Cancer-driving mutations are enriched in genic regions intolerant to germline variation

Dimitrios Vitsios[1,*,+], Ryan S. Dhindsa[2,3,4,*,+], Jonathan Mitchell[1], Dorota Matelska[1], Zoe Zou[1], Joshua Armenia[5], Quanli Wang[2], Ben Sidders[5], Andrew R. Harper[1], Slavé Petrovski[1,6,+]

[1]Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK
[2]Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Waltham, MA, USA
[3]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA
[4]Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, TX, USA
[5]Bioinformatics and Data Science, Research and Early Development, Oncology R&D, AstraZeneca, Cambridge, UK
[6]Department of Medicine, University of Melbourne, Austin Health, Melbourne, Victoria, Australia

*These authors contributed equally

+Correspondence: dimitrios.vitsios@astrazeneca.com (D.V.), ryan.dhindsa@astrazeneca.com (R.S.D.), slav.petrovski@astrazeneca.com (S.P.)

Abstract

1 Large reference datasets of protein-coding variation in human populations have allowed us to
2 determine which genes and genic sub-regions are intolerant to germline genetic variation. There
3 is also a growing number of genes implicated in severe Mendelian diseases that overlap with
4 genes implicated in cancer. Here, we hypothesized that mitotically mutable genic sub-regions
5 that are intolerant to germline variation are enriched for cancer-driving mutations. We introduce
6 a new metric, OncMTR, which uses 125,748 exomes in the gnomAD database to identify genic
7 sub-regions intolerant to germline variation but enriched for hematologic somatic variants. We
8 demonstrate that OncMTR can significantly predict driver mutations implicated in hematologic
9 malignancies. Divergent OncMTR regions were enriched for cancer-relevant protein domains,
10 and overlaying OncMTR scores on protein structures identified functionally important protein
11 residues. Finally, we performed a rare variant, gene-based collapsing analysis on an
12 independent set of 394,694 exomes from the UK Biobank and find that OncMTR dramatically
13 improves genetic signals for hematologic malignancies. Our web app enables easy visualization
14 of OncMTR scores for each protein-coding gene (https://astrazeneca-cgr-
15 publications.github.io/OncMTR-Viewer/).

Introduction

The availability of large-scale human genetic variation reference datasets has revolutionized our ability to identify disease-causing mutations (Karczewski et al., 2020; Wang et al., 2021). Through the effective process of natural selection, variants with severe clinical outcomes are generally depleted in these datasets. We and others have leveraged this paradigm to develop intolerance metrics that quantify the extent to which natural selection constrains germline variation in genes and genic-sub regions (Dhindsa et al., 2020; Petrovski et al., 2013; Samocha et al., 2014; Traynelis et al., 2017). These methods have proven invaluable in prioritizing which of the roughly 20,000 protein-coding variants observed in any given individual are most likely to contribute to disease. Interpreting variants in the context of cancer suffers from similar challenges as interpreting germline variation: cancer cells often carry thousands of somatic mutations, but only some of these drive the oncogenic process. Despite their success in prioritizing germline variants, population genetics-based approaches have yet to be applied in the context of distinguishing between somatic cancer driving mutations and neutral "passenger" mutations.

Many developmental disorder-causing germline mutations occur in essential genic subregions, leading to dysfunction of crucial cellular biology pathways. We postulated that if these same mutations arise mitotically later in life, they will not cause the same developmental disease due to more limited expression of the mutation but could have equally as profound impacts on cellular biology. Consistent with this, there are several examples whereby identical point mutations that cause severe developmental syndromes when mutated in the germline cause cancer when mutated somatically (Hoischen et al., 2014; Petrovski et al., 2016), including identical mutations in *PTEN, ASXL1* (Hoischen et al., 2011)*, EZH2* (Gibson et al., 2012)*,* and others (Kaplanis et al., 2020). Many of these genes are involved in cell proliferation, chromatin remodeling, genome maintenance, and signal transduction pathways. This convergence highlights a subset of genes in the human genome that are crucial to cell biology, whereby disruptive mutations can cause different clinical outcomes depending on their timing, localization, and cellular context.

Here, we hypothesized that regions of genes that are under strong negative selection for germline variation but are exceptionally mitotically mutable would be enriched for variants that increase cancer risk. Identifying germline-constrained but mitotically mutable genic subregions could help prioritize cancer-driving mutations. Here, we focus on missense variants as they are the most observed protein-coding variant class, are becoming increasingly clinically actionable (Hyman et al., 2017), but importantly are also more difficult to interpret than protein-truncating

50   annotated variants. We previously introduced the missense tolerance ratio (MTR), a sliding

51   window-based approach that detects genetic sub-regions depleted of missense variation

52   (Traynelis et al., 2017). In this study, we extended this method to produce a score (OncMTR) to

53   identify germline intolerant but mitotically mutable genic sub-regions by using exome data from

54   125,748 individuals in GnomAD (Karczewski et al., 2020). We demonstrate that OncMTR

55   effectively predicts driver mutations of hematologic malignancies. We also use 394,694 UK

56   Biobank exomes to illustrate the utility of OncMTR in prioritizing variants in genetic discovery for

57   cancer phenotypes. This work introduces a population genetics approach to identify genic

58   subregions enriched for cancer-related somatic missense mutations.

59

60

Results

Putative somatic variants in gnomAD

63   Population-level catalogues of human genetic variation allow for the investigation of selective

64   constraint and mutational patterns in the exome. We used the gnomAD database of 125,748

65   human exomes to survey both germline and somatic variants (Karczewski et al., 2020).

66   Although the gnomAD variant calling pipeline was tuned to detect germline variation, we

67   reasoned that we may also be able to identify somatic variants that reach a sufficiently high

68   variant allele frequency to be detected through their germline variant caller. Inherited

69   heterozygous germline variants are expected to have an allelic ratio close to 50%. We observed

70   that the distribution of median allelic balance (AB_median) values for gnomAD variants followed

71   a bimodal distribution, with one distribution centered around 50% and another, smaller

72   distribution centered around 20% (**Fig. 1A**).

73

Defining OncMTR

75   We previously introduced a sliding window-based metric, the missense tolerance ratio (MTR),

76   that measures purifying selection on missense variation in genic sub-regions (Traynelis et al.,

77   2017). This score demonstrably detects crucial functional domains of proteins that can cause

78   Mendelian disease when mutated in the germline. Motivated by the overlap between mutations

79   associated with Mendelian disease and cancer, we set out to create a cancer-relevant version

80   of MTR (methods) that captures regions that are depleted of germline variation but also

81   enriched for somatic variation. In this study, we defined another variation of the MTR score,

82   namely $MTR^{germline}$. In its construction, $MTR^{germline}$ is restricted to only those variants achieving

83   an AB_median > 0.3. Taking the well-known cancer gene *TP53* as an example, we can observe

84   those genic subregions where the two MTR formulations diverged (**Fig. 1B**). We then define

85   OncMTR as the difference between these two MTR formulations for each codon and using a 31-

86   codon sliding window (**Fig. 1B**). Negative scores correspond to regions with the greatest

87   divergence between germline intolerance and somatic variant enrichment. Overlaying OncMTR

88   scores on the AlphaFold-predicted structure of TP53 (Jumper et al., 2021) illustrated that the

89   strongest negative scores correspond to the DNA-binding domain, which is the domain enriched

