Integrative screening identifies functional variants and VNTRs underlying GWAS signals at the 5p15.33 multi-cancer susceptibility locus

Aidan O'Brien[1,2], Hyunkyung Kong[1], Harsh Patel[1], Michelle Ho[1], Minal B. Patel[1], Jun Zhong[1], Mai Xu[1], Brenen W. Papenberg[1], Katelyn E. Connelly[1], Irene Collins[1], Rebecca Hennessey[1], Rohit Thakur[1], Hayley Sowards[1], Karen Funderburk[1], Thong Luong[1], Oscar Florez-Vargas[1], Timothy Myers[3], Ashley Jermusyk[1], Bryan Gorman[4], Wen Luo[5], Kristine Jones[5], Sudipto Das[6], The Pancreatic Cancer Cohort and Case-Control Consortia (PanScan/PanC4), Melanoma Meta-Analysis Consortium, International Lung Cancer Consortium (ILCCO), Qing Lan[7], Nathaniel Rothman[7], James D. McKay[8], Rayjean J. Hung[9], Christopher I. Amos[10], Mark M. Iles[11], Stella Koutros[12], Maria Teresa Landi[4,13], Matthew H. Law[14,15,16], Rachael Z. Stolzenberg-Solomon[17], Brian Wolpin[18], Manal Hassan[19], Alison P. Klein[20], Samuel O Antwi[21,22], Nick Orr[2], Stephen J. Chanock[23], Sara Lindström[24,25], Jason W. Hoskins[1], Marc-Henri Stern[26], Thorkell Andresson[6], Jianxin Shi[27], Ludmila Prokunina-Olsson[1], Jiyeon Choi[1], Kevin M. Brown[1,28], Laufey T. Amundadottir[1,28]

[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, 20850, USA, [2]The Johnston Cancer Research Centre, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, UK, [3]Laboratory of Genetic Susceptibility, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, 20850, USA, [4]Trans-Divisional Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, 20850, USA, [5]Cancer Genomics Research Laboratory, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, 20850, USA, [6]CCR Protein and Metabolite Characterization Core (PMCC), Frederick National Laboratory for Cancer Research, National Cancer Institute, Frederick, MD, USA, [7]Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, 20850, USA, [8]International Agency for Research on Cancer, World Health Organization, Lyon, France, [9]Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Canada, [10]Department of Internal Medicine, University of New Mexico, Albuquerque, NM, [11]Leeds Institute for Data Analytics, University of Leeds, Leeds, UK, [12]Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA, [13]Integrative Tumor Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, 20850, USA, [14]Statistical Genetics, QIMR Berghofer Medical Research Institute, 300 Herston Road, Brisbane, QLD, 4006, Australia, [15]School of Biomedical Sciences, Faculty of Health, Queensland University of Technology, Brisbane, QLD, 4001, Australia, [16]School of Biomedical Sciences, Faculty of Medicine, University of Queensland, Brisbane, QLD, 4072, Australia, [17]Metabolic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, 20850, USA, [18]Department of Medical Oncology, Dana-Farber Harvard Cancer Center, USA, [19]Department of Gastrointestinal Medical Oncology, Division of Cancer Prevention and Population Sciences, MD Anderson Cancer Center, Houston, TX, USA, [20]Departments of Oncology, Pathology and Medicine, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD, USA, [21]Department of Quantitative Health Sciences, Division of Epidemiology,, Mayo Clinic, Jacksonville, FL, USA, [22]Department of Internal Medicine, Division of Gastroenterology and Hepatology, Mayo Clinic, Jacksonville, FL, USA, [23]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, 20850, USA, [24]Department of Epidemiology, University of Washington, Seattle, WA, USA, [25]Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA, [26]Inserm U1339 UMR3666, DNA Repair and Uveal Melanoma (D.R.U.M.), Institut Curie, PSL Research University, Paris, France, [27]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, 20850, USA, [28]These authors jointly supervised this work

Abstract

Chromosome 5p15.33 harbors several independent association signals which demonstrate antagonistic pleiotropy across cancer types, with causal mechanisms largely unresolved. To identify functional variants and enhancer elements at this locus, we performed statistical fine-mapping followed by massively parallel reporter assays (MPRA) and proliferation based CRISPRi screens. This approach identified eight multi-cancer functional variants (MCFVs) across three GWAS signals. Targeting rs421629 (part of the *CLPTM1L* signal marked by rs465498) with CRISPRi revealed opposing effects on *TERT* expression in pancreatic versus lung cancer cells, consistent with the antagonistic pleiotropy observed for this signal. Furthermore, CRISPRi nominated an intronic *CLPTM1L* variable number tandem repeat (VNTR) as a potent enhancer. Long-read sequencing established VNTR polymorphisms as potential causal variants for the rs465498 signal. We showed that Hippo-pathway transcription factors mediate VNTR enhancer activity in lung and pancreatic cancer cells. Together, these findings indicate that cancer susceptibility at 5p15.33 may be mediated by both SNPs and VNTRs and provide an integrated framework for resolving complex pleiotropic loci.

