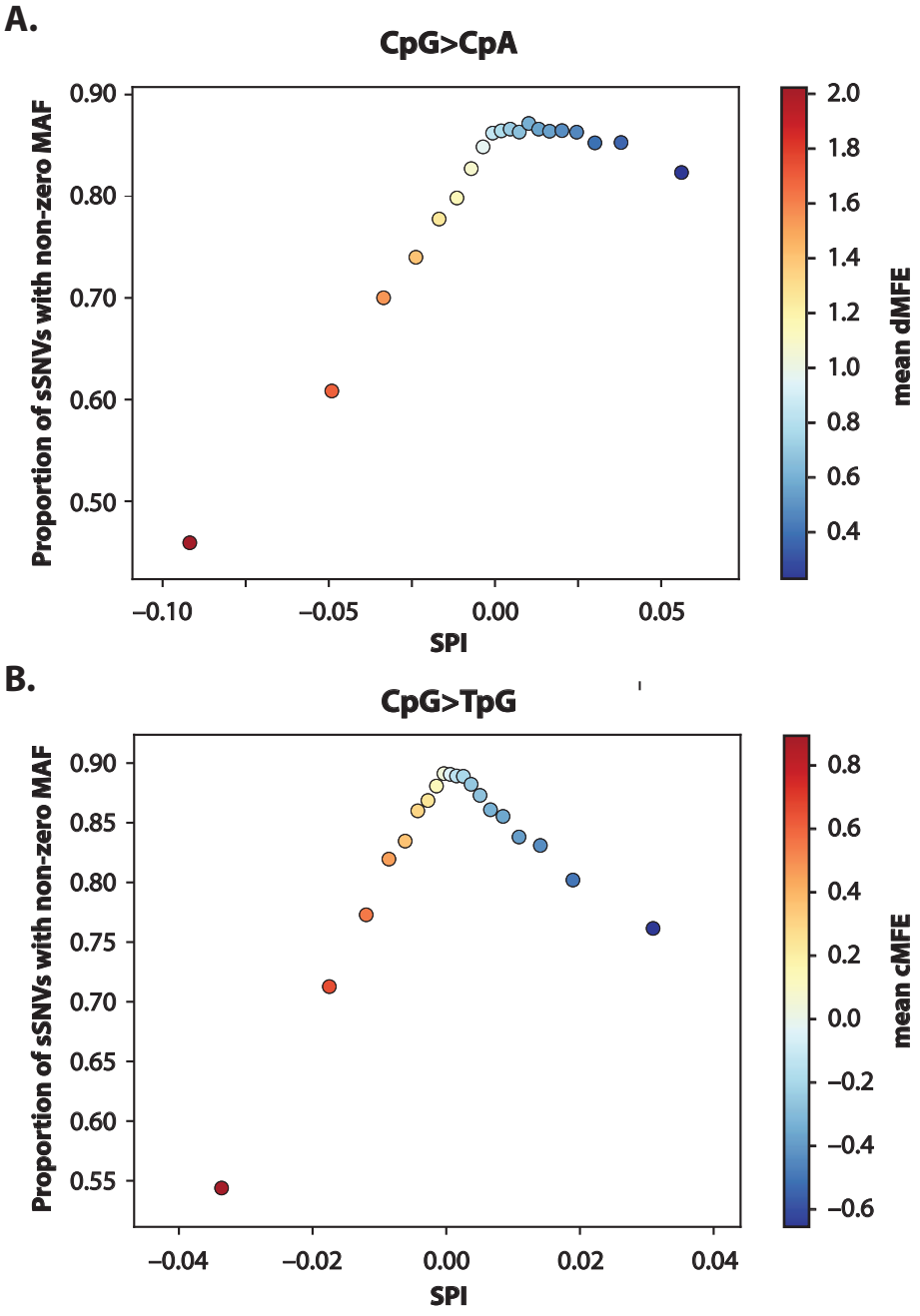
A.
CpG>CpA
Proportion of sSNVs with non-zero MAF
0.90
0.80
0.70
0.60
0.50
−0.10
−0.05
0.00
0.05
SPI
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
mean dMFE
B.
CpG>TpG
Proportion of sSNVs with non-zero MAF
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
−0.04
−0.02
0.00
0.02
0.04
SPI
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6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

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)