90   for mutations known to drive hematologic malignancies (**Fig. 1C**).
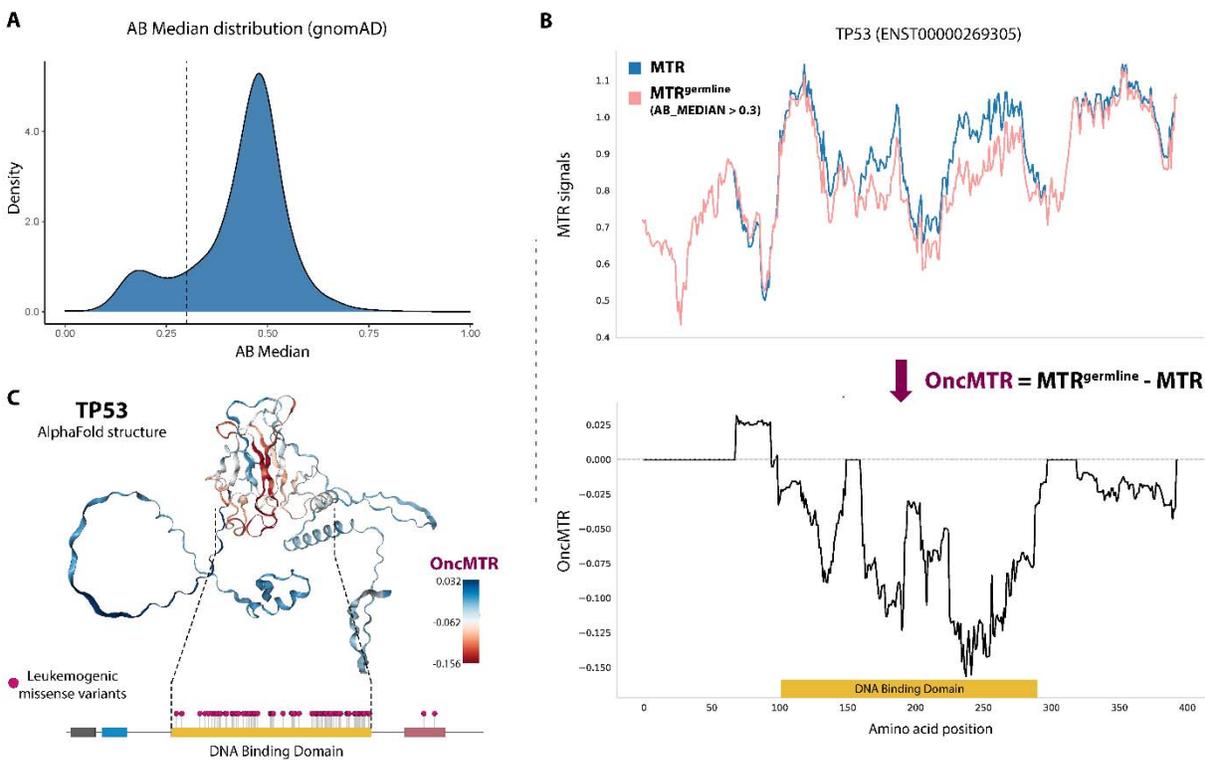
91

92   **Figure 1**. **Defining the OncMTR score**. **(A)** Bi-modal distribution of median allelic
93   balance values for heterozygous variants in the gnomAD database. We defined putative
94   somatic variants as those with an AB median ≤ 0.3 (dashed line). **(B)** The top figure
95   demonstrates the missense tolerance ratio (MTR) distribution of *TP53* when considering
96   all missense variants (blue) and when restricted to only germline variants (i.e., AB
97   Median > 0.3, depicted in pink). We defined OncMTR as the difference between these
98   two distributions (bottom panel). **(C)** OncMTR scores overlaid on the AlphaFold structure
99   for TP53. The most intolerant region maps to the DNA-binding domain of the protein,
100   which is strongly enriched for mutations known to drive hematologic malignancies.

101
102

103   **Using OncMTR to prioritize driver mutations in hematologic malignancies**

104   Motivated by the positive proof-of-concept demonstrated for *TP53*, we next tested whether the

105   MTR and MTR^{germline} distributions differed across other oncogenes included in the Catalogue of

106   Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (CGC). The CGC is divided into

107   two tiers, with Tier 1 containing *bona fide* cancer genes (n=556) and Tier 2 containing genes

108   that have strong indications of playing a role in cancer but with less expansive evidence than

109   Tier 1 (n=137). The difference between MTR and MTR^{germline} distributions per gene, calculated

110   via cross entropy, was significantly higher for Tier 1 genes than a random selection of 556 non-

111   CGC genes (p = 5.7x10^{-31}), the remainder of the exome (p = 2.8x10^{-67}), and Tier 2 genes (p =

112   1.1x10^{-7}) (**Fig. 2A**). The cross entropy was also significantly larger for Tier 2 genes than the

113   remaining genes in the exome (p = 2.6x10^{-4}) (**Fig. 2A**). Together, these results support the

114   hypothesis that mitotically mutable genic sub-regions that are intolerant to germline variation are

115   broadly relevant to cancer.

116
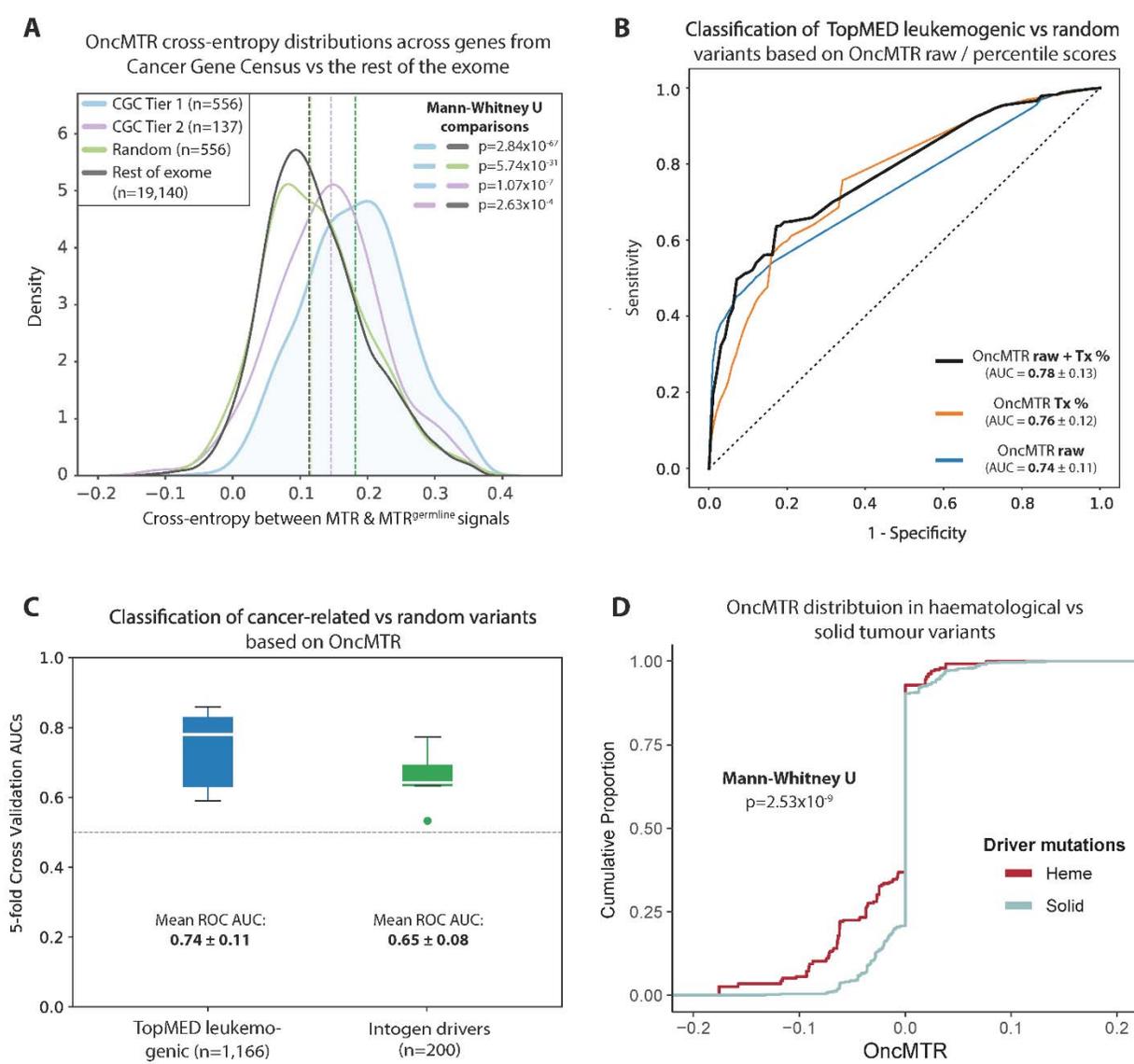
117

118   **Figure 2. OncMTR regions are enriched for somatic variants associated with**
119   **hematologic malignancies. (A) Cross entropy between the** distribution MTR and
120   MTR^{AB} distributions for Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer
121   Gene Census (CGC) genes, a random selection of genes, and the rest of the exome. **(B)**
122   Receiver operator curve (ROC) depicting the ability of random forest models based on
123   either the raw OncMTR score, the OncMTR transcript-level percentile scores ("Tx%"),
124   and a joint model in discriminating between 1,166 leukemogenic variants and a random
125   size-matched set of variants. AUC = area under the curve. **(C)** Mean ROC AUCs (with
126   fivefold cross-validation) of random forest models based on raw OncMTR in predicting
127   variants involved in leukemia (same variant set as figure B) and hematologic driver
128   mutations annotated in IntoGen (Tamborero et al., 2018). The putatively neutral variant
129   sets comprise of random, size-matched selection of variants. (**D**) The OncMTR
130   distributions of driver mutations for hematologic malignancies versus solid tumors is
131   derived from the Cancer Genome Interpreter.