Introduction

While genome-wide association studies (GWAS) have identified hundreds of common risk loci for human cancers, functional characterization of the association signals has lagged behind their identification[1]. Many such loci are thought to be underpinned by causal variants which influence allele-specific *cis*-regulation of nearby target genes[2,3]. Chromosome 5p15.33 is one of the most replicated cancer susceptibility loci, where at least 10 distinct signals have been identified across as many cancer types[4,5]. Successive multi-cancer fine-mapping analyses by Wang and colleagues in 2014[4] and Chen and colleagues in 2021[5] have applied the Association Analysis for SubSETs (ASSET)[6] method to GWAS datasets at the 5p15.33 locus, noting considerable genetic correlation between cancer types. Intriguingly, signals at 5p15.33 demonstrate antagonistic pleiotropy, with risk alleles for one cancer being protective for other cancer types[4,5,7].

At 5p15.33, both *TERT* and *CLPTM1L* are plausible target genes with established roles in tumor development and progression. *TERT* encodes the catalytic subunit of telomerase reverse transcriptase, which maintains telomere length and protects chromosomes from replication-associated erosion. Telomerase activation via upregulation of *TERT* is observed in over 90% of cancers[8–10]. Furthermore, *TERT* represents an established cancer susceptibility gene via both rare and common germline promoter polymorphisms[11–16]. Somatic *TERT* promoter mutations are prognostic indicators for multiple cancers[17–21], and act synergistically with proximal germline polymorphisms (rs2853669) in melanoma[22], bladder cancer[18,23], and glioma[21,24]. Notably, cancers such as pancreatic ductal adenocarcinoma (PDAC) exhibit activated telomerase activity[25] despite lacking canonical *TERT* promoter mutations[26], suggesting alternate *cis*-regulatory activating mechanisms. *CLPTM1L* encodes a transmembrane lipid scramblase which is overexpressed in pancreatic, ovarian, and lung cancers and contributes to RAS-dependent transformation[27], acting as an anti-apoptotic factor upon endoplasmic reticulum stress[28–31]. Recent functional analyses[32,33] and transcriptome-wide association studies[34] further implicate *CLPTM1L* expression as

46 a potential mediator of risk for uveal melanoma and pancreatic cancer, respectively.
47 Together, these observations raise the possibility that 5p15.33 risk alleles converge
48 through expression of *TERT*, *CLPTM1L*, or both, motivating a high-throughput, cross-
49 cancer functional screen of cancers associated with the 5p15.33 locus using
50 methodologies sensitive to downstream effects on each gene.

51 Post-GWAS studies focusing on specific association signals and cancer types have
52 characterized *cis*-regulatory effects of functional sequence variants on both *TERT* and
53 *CLPTM1L* as effector genes[7,12,32,35–40]. More recently, data presented by Florez-Vargas
54 and colleagues[41] suggests that variable number tandem repeats (VNTRs) may
55 represent an important class of credible causal variants (CCVs) at 5p15.33[42–44],
56 indicating that SNP-centric models of GWAS causality may overlook more cryptic
57 classes of functional variants. With no extensive cross-cancer screen of regulatory
58 elements and CCVs performed for the larger region to date, we set out to
59 comprehensively characterize *cis*-regulatory elements and CCVs to better understand
60 the mechanisms underpinning cancer risk and pleiotropy at this important multi-cancer
61 susceptibility locus.

62 To systematically characterize functional variants and *cis*-regulatory elements important
63 for putative 5p15.33 effector genes, we performed an integrative functional screen at
64 this locus. First, we fine-mapped GWAS datasets from melanoma, lung, pancreatic, and
65 bladder cancers and compiled correlated variants from previously published signals to
66 derive a panel of cross-cancer CCVs. We then implemented a massively parallel
67 reporter assay (MPRA) to assess allele-specific regulatory activity for CCVs in cell lines
68 representative of the four fine-mapped cancers. Next, we agnostically identified *cis*-
69 regulatory elements using a tiled CRISPR-inhibition (CRISPRi) proliferation screen in
70 cell lines from the same four cancer types. By integrating MPRA and CRISPRi results,
71 we derived multi-cancer functional variants (MCFV) which mediate their effect through
72 *cis*-regulation of *TERT* and/or *CLPTM1L* across cancer types. In addition, CRISPRi
73 screening nominated a variable number tandem repeat (VNTR) within intron 9 of
74 *CLPTM1L* as a potent enhancer element. Long-read sequencing, followed by imputation
75 of VNTR alleles into a pancreatic cancer GWAS dataset, revealed that polymorphisms

of this element contribute to the credible set of variants underlying the rs465498 signal at *CLPTM1L*. Subsequent functional analyses revealed VNTR-mediated regulation of both *TERT* and *CLPTM1L* through Hippo signaling pathway transcription factors.