132
133

134   Distinguishing between cancer-causing driver mutations and neutral passenger mutations

135   remains a central challenge in cancer genomics. We thus tested whether OncMTR could help

136   prioritize somatic mutations that cause hematologic malignancies. We found that the OncMTR

137   scores of a previously defined list of 1,166 leukemogenic driver mutations (Bick et al., 2020)

138   (**Supplementary Table 1**) were significantly lower than a size-matched set of random variants

139   (Mann Whitney U p=2.97x10^{-86}; **Supplementary Fig. 1A**). A random forest model using

140   OncMTR achieved an area under the receiving operator curve (AUC) of 0.74 in discriminating

141   between these leukemogenic variants and the random set (**Fig. 2B**). We also calculated

142   transcript-level percentiles for the MTR scores, in which lower percentiles corresponded to lower

143   OncMTR scores. The AUC or the OncMTR transcript percentiles was 0.76, and a combined

144   model that incorporated both the raw OncMTR scores and transcript percentiles achieved an

145   even higher AUC of 0.78 (**Fig. 2B**).

146   To further assess the capacity of OncMTR to prioritize driver mutations, we trained

147   random forest models with raw OncMTR scores using 5-fold cross-validation. The mean AUC

148   for predicting leukemogenic variants was 0.74 (**Fig. 2C**). We next compared the performance of

149   OncMTR in distinguishing between a set of random variants and 200 established driver

150   mutations implicated in acute lymphocytic leukemia (ALL), acute myeloid leukemia (AML),

151   chronic lymphocytic leukemia (CLL), diffuse large B-cell lymphoma (DLBCL), or multiple

152   myeloma (MM), achieving an AUC of 0.65 (**Fig. 2C**) and having significantly disparate OncMTR

153   distributions from each other (Mann Whitney U p=4.89x10^{-5}; **Supplementary Fig. 1B** and

154   **Supplementary Table 2**). Logistic regression-based classifiers achieved similar, albeit

155   marginally lower, AUCs than the random forest models (with AUCs of 0.73 and 0.62 for the two

156   variant sets, respectively), likely due to a small degree of non-linear distribution of OncMTR

157   scores (**Supplementary Fig. 2**). Altogether, these results demonstrate the utility of our

158   population genetics-based approach in identifying genic sub-regions relevant to hematologic
159   malignancies.

160   Because the somatic mutations used to calculate OncMTR arose in the blood, we
161   expected that OncMTR would more reliably prioritize driver mutations in hematologic
162   malignancies than in solid tumors. As expected, the OncMTR scores of driver mutations
163   implicated in heme malignancies were significantly lower (Mann Whitney U p=2.53x10^{-9}; **Fig.**
164   **2D**). To determine whether OncMTR performs better for certain subtypes of heme malignancies,
165   we compared OncMTR distributions of putative driver and passenger mutations identified in a
166   recent comprehensive *in silico* saturation mutagenesis experiment (Muiños et al., 2021). This
167   dataset includes simulated variants across 3 genes for CLL, 9 genes for AML, 2 genes for non-
168   Hodgkin lymphoma, 5 genes for lymphoma, 6 genes for multiple myeloma, and 2 genes for ALL
169   (**Supplementary Table 10**). The OncMTR scores of predicted driver mutations were
170   significantly lower than those of passenger mutations for each cancer subtype, though we
171   observed the strongest separation in CLL (Wilcoxon p<2x10^{-308}) and AML (Wilcoxon p=1.4x10^{-155}
172   ) (**Supplementary Fig. 3**).

173   We next assessed whether OncMTR can successfully distinguish between ClinVar
174   pathogenic and benign somatic variants. Logistic regression classification between pathogenic
175   and benign or random variants across all protein-coding genes reached an AUC of 0.60 and
176   0.58, respectively (**Supplementary Fig. 4**; P=815 unique pathogenic vs B=58 unique benign
177   variants; a set R [random] of equal size to P was sampled to compile the random variants - see
178   also Methods). We next restricted the set of pathogenic somatic variants to those occurring in
179   genes associated with hematologic malignancies and compared to benign or random variants.
180   The AUC was 0.62 in distinguishing between pathogenic and benign variants in hematologic
181   malignancy genes (P=64 vs B=20) and 0.67 when comparing to benign variants across the
182   entire exome (P=64 vs B=58). The AUCs for pathogenic hematologic malignancy variants
183   versus random variants were 0.61 for random variants restricted to heme genes (P=64 vs R=64)
184   and 0.64 for random variants pulled from all protein-coding genes (P=815 vs R=815)
185   (**Supplementary Fig. 4**). These results provide support to this blood-based sequencing version
186   of OncMTR being more powerful in identifying pathogenic mutations implicated with heme
187   malignancies.

188   Finally, to further explore OncMTR's power to agnostically detect putative oncogenic
189   regions, we scanned all protein-coding genes in ClinVar in search of transcripts that are
190   preferentially enriched for ClinVar pathogenic somatic variants in regions with OncMTR scores
191   at the bottom 20-percentile of the full OncMTR distribution (see Methods). We identified 101

192 such transcripts from 24 unique genes (Fisher's exact test p<0.05; **Supplementary Table 11**),

193 with several known cancer driver genes captured, such as *TP53*, *IDH1*, *ALK* and *HNRNPA1*

194 (Martínez-Jiménez et al., 2020). Many of the top ranked genes are implicated in hematologic

195 malignancies, including *MYC*, *MSH2*, and *FBXW7* (**Supplementary Fig. 5**) (Bhatia et al., 1993;

196 King et al., 2013; Whiteside et al., 2002).

197

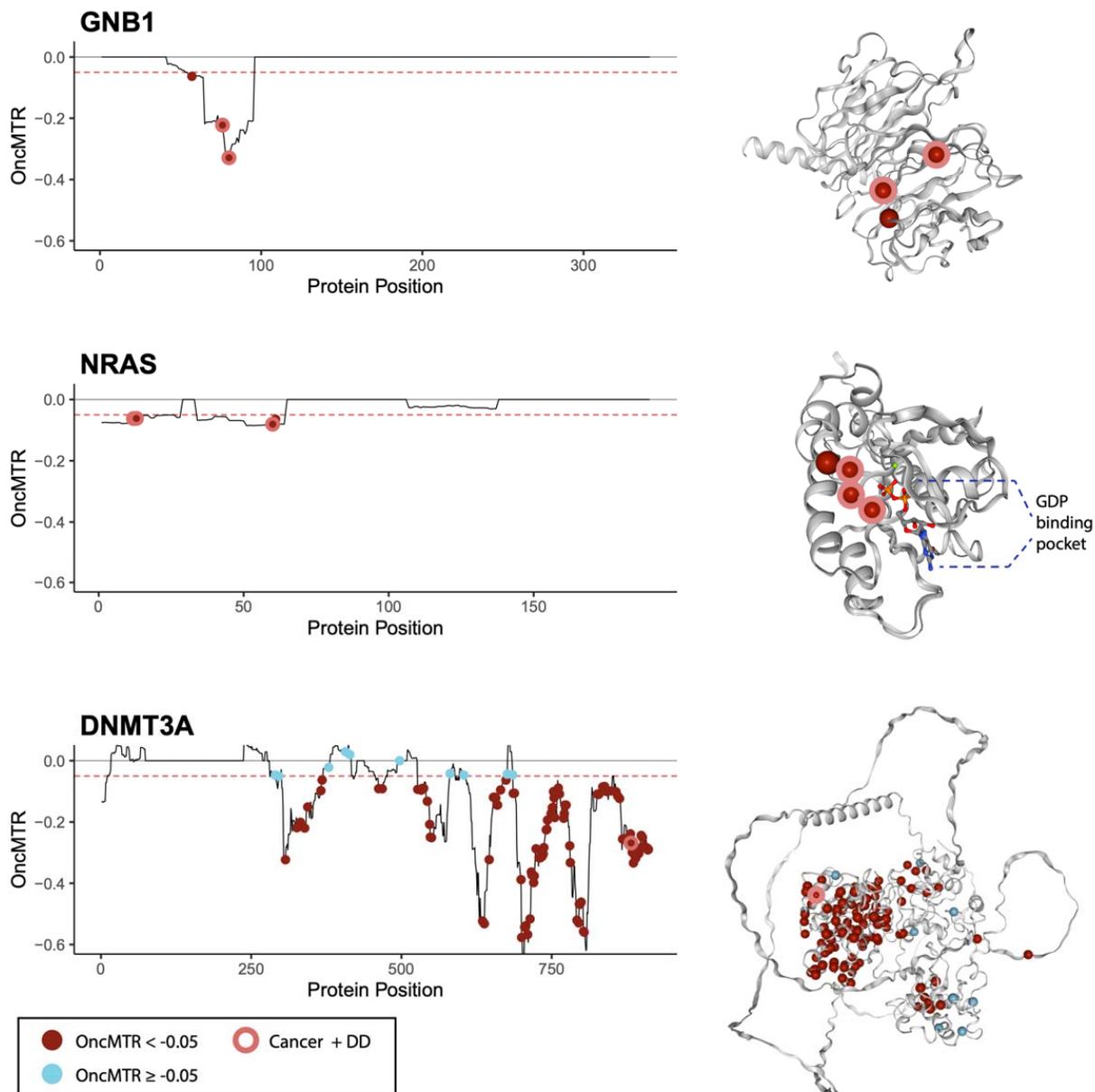198 **Genes carrying mutations implicated in both human Mendelian disease and cancer**

199 The underlying hypothesis in deriving OncMTR is that certain genic regions are critically

200 important to human biology, and thus germline mutations in these regions cause severe

201 Mendelian phenotypes, whereas identical somatic mutations–occurring later in life and localized

202 to specific tissue(s)–in these regions may have an oncogenic effect. To evaluate this, we plotted

203 OncMTR distributions for three genes implicated in both neurodevelopmental disease and

204 leukemia: *GNB1, NRAS*, and *DNMT3A* **(Fig. 3 A-C** and **Supplementary Table 4**).

205 Germline de novo mutations in *GNB1* cause a severe developmental syndrome

206 characterized by intellectual disability (ID) and other variable features, including hypotonia,

207 seizures, and poor growth (Petrovski et al., 2016). Somatic mutations in this gene have been

208 associated with ALL, CLL, and myelodysplastic syndrome (Yoda et al., 2015). Three of the four

209 somatic driver mutations in this gene overlap with *de novo* mutations implicated in

210 developmental delay (p.Asp76Gly, p.Ile80Thr, and p.Ile80Asn) (**Fig. 3A**) (Petrovski et al., 2016).

211 All four mutated residues reside in a low OncMTR region (OncMTR < -0.05) of the gene, which

212 corresponds to the Gβ-protein surface that interacts with Gα subunits and downstream effectors

213 (**Fig. 3A**).

214 *NRAS* encodes a RAS protein with intrinsic GTPase activity that has been implicated in

215 multiple hematologic and solid malignancies (Oliveira et al., 2007). There are 28 somatic

216 missense variants in this gene at four distinct amino acid positions associated with juvenile

217 myelomonocytic leukemia and AML, and all residing in low OncMTR regions (**Fig. 3B**) (Bick et

218 al., 2020). Two of these mutations have also been reported as causal germline *de novo*

219 mutations for Noonan syndrome, a developmental delay syndrome that includes congenital

220 heart defects, short stature, and other features (p.Gly13Asp, p.Gly60Glu) (**Fig. 3B**) (Cirstea et

221 al., 2010; Matsuda et al., 2007).

222 *DNMT3A* encodes a DNA methyltransferase essential for DNA methylation during

223 human embryogenesis and, when mutated somatically, increases risk of acute myeloid

224 leukemia (Kosaki et al., 2017). In a large study on clonal hematopoiesis of indeterminate

225 potential (CHIP), *DNMT3A* was found to harbor the largest proportion of CHIP variants of all

226　CHIP-associated genes (Jaiswal et al., 2017), suggesting it is highly mitotically mutable. In line

227　with this, the OncMTR distribution of this gene is highly enriched for negative values, even

228　compared to *GNB1* and *NRAS* (**Fig. 3C**). The R882 amino acid residue of *DNMT3A*

229　corresponds to a DNA-binding residue that is a major somatic mutation hotspot in CHIP and

230　AML (Kosaki et al., 2017). *De novo* germline mutations at this residue are associated with an

231　overgrowth syndrome called Tatton-Brown-Rahman syndrome characterized by tall stature and

232　impaired intellectual development (Tatton-Brown et al., 2014). Mutations at the R882 residue

233　are thought to interfere with DNA binding, resulting in functional impairment of the protein and

234　aberrant DNA methylation patterns (Zhang et al., 2018). As expected, we identify that the

235　leukemogenic variants in this gene are enriched in low OncMTR regions (**Fig. 3C**). Altogether,

236　these results support the notion that some critically important genic sub-regions are

237　exceptionally mitotically mutable, and mutations in these regions result in different phenotypic

238　outcomes depending on timing and cellular context (Hoischen et al., 2011).

Figure 3. OncMTR distributions for genes implicated in both cancer and Mendelian disease. (A-C) OncMTR scores for *GNB1* (A)*, NRAS* (B), and *DNMT3A* (C) with corresponding protein structures from PDB (for NRAS, PDB ID: 6zio) or predicted by AlphaFold (Jumper et al., 2021). Points on the OncMTR plots and spheres on the protein structures indicate pathogenic somatic mutations included in TopMED leukemogenic variant set. Red points indicate variants with OncMTR < -0.05. Points with a pink outline indicate somatic leukomogenic variants that are also known to cause developmental delay (DD) when mutated *de novo* in the germline. *De novo* mutations were aggregated from the Online Mendelian Inheritance of Man database.

Enrichment of low OncMTR scores in protein domains

One strength of the sliding window approach implemented in OncMTR is that its estimates are independent of biological boundaries, such as annotated protein domains, which are not always

254   well-annotated. However, it is known that cancer-causing missense mutations tend to cluster in

255   certain functional domains. We thus tested whether Pfam domains and domain superfamilies

256   were enriched for low OncMTR regions (defined as OncMTR < -0.05). Across human protein-

257   coding genes, low OncMTR regions were significantly enriched for several protein domains

258   previously implicated in cancer, such as homeodomains (Fisher's exact adjusted p-

259   value=4.9×10^{-46}), protein kinase domains (Fisher's exact adjusted p-value=5.25×10^{-110}), RING

260   domains (Fisher's exact adjusted p-value=3.22×10^{-48}), and others (**Figure 4 A,B** and

261   **Supplementary Tables 5,6**). Furthermore, we found that proteins that had functional domains

262   enriched for low OncMTR scores are significantly enriched in genes with TOPMed

263   leukemogenic variants and known cancer hotspots (Chang et al., 2018) (**Figure 4C** and

264   **Supplementary Tables 1-3;7-9**). Among these two lists of genes, zinc finger motifs were found

265   to be the most strongly enriched for low OncMTR scores (**Figures 4D-F**; most significant

266   adjusted p-value=2.3×10^{-52} from the union list, based on Fisher's exact test), in line with their

267   well-established role in cancer development (Cassandri et al., 2017). Remarkably, although the

268   calculation of OncMTR is agnostic to domain annotations, it independently identifies cancer-

269   relevant functional genic sub-regions.