Collectively, our work applies and extends a generalizable framework for integrative high-throughput functional screens to assess causal regulatory mechanisms at complex GWAS loci. By expanding our scope to repeat polymorphisms, we demonstrate that VNTRs and SNPs may jointly underlie cancer susceptibility signals through *cis*-regulation of both *TERT* and *CLPTM1L*. In doing so, our work bridges a gap between statistical genetics and functional biology, providing new insights into the molecular basis of cancer heritability at one of the most recurrent and complex cancer susceptibility loci identified through GWAS.

Results

Fine-mapping reveals multiple credible sets at 5p15.33

To comprehensively compile credible causal variants (CCVs) for massively parallel reporter assay (MPRA) analysis, we performed ancestry-specific fine-mapping of GWAS summary statistics from four 5p15.33-implicated cancer types in populations of European ancestry using SuSiE[45] to derive 95% credible causal sets (PDAC: 9,040 cases and 12,496 control subjects[46], melanoma: 36,760 cases and 375,188 controls[47], bladder cancer: 13,790 cases and 343,502 controls[48] and lung adenocarcinoma: 29,266 cases and 56,450 controls[49]). We also assessed a lung adenocarcinoma GWAS of 21,658 cases and 150,676 controls of East Asian[50] (EAS) ancestry (**Table 1, Fig. 1, Supplementary Figs. 1&2)**. These credible sets were denoted by their lead variant (highest posterior inclusion probability, PIP) and assigned to individual pleiotropic ASSET signals reported by Chen and colleagues[5] based on pairwise LD between credible set lead variants and ASSET tag variants ($r^2 \geq 0.6$, 1000 Genomes Project Phase 3, 1kGP p3[51], **Fig. 2, Methods**).

Across studies, SuSiE resolved 11 cancer-specific credible sets (ranging from 1–19 variants; **Table 1, Supplementary Fig. 2**), with 10/11 credible sets recapitulating previously reported ASSET signals (**Table 1**). Lung EUR yielded four credible sets, consistent with previously reported signals[49]. Lung EAS, by contrast, yielded a single credible set marked by rs2736100. Fine-mapping for bladder cancer resulted in a single credible set mapping to the signal marked by rs10069690. Additionally, SuSiE fine-mapping nominated a melanoma signal marked by rs2853667 that was not captured by ASSET (**Table 1**). The rs465498-tagged signal mapping across *CLPTM1L,* notably yielding a larger number of CCVs relative to other signals, was supported in PDAC, melanoma, and lung EUR GWAS, with antagonistic pleiotropy consistent with previous analyses (**Supplementary Fig. 2f**). SuSiE did not return a credible set for several previously reported association signals (PDAC rs2736098[52], lung EAS rs1316780 and rs62332591[50]; **Supplementary Fig. 2**.). In two instances the best-matching ASSET

signal was supported only by modest LD with these SNPs (PDAC rs2735940 with rs2853677, r^2 = 0.62; melanoma rs2736100 with rs7705526, r^2 = 0.51; **Table 1**).

To ensure comprehensive coverage for our MPRA design, we supplemented ASSET signals and cancer-specific CCVs with ancestry-matched proxy variants from 1kGP p3 ($r^2 \geq 0.6$) to derive a final MPRA panel of 116 5p15.33 CCVs (**Fig. 1a; Supplementary Fig. 1, Supplementary Table 1**). Given considerable cross-correlation of SuSiE credible sets (**Supplementary Fig.3**), subsequent functional annotation of CCVs was stratified by ASSET signal, with SuSiE results used supplementarily to ensure comprehensive functional annotation across pleiotropic signals.