270

271
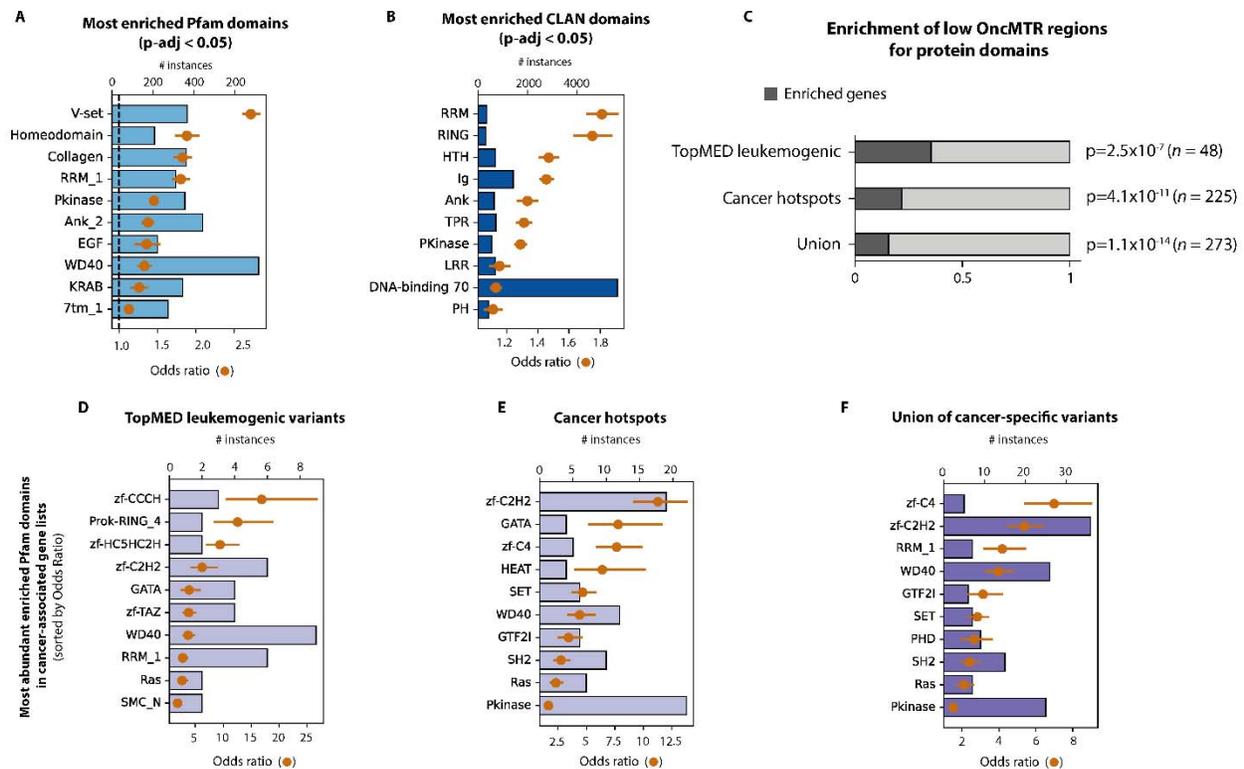
272   **Figure 4. Overlap between OncMTR regions and protein domains. (A)** Pfam protein
273   domains most strongly enriched with low OncMTR regions (OncMTR < -0.05). **(B)** Pfam
274   domain clans most strongly enriched with low OncMTR regions. The DNA-binding
275   superfamily set was defined in a prior publication (Bahrami et al., 2015). **(C)** Proportions
276   of genes enriched with low OncMTR scores in annotated protein domains in various
277   cancer-related gene sets: genes carrying TopMED leukemogenic variants, annotated
278   cancer hotspots, as well as the union of these three lists. **(D-F)** The most abundant Pfam
279   domains enriched with low OncMTR regions in proteins encoded by the labeled sets of
280   cancer genes. Error bars in each panel represent 95% confidence intervals. P-values
281   were calculated with Fisher's exact test and adjusted via Bonferroni correction.
282

283   **Informing rare-variant collapsing analysis with OncMTR**

284   With increasing adoption of next generation sequencing to generate case-control cohorts, rare

285   variant collapsing analysis has emerged as a powerful approach to detect disease-associated

286   genes for both rare and complex disorders. In this approach, the proportion of cases with a

287   qualifying variant is compared to the proportion of controls with a qualifying variant in the same

288   gene. We have previously shown that incorporating an MTR filter in defining QVs dramatically

289   improves rare variant collapsing analyses (Wang et al., 2021). In that phenome-wide

290   association study (PheWAS) on approximately 300,000 exomes in the UK Biobank, the

291   collapsing analyses detected seven genes associated with hematologic malignances (Wang et

292   al., 2021). Here, we sought to test whether OncMTR would further improve collapsing analysis

293   signals for hematologic malignancy associations by performing a collapsing analysis on 394,694

294   European exomes contained in the UK Biobank focused on 1,394 chapter IX (neoplasm)

295   phenotypes. We defined a total of eight collapsing models with and without OncMTR filters

296   (**Supplementary Table 12**). Imposing an OncMTR filter of -0.05 (i.e., only considering missense

297   QVs that fall below this threshold) significantly increased the effect sizes of gene-phenotype

298   associations (p < 0.0001) for each model (**Fig. 5A**). We observed genome-wide significant

299   ($p<1x10^{-8}$) associations between several heme malignancies and *DNMT3A, FBXW7, IDH2,*

300   *IGLL5, JAK2, SF3B1, SRSF2, TET2,* and *TP53*, in certain cases the effect sizes were 10-fold

301   greater than without adopting the OncMTR filter (**Fig. 5B**). We also found that the association

302   between *TP53* and CLL only reached significance in models including our OncMTR filter; for

303   example, in the 'raredmg' model, this association had a p-value of $1.2x10^{-7}$ (odds ratio [OR] =

304   8.8; 95% confidence interval [CI]: 4.8-16.0), whereas in the 'raredmgoncmtr' model, the same

305   association reached a p-value of $3.4x10^{-10}$ (OR = 33.2; 95%CI: 16.1-68.7). Thus, applying the

306   OncMTR filter effectively reduces background variation in the setting of gene-level collapsing

307   analysis for haematological malignancy phenotypes and we advise future large-scale

308   haematological malignancy discovery studies to consider adopting OncMTR filter for improved

309   signal detection.
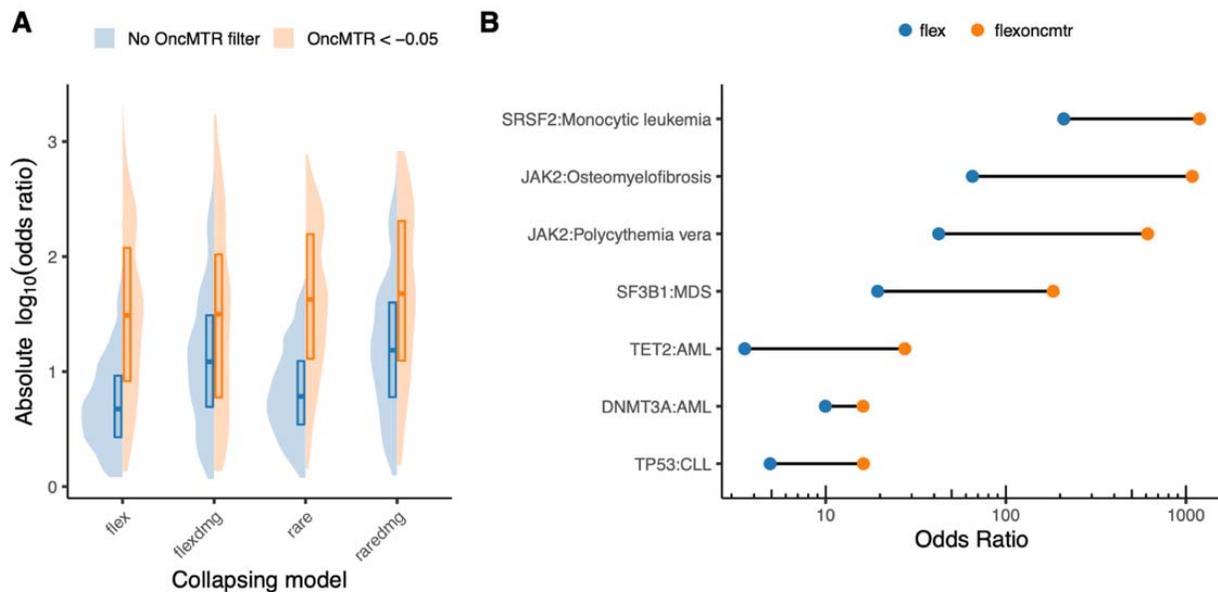
Figure 5. Collapsing analyses using OncMTR. (A) Effect sizes of gene-phenotype associations derived from a gene-level collapsing analysis performed on neoplasm phenotypes in 394,694 UK Biobank exomes. Collapsing models are defined in Supplemental Table 12. **(B)** Changes in odds ratios observed for selected gene-phenotype associations. MDS = myelodysplastic syndrome; AML = acute myeloid leukemia; CLL = chronic lymphocytic leukemia.

Discussion

Determining the clinical relevance of missense variants in oncogenes remains a central challenge in cancer genetics (Chang et al., 2018; Hyman et al., 2017). Motivated by the observation that missense variants in certain genic sub-regions can cause severe Mendelian disease when mutated in the germline and cancer when mutated somatically, we introduced a population genetics-based framework called OncMTR to quantitate the divergence between germline constraint and somatic mutability across the human exome.

First, we demonstrated that oncogenes are enriched for these critically important regions that do not tolerate germline missense variants but harbor somatic mutations. We then illustrated that OncMTR can effectively distinguish between leukemogenic driver mutations and passenger mutations. Although OncMTR is calculated using a sliding window without any input of domain annotations, we found that genic sub-regions that have low OncMTR scores are significantly enriched for protein domains known to be relevant to cancer. Illustrative of our hypothesis was the observation that identical point mutations implicated in both severe Mendelian disease and leukemia in the genes *GNB1, NRAS,* and *DNMT3A* occur in low OncMTR regions. Finally, we found that incorporating OncMTR in a gene-level collapsing

335   analysis on hematologic malignancy phenotypes using 394,694 UKB exomes improved the

336   signal-to-noise ratio for detecting hematologic malignancy associations. We have also

337   developed web server for visualization of OncMTR scores for each human protein-coding gene:

338   https://astrazeneca-cgr-publications.github.io/OncMTR-Viewer/.

339   Our findings have important implications for the disease biology of both severe

340   Mendelian disorders and cancer. The convergence of genes and genic sub-regions between

341   these two disease areas suggest that similar biological processes play a fundamental role in

342   these two groups of phenotypes. Indeed, cellular proliferation, chromatin remodeling, cell

343   migration, and other cancer-relevant processes have been implicated in neurodevelopmental

344   diseases (De Rubeis et al., 2014; Dhindsa et al., 2021; Feng et al., 2019; Kaplanis et al., 2020).

345   Furthermore, our work supports the notion that mutations in these genes have different

346   phenotypic manifestations based on timing (i.e., zygote versus adulthood), localization

347   (systemic versus hematological), and cellular context.

348   There exist many other approaches that aim to predict which genic sub-regions are

349   relevant to cancer. These methods tend to look for nonrandom clustering patterns of somatic

350   mutations in either the linear protein sequence or three-dimensional space (Porta-Pardo et al.,

351   2017). To the best of our knowledge, none of these approaches integrate population-level

352   inferences of genic constraint. OncMTR could improve the predictive performance of other,

353   orthogonal driver mutation prediction approaches, as a recent *in silico* saturation mutagenesis

354   experiment demonstrated the strength of incorporating multiple lines of evidence in prioritizing

355   driver mutations (Muiños et al., 2021).

356   One limitation of OncMTR in its current formulation is that it does not reflect the broader

357   range of solid tumor malignancies since it is based on somatic mutations observed in blood-

358   based sequencing. In future work, the general framework introduced in this study could be

359   applied to sufficiently large tumor-normal sequence datasets as those numbers increase.

360   Furthermore, we used gnomAD because it represents the largest collection of publicly available

361   aggregated allele frequency data. However, gnomAD variants were all called using a germline

362   variant caller. While we demonstrated that we could detect somatic variants in this database, we

363   were likely limited to those that reached a sufficiently high variant allele frequency to be

364   detected. Use of somatic variant callers adopted on these large-scale datasets could further

365   improve the sensitivity of OncMTR.

366

367

Methods

Reconstructing the Missense Tolerance Ratio with 125K samples from gnomAD

We first reconstruct the Missense Tolerance Ratio (MTR) using a cohort of 125,748 exomes from the gnomAD consortium (v2, GRCh38 liftover). The formula for deriving the window-based MTR scores has been introduced in the original paper (Traynelis et al., 2017):

$$MTR = \frac{\dfrac{missense\ (\boldsymbol{obs})}{missense\ (\boldsymbol{obs})\ +\ synonymous\ (\boldsymbol{obs})}}{\dfrac{missense\ (\boldsymbol{exp})}{missense\ (\boldsymbol{exp})\ +\ synonymous\ (\boldsymbol{exp})}}$$

where the numerator represents the observed proportion of missense variants among the total observed protein-coding variation. The numerator is then scaled by the same proportion computed from the collection of all possible protein-coding variants in the corresponding protein-coding window. A window size of 31 codons has been employed for calculating MTR based on the gnomAD cohort, in agreement with the previously published score (Traynelis et al., 2017).

The expected proportion of missense variants in a given protein-coding window was calculated by annotating all possible variants of a protein-coding transcript with SnpEff 4.3t using GRCh38.92 as the reference annotation and assuming all events were equally likely to occur. Annotation with SnpEff focused on single nucleotide variants (SNVs) that were flagged as PASS variants in the original gnomAD release (v2). Variants annotated as 'missense_variant' or 'missense_variant&splice_region_variant' by SnpEff represent the set of 'missense' variants in the MTR formula. Variants annotated as 'synonymous_variant', 'stop_retained_variant', 'splice_region_variant&stop_retained_variant' or 'splice_region_variant&synonymous_variant' by SnpEff were considered as the 'synonymous' variants in the same formula.

OncMTR score construction

Using MTR as our basis, we construct the OncMTR score (i.e. Oncology MTR score) to capture protein-coding subregions that are depleted of germline missense variants but observe somatic mutations. We observe that the total distribution of AB_MEDIAN values across all gnomAD variants (**Fig. 1A**) is bimodal, with the main peak centered close to 0.5 and a second one emerging for values approximately around 0.2. The AB_MEDIAN metric represents the allelic ratio between the alleles for each variant, with values close to 0.5 reflecting an equal number of

398   copies being inherited from each parent in heterozygous settings, while truly biological variants

399   that approach zero increasingly reflect variants that more likely arose somatically.

400

401   We leverage this observation to construct an alternative version of the original MTR score:

402   excluding any putative somatic variants and employing only germline variants from the gnomAD

403   dataset. We achieve that by selecting only variants with AB_MEDIAN > 0.3, constructing the

404     $MTR^{germline}$ version of the score. OncMTR is then defined as the difference of the original MTR

405   score from the $MTR^{germline}$ version:

406
$$OncMTR = MTR^{germline} - MTR$$

407   Negative OncMTR values (i.e. $MTR^{germline}$ < MTR) represent regions that are depleted of

408   germline variants and are instead enriched for somatic variation, thus allowing to highlight

409   putative oncogenic subregions in protein coding genes.

410

Compilation of variant sets

412   We used a pre-compiled set of variants known to be drivers of haematologic malignancies in a

413   total of 160 genes (Jaiswal et al., 2014). This list was generated from recurrent haematologic

414   somatic mutations in the literature and COSMIC, excluding genes with a relatively high

415   proportion of loss-of-function germline mutations. A second, smaller pre-compiled list, focused

416   on genes which were recurrent drivers specifically for myeloid malignancies (Bick et al., 2020).

417   A third validation set included a list of annotated driver mutations provided through the IntoGen

418   database (Tamborero et al., 2018). We restricted this set to "Tier 1" (highest confidence) driver

419   mutations observed in hematologic malignancies, which included ALL, AML, CLL, DLBC, and

420   MM.

421

Classification of oncogenic variant sets with OncMTR

423   We have performed classification of different oncogenic variant sets (TOPMed leukemogenic

424   and Intogen drivers) against random variant sets of equal size. We employ two supervised

425   models for the binary classification task: Logistic Regression with 'max_iter'=1000 and a

426   Random Forest classifier with 'max_depth'=2, to avoid overfitting on the training set. Each

427   classification was performed as a 5-fold cross-validation task and the mean Area Under Curve

428   (AUC) from all folds is reported to reflect the total average performance of each learning task.

429   The implementations of Logistic Regression and Random Forest were derived from the sklearn

430   Python package (v0.22.1).

431

432 We also estimated the optimal OncMTR cut point for each classification by calculating the

433 Youden's index from each learning task. The average Youden index from all classification tasks

434 performed with Logistic Regression was Y_{LR} = -0.0409 (standard deviation: 0.00126) while for

435 Random Forest it was Y_{RF} = -0.0614 (standard deviation: 0.00057). The mean of the two

436 averages of Youden indexes is -0.05115 or -0.05, after rounding it up to one decimal point for

437 simplicity. We thus consider OncMTR values below -0.05 to have the most distinctive power.

438

439 **Identifying OncMTR regions significantly enriched for ClinVar somatic variants**

440 For this analysis, we use all ClinVar somatic variants (ORIGIN=2) from the GRCh38 release

441 (last accessed on 9 June 2019), focusing on those annotated as missense or synonymous. We

442 consider as pathogenic variants those annotated as "Pathogenic" or "Likely_pathogenic" and as

443 benign those annotated as "Benign" or "Likely_benign" (based on ClinVar). Classification

444 between pathogenic and benign (or random) variant sets was performed with a logistic

445 regression classifier with 5-fold cross validation (sklearn, Python package v0.22.1). When

446 restricting the classification to heme-implicated genes, we derived those gene sets based on the

447 Intogen annotation (**Supplementary Table 10**).

448

449 In order to identify genes/transcripts across the exome that are preferentially enriched for

450 ClinVar somatic pathogenic variants in regions with low OncMTR scores we employ Fisher's

451 exact test. Specifically, we scan across each transcript and identify what percentage of the

452 codons in each transcript achieve an OncMTR score at the bottom 20-percentile of the full

453 OncMTR distribution (across the entire transcript). Then, we check whether known pathogenic

454 or likely pathogenic ClinVar missense variants preferentially land in these codons (i.e.