MPRA identifies functional *cis*-regulatory CCVs

To assess allele-specific *cis*-regulatory activity for each of the 116 fine-mapped CCVs, we utilized an MPRA strategy based on previously established workflows[53–56] across eight cancer cell lines representative of the four fine-mapped cancers (pancreas: PANC-1 and MIA PaCa-2; melanoma: UACC903 and UACC1113; bladder: 5637 and UM-UC-3; lung: A549 and NCI-H520). After quality control (see **Methods, Supplementary Fig.4**), 61.21% of tested CCVs demonstrated a significant allelic effect (FDR ≤ 0.01) in at least one cell line (**Supplementary Table 2**), although effect sizes were generally modest (median allelic $\log_2$ fold change for significant CCVs was 0.067, **Supplementary Fig.5**). We therefore prioritized CCVs using additional stringency metrics in a step-wise manner: (i) active element status for the 145 bp sequence harboring each variant determined by divergence of either allele from a distribution of scrambled controls (n = 79 of 109 CCVs passing MPRA design and QC; see **Methods; Supplementary Fig. 6**) (ii) FDR ≤ 0.01 for allelic effects (n = 71), (iii) absolute allelic Δ RNA/DNA TPM ≥ 15% (n = 13). Variants meeting all three criteria in at least one cell line were designated MPRA-hits (n = 13; **Table 2**). Effect sizes were concordant with independent MPRA experiments of overlapping CCVs in UACC903 cells by Choi and colleagues[55] (Pearson's r = 0.88, n = 5), and lung cancer cell lines by Long and colleagues[54] (A549 r = 0.84, NCI-H520 r = 0.82, n = 17). MPRA-functional CCVs mapped to five ASSET signals (marked by rs465498: n = 7 CCVs; rs11414507: 2; rs7705526: 1; rs7726159: 2;

144 rs451360: 1; **Supplementary Table 3).** We annotated CCVs with functional scores from
145 RegulomeDB[57], FORGEdb[58], and assessed disruption/creation of transcription factor
146 binding sites with motifbreakR[59] (**Supplementary Table 4**), but did not observe robust
147 enrichment of MPRA hits among higher-scoring or motif-disrupting candidates (see
148 **Supplementary Results**), underscoring the value of direct functional screening at this
149 locus.

150 A tiled CRISPRi screen at 5p15.33 supported MPRA-hits and 151 revealed additional regulatory elements

152 Because episomal MPRA constructs are not integrated into the endogenous chromatin
153 context, we complemented this approach with an agnostic, tiled CRISPRi proliferation
154 screen of 25,778 sgRNAs across this locus (chr5:1,233,287-1,380,000, GRCh37)
155 ensuring robust coverage of CCVs across the sgRNA pool (**Supplementary Table 5,**
156 **Supplementary Results**). We used representative cell lines from the same four cancer
157 types engineered to stably express dCas9-KRAB-ZIM3, retaining seven of the eight
158 lines described above and substituting TCCSUP for 5637 for bladder cancer (see
159 **Methods**)**.** After confirming both *TERT* and *CLPTM1L* as proliferation-essential genes
160 in these cell lines (consistent with data from DepMap[60], knockdown data not presented),
161 CRISPRi screens were carried out over 25 cumulative cell doublings with sampling at
162 T_0, T_1 (~12.5 cell doublings) and T_2 (~25 doublings). Resultant screen data was
163 analyzed with MAGeCK[61], demonstrating robust depletion of sgRNAs targeting the
164 promoters of genes selected for proliferation-essentiality and cancer-type susceptibility
165 status, confirming screen efficacy (**Methods, Supplementary Tables 6-7**). We next
166 used CRISPR-SURF[62] to deconvolve signals from 5p15.33-tiled sgRNAs across
167 datapoints into 2,718 proliferation-essential elements at FDR ≤ 0.05 (median length 470
168 bp; **Supplementary Table 8**), including strong signals intersecting the *TERT* and
169 *CLPTM1L* promoters. This nominated a large fraction of the locus as proliferation-
170 essential (median ~40–50% per cell line per timepoint; **Supplementary Table 9**),
171 whereas both melanoma lines showed substantially reduced sensitivity (6.1–17.6% of
172 the locus across timepoints).

173 To nominate CCVs with the strongest functional support across the four cancers, we
174 integrated MPRA hits with CRISPRi screening data stratified by pleiotropic ASSET
175 signals[5] **(Fig. 2)**. We defined CCVs as *functional variants* for a given cancer type upon
176 concurrent observation of (i) MPRA-hit status and (ii) colocalization with a CRISPR-
177 SURF-identified regulatory element (FDR ≤ 0.05) in at least one representative cell line
178 for a given cancer type (**Fig. 2**). We then defined variants as *multi-cancer functional*
179 *variants (MCFVs)* upon annotation as functional variants in more than one cancer type,
180 nominating eight MCFVs. Stratifying MCFVs by corresponding ASSET signal (pairwise
181 $r^2 \geq 0.6$; 1000 Genomes Phase 3 EUR; Fig. 2), the rs465498 signal in *CLPTM1L* again
182 accounted for the largest set, yielding four MCFVs: rs421629 ($r^2 = 0.98$ with rs465498),
183 rs11133729 ($r^2 = 0.99$), rs31487 ($r^2 = 0.97$), and rs37005 ($r^2 = 0.75$). Notably, rs421629,
184 rs11133729, and rs31487 were included in the PDAC SuSiE credible set, and rs31487
185 additionally mapped to the melanoma credible set (**Fig. 2**). Three MCFVs were detected
186 within intron 2 of *TERT* (rs7705526 signal: rs4449583 ($r^2 = 0.78$); rs11414507 signal:
187 rs74682426 ($r^2 = 0.98$) and rs56023411 ($r^2 = 0.62$); **Fig. 2**). Neither of these MCFVs
188 were captured in SuSiE credible sets but were associated with individual cancer types
189 (rs4449583: $P_{melanoma} = 2.04\times10^{-15}$, $P_{lung_EUR} = 1.49\times10^{-20}$, $P_{lung_EAS} = 9.82\times10^{-43}$;
190 rs56023411: $P_{melanoma} = 1.27\times10^{-17}$). Another *TERT* intron 2 MCFV, rs62332591 was
191 not correlated to any 2021 ASSET signal[5]) at $r^2 \geq 0.6$, but was most highly correlated to
192 signal rs2853677 from the 2014 ASSET analysis[4]($r^2 = 0.49$).