455 corresponding to low OncMTR scores) compared to the rest of the transcript. We apply a

456 Fisher's exact test (FET) to evaluate the enrichment of each set of regions, i.e., those with low

457 OncMTR scores vs the rest of the transcript. Eventually, we rank each transcript based on the

458 odds ratio and significance of the FET enrichments (**Supplementary Table 11**).

459

460 **Enrichment of low OncMTR scores in protein domains**

461 To describe the functional context of OncMTR, we calculated enrichment of constrained regions

462 in protein domain families. Residues within each canonical transcript (as defined by UniProtKB)

463 were divided into two classes based on their corresponding OncMTR scores: below -0.05

464 (constrained; as defined by Youden's index) and greater or equal -0.05 (relaxed). Domain and

465 clan annotations for the human proteome were taken from the Pfam 34.0 database. DNA-

466   binding domains were pulled from a previous compendium (Bahrami et al., 2015). The final set

467   of the canonical human proteome consisted of 18,313 annotated proteins. Enrichments of the

468   constrained regions in protein domains were tested with the Fisher's exact test followed by

469   Bonferroni correction, and significance level of adjusted *p*-value of 0.05.

470

471   **UK Biobank Collapsing analysis**

472   Collapsing analyses were performed using the 394,694 exomes available in the UK Biobank

473   (UKB) cohort (Bycroft et al., 2018). The UKB is a prospective study of approximately 500,000

474   participants aged 40–69 years at time of recruitment. Participants were recruited in the UK

475   between 2006 and 2010 and are continuously followed. The average age at recruitment for

476   sequenced individuals was 56.5 years and 54% of the sequenced cohort is of female genetic

477   sex. Participant data include health records that are periodically updated by the UKB, self-

478   reported survey information, linkage to death and cancer registries, collection of urine and blood

479   biomarkers, imaging data, accelerometer data and various other phenotypic end points. All

480   study participants provided informed consent and the UK Biobank has approval from the North-

481   West Multi-centre Research Ethics Committee (MREC; 11/NW/0382).

482   We performed a gene-based collapsing analysis on 1,394 chapter IX (neoplasm)

483   phenotypes adopting our previously described approach (Wang et al., 2021). We implemented a

484   total of eight dominant collapsing models with and without OncMTR filters (**Supplementary**

485   **Table 12**). Using SnpEff (Cingolani et al., 2012), we defined PTVs as variants annotated as

486   exon_loss_variant, frameshift_variant, start_lost, stop_gained, stop_lost,

487   splice_acceptor_variant, splice_donor_variant, gene_fusion, bidirectional_gene_fusion,

488   rare_amino_acid_variant, and transcript_ablation. We defined missense as:

489   missense_variant_splice_region_variant, and missense_variant. Non-Synonymous variants

490   included: exon_loss_variant, frameshift_variant, start_lost, stop_gained, stop_lost,

491   splice_acceptor_variant, splice_donor_variant, gene_fusion, bidirectional_gene_fusion,

492   rare_amino_acid_variant, transcript_ablation, conservative_inframe_deletion,

493   conservative_inframe_insertion, disruptive_inframe_insertion, disruptive_inframe_deletion,

494   missense_variant_splice_region_variant, missense_variant, and protein_altering_variant. We

495   derived allele frequencies from gnomAD (Karczewski et al., 2020). The raredmg,

496   raredmg_OncMTR, flexdmg, and flexdmg_oncMTR models incorporated a REVEL cutoff of

497   REVEL >= 0.5 to restrict to putatively damaging missense variants (Ioannidis et al., 2016).

498   To compute p-values, the carriers of at least one qualifying variant (QV) in a gene were

499   compared to the non-carriers. The difference in the proportion of cases and controls carrying

500   QVs in a gene was tested using a Fisher's exact two-sided test. we applied the following quality

501   control filters: minimum coverage 10X; annotation in CCDS transcripts (release 22;

502   approximately 34 Mb); at most 80% alternate reads in homozygous genotypes; percent of

503   alternate reads in heterozygous variants ≤ 0.25 and ≥ 0.8; binomial test of alternate allele

504   proportion departure from 50% in heterozygous state P < 1 × 10^{-6}; GQ ≤ 20; FS ≥ 200 (indels) ≥

505   60 (SNVs); MQ ≤ 40; QUAL ≤ 30; read position rank sum score ≤ –2; MQRS ≤ –8; DRAGEN

506   variant status = PASS; the variant site achieved ten-fold coverage in ≤ 25% of gnomAD

507   exomes, and if the variant was observed in gnomAD exomes, the variant achieved exome z-

508   score ≤ –2.0 and exome MQ ≤ 30. We excluded 46 genes that we previously found associated

509   with batch effects (Wang et al., 2021).

510   For all models, we applied the following quality control filters: minimum coverage 10X;

511   annotation in CCDS transcripts (release 22; approximately 34 Mb); at most 80% alternate reads

512   in homozygous genotypes; percent of alternate reads in heterozygous variants ≤ 0.25 and ≥ 0.8;

513   binomial test of alternate allele proportion departure from 50% in heterozygous state P < 1 × 10-

514   6; GQ ≤ 20; FS ≥ 200 (indels) ≥ 60 (SNVs); MQ ≤ 40; QUAL ≤ 30; read position rank sum score

515   ≤ –2; MQRS ≤ –8; DRAGEN variant status = PASS; the variant site achieved ten-fold coverage

516   in ≤ 25% of gnomAD exomes, and if the variant was observed in gnomAD exomes, the variant

517   achieved exome z-score ≤ –2.0 and exome MQ ≤ 30. We excluded 46 genes that we previously

518   found associated with batch effects10.

519   We defined the study-wide significance threshold as p<1×10^{-8}. We have previously

520   shown using an n-of-1 permutation approach and the empirical null synonymous model that this

521   threshold corresponds to a false positive rate of 9 and 2, respectively, out of ~346.5 million tests

522   for binary traits in the setting of collapsing analysis PheWAS (Wang et al., 2021).

523

Acknowledgements

524

525 We thank Oliver Backhouse for useful discussions and his feedback on this work. We thank
526 Lawrence Middleton for his assistance in developing the OncMTR web portal.
527
528

Author Contributions

529

530 D.V., R.S.D., and S.P. designed the study. D.V., R.S.D., J.M., D.M., J.A., Q.W., B.S., A.R.H.,
531 and S.P. performed analyses and statistical interpretation. Q.W. performed bioinformatic
532 processing. D.V., R.S.D., and S.P. wrote the manuscript. D.V. and D.M. developed the web
533 portal. D.V., R.S.D., J.M., D.M., Z.Z., J.A., Q.W., B.S., A.R.H, and S.P. reviewed the manuscript.
534
535

Competing Interests

536

537 D.V., R.S.D., J.M., D.M., Z.Z., J.A., Q.W., B.S., A.R.H, and S.P. are current employees and/or
538 stockholders of AstraZeneca.
539
540

References

Bahrami, S., Ehsani, R., and Drabløs, F. (2015). A property-based analysis of human transcription factors. BMC Research Notes 8, 82.

Bhatia, K., Huppi, K., Spangler, G., Siwarski, D., Iyer, R., and Magrath, I. (1993). Point mutations in the c-Myc transactivation domain are common in Burkitt's lymphoma and mouse plasmacytomas. Nat Genet 5, 56–61.

Bick, A.G., Weinstock, J.S., Nandakumar, S.K., Fulco, C.P., Bao, E.L., Zekavat, S.M., Szeto, M.D., Liao, X., Leventhal, M.J., Nasser, J., et al. (2020). Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature 586, 763–768.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209.

Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G., and Raschellà, G. (2017). Zinc-finger proteins in health and disease. Cell Death Discov. 3, 1–12.

Chang, M.T., Bhattarai, T.S., Schram, A.M., Bielski, C.M., Donoghue, M.T.A., Jonsson, P., Chakravarty, D., Phillips, S., Kandoth, C., Penson, A., et al. (2018). Accelerating Discovery of Functional Mutant Alleles in Cancer. Cancer Discov 8, 174–183.