193 Having nominated a panel of MCFVs, we next used SNP-centric individual CRISPRi
194 guides to link functional variants to their effector gene(s) (**Fig. 3a-b**). Guided by
195 consistent antagonistic pleiotropy between PDAC and lung cancer in published ASSET
196 analyses[4,5], we performed CRISPRi to target the respective regions surrounding
197 MCFVs in representative cell lines (PANC-1 for PDAC; A549 for lung cancer) and
198 assessed effects on *TERT* and *CLPTM1L* expression by TaqMan RT-qPCR analysis
199 (**Fig. 3c**; **Methods**). Targeting MCFVs assigned to both the rs7705526 (blue, **Fig. 3**)
200 and rs11414507 (red, **Fig. 3**) ASSET signals in *TERT*'s intron 2 robustly decreased
201 *TERT* expression in tested cells, with weak or no effects on *CLPTM1L*. By contrast,
202 three MCFVs assigned to the rs465498 signal (purple, **Fig. 3**), comprising CCVs intronic
203 and upstream of *CLPTM1L*, produced more modest and heterogeneous transcriptional

204 effects. Perturbation of the respective regions surrounding MCFVs rs31487 and
205 rs37005 reduced *TERT* expression (with rs37005 also decreasing *CLPTM1L* in A549
206 cells), whereas perturbation of the region containing rs421629 (within *CLPTM1L* intron
207 16) yielded opposing *TERT* effects between cell types (increased in PANC-1 but
208 decreased in A549 for a subset of sgRNAs) (**Fig. 3c**).

209 A variable number tandem repeat (VNTR) polymorphism in 210 *CLPTM1L* is part of the signal in *CLPTM1L* tagged by rs465498

211 SNP rs11133729 was identified as a MCFV mapping to the ASSET signal tagged by
212 rs465498 (r^2 = 0.99, **Fig. 4a**). Located in intron 9 of *CLPTM1L,* this region was
213 annotated by prominent enhancer/promoter marks across ENCODE cell lines (**Fig. 4a)**
214 and was nominated as proliferation-essential in 7/8 cell lines (**Fig. 4b**). This SNP maps
215 within the bounds of a variable number tandem repeat (VNTR) (2,728 bp:
216 chr5:1,326,295-1,329,022, GRCh38, denoted VNTR1), itself located 371 bp upstream of
217 a second VNTR (669 bp, denoted VNTR2, chr5:1,329,508-1,330,176, GRCh38), with
218 both VNTRs comprising distinct 29 bp consensus repeat units (RU). Given emergent
219 appreciation of repeat polymorphisms as underlying drivers of GWAS association
220 signals[41,63–65], we next aimed to determine whether *CLPMT1L* intron 9 VNTR
221 polymorphisms might represent CCVs at 5p15.33. To achieve this, we utilized a joint
222 genotyping strategy to infer read-phased VNTR and SNP haplotypes from PacBio
223 amplicon reads within *CLPTM1L* intron 9 across 484 1kGP EUR samples as well as
224 DNA samples from 940 patients diagnosed with PDAC and 870 controls from the
225 PanScan I[66] and II[67] GWAS phases (see **Methods, Supplementary Tables 10-12**). In
226 the PDAC GWAS cohort, we resolved 459 distinct VNTR1 alleles spanning lengths of
227 2,149–4,560 bp, with a broadly trimodal frequency distribution (peaks at 2,586 bp, AF =
228 0.07; 2,789 bp, AF = 0.31; 2,934 bp, AF = 0.31; **Fig. 4c**), and VNTR2 alleles spanning
229 527–1,830 bp with a bimodal distribution (peaks at 730 bp, AF = 0.42; 933 bp, AF =
230 0.36; **Fig. 4c**).

231 Genetic association tests for VNTRs have often grouped alleles as long vs. short to
232 overcome the challenges in genotyping with traditional sequencing and PCR-based

233 methods[41,44,68]. Given that TRGT outputs full VNTR sequences as alleles, we assessed
234 VNTR associations with individual risk signals by encoding VNTR genotypes from the
235 1,810 PanScan samples in four different ways, with increasing levels of aggregation, as
236 follows (**Fig. 4c**); (i) discrete VNTR alleles (each unique haploid VNTR sequence as its
237 own allele, VNTR1_allele_n, VNTR2_allele_n, both represented by allele length
238 followed by unique index number (n)), (ii) aggregate VNTR lengths (collapsing alleles of
239 identical length regardless of sequence, VNTR1_length, VNTR2_length), (iii) data-
240 driven length classes defined by applying k-means clustering to observed allele-length
241 distributions (k selected based on visual inspection of allele-length distributions,
242 VNTR1_cluster, VNTR2_cluster, yielding k = 4 for VNTR1 (cluster 1: 2,149–2,673 bp;
243 cluster 2: 2,702–2,872 bp; cluster 3: 2,876–3,224 bp; cluster 4: 3,253–4,560 bp) and k =
244 3 for VNTR2 (cluster 1: 527–817 bp; cluster 2: 846–1,020 bp; cluster 3: 1,049–1,803 bp)
245 and (iv) a binary long/short (VNTR1_LS, VNTR2_LS) encoding obtained by collapsing
246 adjacent length classes within each VNTR (VNTR1: short = clusters 1–2, long = clusters
247 3–4; VNTR2: short = clusters 1–2, long = clusters 2-3)). We found VNTR1_LS long
248 encodings were linked to the PDAC risk haplotype marked by the rs465498 ASSET
249 signal (rs31490-A, **Fig. 4d**; pairwise LD between VNTR alleles and rs465498; VNTR1
250 long: $r^2 = 0.83$, $D' = 0.99$; VNTR2 long: $r^2 = 0.70$, $D' = 0.99$), which motivated
251 subsequent analyses to determine whether VNTR alleles comprise credible causal
252 variants for this signal.

253 VNTR alleles associate with PDAC risk

254 PacBio-derived VNTR alleles were first tested for association with PDAC risk in 1,810
255 directly sequenced PanScan I and II DNA samples (940 PDAC and 870 controls, see
256 **Methods**). Multiple VNTR1 encodings were associated with PDAC (**Supplementary**
257 **Table 13**). Among specific VNTR1 alleles, a VNTR1 2,934 bp allele (2934_30) exhibited
258 the most significant association (VNTR1_allele 2934_30; AF = 0.14, OR = 1.52, P =
259 3.25×10^{-5}) when compared to all other alleles. Aggregate VNTR1 long alleles also
260 demonstrated significant association with PDAC (VNTR1_LS long; AF = 0.52, OR =
261 1.22, $P = 2.54\times10^{-3}$), with VNTR1-cluster 3 demonstrating comparable results
262 (VNTR1_cluster 3; AF = 0.44, OR = 1.20, $P = 7.73\times10^{-3}$). By contrast, both common

263 and infrequent VNTR2 allele encodings yielded only nominally significant associations,
264 including VNTR2-long (VNTR2_LS long; AF = 0.55, OR = 1.15, P = 0.05) and a VNTR2
265 962-bp allele (VNTR2_allele 962; AF = 0.006, OR = 1.50, P = 0.01).

266 To increase statistical power, we imputed VNTR genotypes into a larger PDAC GWAS
267 dataset. Using PacBio-derived VNTR encodings and SNP calls from the 1,810 PanScan
268 I & II samples described above, we imputed each of the four VNTR genotype encodings
269 (**Fig. 4c**) as distinct multiallelic variants into the remaining samples from the
270 PanScan/PanC4 PDAC meta-GWAS[66,67] (including an additional 8,066 cases and
271 11,208 controls; **Methods**). After imputation, we filtered VNTR and SNP alleles (MAF ≥
272 0.01, imputation quality score R2 ≥ 0.3) prior to application of the same logistic model
273 applied to directly genotyped samples and subsequent meta-analysis of each cohort. In
274 agreement with the original GWAS, SNP rs31490 exhibited the most significant
275 association[46] (**Fig. 5b;** OR = 1.20, $P = 2.33\text{x}10^{-16}$). Results for imputed VNTR alleles
276 largely replicated the analysis of directly genotyped samples **(Supplementary Table**
277 **14)**. Of the four VNTR encodings, VNTR1 cluster 3 emerged as the most significant
278 VNTR association (OR = 1.20, $P = 8.86\text{x}10^{-15}$). We also replicated significant effects for
279 long/short VNTR1 alleles (OR = 1.20, $P = 4.00\text{x}10^{-15}$). Of the discrete VNTR1 alleles,
280 the 2,934 bp sequence (2934_30) which previously comprised the most significant
281 VNTR association in the pilot analysis retained genome-wide significance (OR = 1.25, P
282 $= 1.19\text{x}10^{-9}$). Collective VNTR2 lengths of 933 bp were the only significant (at $P \leq 5\text{x}10^{-}$
283 ^{8}) VNTR2 encoding (OR = 1.17, $P = 5.36\text{x}10^{-11}$).