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

Cirstea, I.C., Kutsche, K., Dvorsky, R., Gremer, L., Carta, C., Horn, D., Roberts, A.E., Lepri, F., Merbitz-Zahradnik, T., König, R., et al. (2010). A restricted spectrum of NRAS mutations causes Noonan syndrome. Nat Genet 42, 27–29.

De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515, 209–215.

Dhindsa, R.S., Copeland, B.R., Mustoe, A.M., and Goldstein, D.B. (2020). Natural Selection Shapes Codon Usage in the Human Genome. The American Journal of Human Genetics 107, 83–95.

Dhindsa, R.S., Zoghbi, A.W., Krizay, D.K., Vasavda, C., and Goldstein, D.B. (2021). A Transcriptome-Based Drug Discovery Paradigm for Neurodevelopmental Disorders. Annals of Neurology 89, 199–211.

Feng, Y.-C.A., Howrigan, D.P., Abbott, L.E., Tashman, K., Cerrato, F., Singh, T., Heyne, H., Byrnes, A., Churchhouse, C., Watts, N., et al. (2019). Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. The American Journal of Human Genetics 105, 267–282.

Gibson, W.T., Hood, R.L., Zhan, S.H., Bulman, D.E., Fejes, A.P., Moore, R., Mungall, A.J., Eydoux, P., Babul-Hirji, R., An, J., et al. (2012). Mutations in EZH2 Cause Weaver Syndrome. Am J Hum Genet 90, 110–118.

577   Hoischen, A., van Bon, B.W.M., Rodríguez-Santiago, B., Gilissen, C., Vissers, L.E.L.M., de Vries, P.,
578   Janssen, I., van Lier, B., Hastings, R., Smithson, S.F., et al. (2011). De novo nonsense mutations in ASXL1
579   cause Bohring-Opitz syndrome. Nat Genet *43*, 729–731.

580   Hoischen, A., Krumm, N., and Eichler, E.E. (2014). Prioritization of neurodevelopmental disease genes by
581   discovery of new mutations. Nat Neurosci *17*, 764–772.

582   Hyman, D.M., Taylor, B.S., and Baselga, J. (2017). Implementing Genome-Driven Oncology. Cell *168*,
583   584–599.

584   Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q.,
585   Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of
586   Rare Missense Variants. The American Journal of Human Genetics *99*, 877–885.

587   Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel,
588   C.H., Burtt, N., Chavez, A., et al. (2014). Age-Related Clonal Hematopoiesis Associated with Adverse
589   Outcomes. New England Journal of Medicine *371*, 2488–2498.

590   Jaiswal, S., Natarajan, P., Silver, A.J., Gibson, C.J., Bick, A.G., Shvartz, E., McConkey, M., Gupta, N.,
591   Gabriel, S., Ardissino, D., et al. (2017). Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular
592   Disease (Massachusetts Medical Society).

593   Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R.,
594   Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold.
595   Nature *596*, 583–589.

596   Kaplanis, J., Samocha, K.E., Wiel, L., Zhang, Z., Arvai, K.J., Eberhardt, R.Y., Gallone, G., Lelieveld, S.H.,
597   Martin, H.C., McRae, J.F., et al. (2020). Evidence for 28 genetic disorders discovered by combining
598   healthcare and research data. Nature *586*, 757–762.

599   Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia,
600   K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from
601   variation in 141,456 humans. Nature *581*, 434–443.

602   King, B., Trimarchi, T., Reavie, L., Xu, L., Mullenders, J., Ntziachristos, P., Aranda-Orgilles, B., Perez-
603   Garcia, A., Shi, J., Vakoc, C., et al. (2013). The ubiquitin ligase FBXW7 modulates leukemia-initiating cell
604   activity by regulating MYC stability. Cell *153*, 1552–1566.

605   Kosaki, R., Terashima, H., Kubota, M., and Kosaki, K. (2017). Acute myeloid leukemia-associated
606   DNMT3A p.Arg882His mutation in a patient with Tatton-Brown–Rahman overgrowth syndrome as a
607   constitutional mutation. American Journal of Medical Genetics Part A *173*, 250–253.

608   Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L.,
609   Pich, O., Bonet, J., Kranas, H., et al. (2020). A compendium of mutational cancer driver genes. Nat Rev
610   Cancer *20*, 555–572.

611   Matsuda, K., Shimada, A., Yoshida, N., Ogawa, A., Watanabe, A., Yajima, S., Iizuka, S., Koike, K., Yanai, F.,
612   Kawasaki, K., et al. (2007). Spontaneous improvement of hematologic abnormalities in patients having
613   juvenile myelomonocytic leukemia with specific RAS mutations. Blood *109*, 5477–5480.

614   Muiños, F., Martínez-Jiménez, F., Pich, O., Gonzalez-Perez, A., and Lopez-Bigas, N. (2021). In silico
615   saturation mutagenesis of cancer genes. Nature *596*, 428–432.

616   Oliveira, J.B., Bidère, N., Niemela, J.E., Zheng, L., Sakai, K., Nix, C.P., Danner, R.L., Barb, J., Munson, P.J.,
617   Puck, J.M., et al. (2007). NRAS mutation causes a human autoimmune lymphoproliferative syndrome.
618   PNAS *104*, 8953–8958.

619   Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic Intolerance to
620   Functional Variation and the Interpretation of Personal Genomes. PLOS Genetics *9*, e1003709.

621   Petrovski, S., Küry, S., Myers, C.T., Anyane-Yeboa, K., Cogné, B., Bialer, M., Xia, F., Hemati, P., Riviello, J.,
622   Mehaffey, M., et al. (2016). Germline De Novo Mutations in GNB1 Cause Severe Neurodevelopmental
623   Disability, Hypotonia, and Seizures. The American Journal of Human Genetics *98*, 1001–1010.

624   Porta-Pardo, E., Kamburov, A., Tamborero, D., Pons, T., Grases, D., Valencia, A., Lopez-Bigas, N., Getz, G.,
625   and Godzik, A. (2017). Comparison of algorithms for the detection of cancer drivers at subgene
626   resolution. Nat Methods *14*, 782–788.

627   Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A.,
628   Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo
629   mutation in human disease. Nat Genet *46*, 944–950.

630   Tamborero, D., Rubio-Perez, C., Deu-Pons, J., Schroeder, M.P., Vivancos, A., Rovira, A., Tusquets, I.,
631   Albanell, J., Rodon, J., Tabernero, J., et al. (2018). Cancer Genome Interpreter annotates the biological
632   and clinical relevance of tumor alterations. Genome Medicine *10*, 25.

633   Tatton-Brown, K., Seal, S., Ruark, E., Harmer, J., Ramsay, E., del Vecchio Duarte, S., Zachariou, A., Hanks,
634   S., O'Brien, E., Aksglaede, L., et al. (2014). Mutations in the DNA methyltransferase gene DNMT3A cause
635   an overgrowth syndrome with intellectual disability. Nat Genet *46*, 385–388.

636   Traynelis, J., Silk, M., Wang, Q., Berkovic, S.F., Liu, L., Ascher, D.B., Balding, D.J., and Petrovski, S. (2017).
637   Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant
638   interpretation. Genome Res. *27*, 1715–1729.

639   Wang, Q., Dhindsa, R.S., Carss, K., Harper, A.R., Nag, A., Tachmazidou, I., Vitsios, D., Deevi, S.V.V.,
640   Mackay, A., Muthas, D., et al. (2021). Rare variant contribution to human disease in 281,104 UK Biobank
641   exomes. Nature 1–9.

642   Whiteside, D., McLeod, R., Graham, G., Steckley, J.L., Booth, K., Somerville, M.J., and Andrew, S.E.
643   (2002). A homozygous germ-line mutation in the human MSH2 gene predisposes to hematological
644   malignancy and multiple café-au-lait spots. Cancer Res *62*, 359–362.

645   Yoda, A., Adelmant, G., Tamburini, J., Chapuy, B., Shindoh, N., Yoda, Y., Weigert, O., Kopp, N., Wu, S.-C.,
646   Kim, S.S., et al. (2015). Mutations in G protein β subunits promote transformation and kinase inhibitor
647   resistance. Nat Med *21*, 71–75.

648   Zhang, Z.-M., Lu, R., Wang, P., Yu, Y., Chen, D., Gao, L., Liu, S., Ji, D., Rothbart, S.B., Wang, Y., et al.
649   (2018). Structural basis for DNMT3A-mediated de novo DNA methylation. Nature *554*, 387–391.

650