284 Conditioning the association analysis on VNTR1 cluster 3 by including corresponding
285 VNTR encoding dosages as a fixed covariate resulted in marked attenuation of the
286 signal for lead SNP rs31490 ($P_{VNTR1_cluster3} = 9.30\text{x}10^{-3}$). Conditioning on VNTR2 lengths
287 of 933 bp resulted in more modest attenuation of the rs31490 signal ($P_{\text{VNTR2_933bp}}$ =
288 $4.13\text{x}10^{-6}$). Conditioning the analysis on rs31490 likewise attenuated the VNTR1 cluster
289 3 association ($P_{rs31490} = 0.18$). Alternatively, conditioning on SNPs known to tag PDAC
290 signals confirmed linkage with the rs31490 haplotype (P-values for VNTR1 cluster 3
291 conditioned on PDAC signal SNPs: $P_{\text{conditional_rs451360}} = 1.37\text{x}10^{-5}$, $P_{\text{conditional_rs2735940}}$ =
292 $3.62\text{x}10^{-9}$, $P_{\text{conditional_rs2853677}} = 2.79\text{x}10^{-7}$, **Supplementary Table 14**).

VNTR1 is a potent enhancer at the 5p15.33 locus

Given strong evidence that VNTR1 polymorphisms comprise plausible CCVs as part of the rs31490/rs465498-tagging haplotype (**Fig. 4d**), we next sought to characterize the biological mechanism underlying VNTR1-enhancer activity. *In-silico* analysis of VNTR1-transcription factor binding sites using TRAP[69] revealed significant enrichment for TEA Domain Family Member 1 (TEF-1) motifs ($P_{nominal} < 5.0\text{x}10^{-4}$, **Fig. 5a**, **Supplementary Table 15**). TEF-1, encoded by *TEAD1,* is a key transcription factor member of the Hippo signaling pathway involved in regulation of apoptosis and proliferation[70–73].

Similar to our approach for small sequence variants, we individually assessed VNTR1 using CRISPRi and individual luciferase reporter assays to assess functionality. CRISPRi repression of VNTR1 reduced expression of both *TERT* and *CLPTM1L* relative to negative control sgRNA in PANC-1 and A549 cells (**Fig. 5c**). VNTR1-sgRNA-A, which directly targets the region harboring rs11133729 in VNTR1 , reduced *TERT* (fold-change = 0.19, $P = 6.69\times10^{-5}$) and *CLPTM1L* (fold-change = 0.58, $P = 1.30\times10^{-4}$) expression in A549 cells, and similarly decreased *TERT* (fold-change = 0.59, $P = 2.78\times10^{-4}$) and *CLPTM1L* (fold-change = 0.72, $P = 1.43\times10^{-2}$) expression in PANC-1 cells. Luciferase reporter assays of VNTR1 alleles representative of each of the cluster-genotype encodings confirmed potent enhancer activity in both cell lines (median fold change vs. negative control: PANC-1 = 74; A549 = 609; **Fig. 5d**). Enhancer strength varied across constructs representative of VNTR1 cluster encodings: relative to the shortest tested allele (2,586 bp; assigned to cluster 1), a 2,933 bp sequence assigned to VNTR1 cluster 3 demonstrated higher activity in PANC-1 (fold-change = 1.28, $P = 2.69\times10^{-4}$) and A549 (fold-change = 1.20, $P = 2.23\times10^{-2}$) cells.

Having established VNTR1 as a potent enhancer for *TERT* and *CLPTM1L*, we proceeded to characterize the transcription factors mediating this effect. We performed DNA pulldown experiments coupled with mass spectrometry by incubating biotinylated VNTR1 repeat units with MIA PaCa-2 nuclear lysate. This analysis identified robust binding of TEF-1 peptides to VNTR1 sequences (**Fig. 5f, Supplementary Table 16**). Notably, competitor experiments led to a striking depletion of TEF-1 (**Fig. 5g**), indicating

322 highly sequence-specific protein-DNA interactions. This finding was validated by oligo
323 pulldown experiments followed by western blot showing robust binding of TEF-1
324 (*TEAD1*) as well as additional Hippo pathway family members TEF-3 (*TEAD4*) and WW
325 domain-containing transcription regulator protein 1 (*WWTR1*; also named *TAZ*) to the
326 VNTR1 repeat motif (**Fig. 5h**). Finally, we sought to determine to what degree VNTR1-
327 luciferase activity was attributable to Hippo-pathway TF signaling. Disruption of TEF-
328 YAP complexes with verteporfin, a small molecule inhibitor of TEF-YAP binding[74,75],
329 markedly reduced VNTR1-driven reporter activity across constructs in both cell lines
330 (**Fig. 5e**), indicating that VNTR1 enhancer function depends on Hippo-pathway
331 transcriptional complexes. Collectively, these data support VNTR1 polymorphisms as
332 MCFVs at 5p15.33 with regulatory implications for both *TERT* and *CLPTM1L*.

333 # Discussion

334 Here, we provide a comprehensive dissection of functional *cis*-regulatory CCVs at
335 5p15.33. By identifying functional variants which mediate causal effects on both target
336 gene expression and cell proliferation, our work furthers understanding of the
337 mechanisms underpinning susceptibility for multiple cancers. We derived an integrated
338 functional map for 116 5p15.33 CCVs and identified eight MCFVs across three distinct
339 GWAS signals, demonstrating allele preferential *cis*-regulatory effects that localized to
340 proliferation-essential regulatory elements across cancer cell lines from four cancer
341 types. Additionally, we build on findings from transcriptome-wide association studies[34]
342 and post-GWAS functional analyses which support the role of *CLPTM1L* as a possible
343 target gene for the GWAS signals at 5p15.33.

344 Antagonistic pleiotropy between cancers has long been a confounding factor at
345 5p15.33. Context-dependent telomerase biology has been hypothesized as a possible
346 mediator of such effects, although direct analysis of telomere length across cancer
347 types has yielded inconclusive results[76–84]. Our analysis marks a major step towards
348 understanding how these differences may be mediated at a *cis*-regulatory level. For all
349 eight MCFVs we identified, MPRA allelic effects demonstrated consistent directionalities
350 across cancer types, suggesting that differences in phenotypic outcome may largely be
351 driven by effects of downstream target genes and pathways, epigenetic factors, or
352 potentially by interactions with environmental factors. Our tiled CRISPRi screen
353 supported this model, with MPRA-functional CCVs intersecting elements nominated as
354 putative enhancers. Focused interrogation of these variants with CRISPRi proved
355 largely consistent with this model except for rs421629. Our findings suggest that this
356 SNP occupies a possible context-dependent regulatory element whose net effect on
357 *TERT* may switch between repressive and activating effects depending on the cell type.
358 An alternative, non-mutually exclusive possibility is that perturbation of this element
359 reconfigures the local regulatory circuit, such that inhibiting rs421629-dependent activity
360 unmasks or engages distinct enhancers or promoters in a cell type-specific manner.
361 This reversal provides a plausible *cis*-regulatory mechanism for antagonistic pleiotropy

362 at 5p15.33. Motif analysis indicated that the PDAC/melanoma risk allele (rs421629-G)
363 may introduce high affinity binding motifs for NFATC2, raising the possibility that allele-
364 specific TF recruitment, combined with lineage-specific cofactor and chromatin
365 environment, may toggle the element between enhancer- and repressor-like states.
366 While additional work is required, our data highlights how cell type-specific wiring of a
367 single *cis*-regulatory element may contribute to antagonistic pleiotropy at 5p15.33.

368 Our analysis identified a functional VNTR polymorphism as an additional CCV for the
369 signal at *CLPTM1L* marked by rs465498, where both VNTR and SNP alleles may jointly
370 underpin the observed association. We characterized VNTR1 as a potent enhancer of
371 both *TERT* and *CLPTM1L* with activity driven by Hippo pathway TFs. Notably given
372 antagonistic pleiotropy for this signal, Hippo pathway factors have been implicated in
373 both oncogenic and tumor suppressive mechanisms across cancer types[85]. We also
374 note four MCFVs in intron 2 of *TERT* intersecting repeat elements, corresponding to
375 several independent association signals. Our analysis supports recent studies which
376 posit VNTRs as a considerable source of missing heritability[86,87]. While post-GWAS
377 analyses are largely focused on the identification of causal SNPs and their relationships
378 to target genes, there are several examples of GWAS signals tagging more complex
379 variants (e.g., in *TERT*[41], *CTRB2*[88], *EGR2*[65]). Repeat elements are pervasive throughout
380 the genome, with a substantial body of evidence supporting their role as *cis*-regulatory
381 elements[89–91]. Recent studies have estimated VNTR polymorphisms to underly
382 prominent glaucoma and colorectal cancer GWAS signals[63]. Moreover, studies have
383 estimated 11.4% of structural variants and simple tandem repeats to be in LD with
384 SNPs associated with 701 GWAS traits[86]. Identification of cryptic functional alleles is
385 particularly pertinent at complex loci such as 5p15.33, where multiple independent
386 signals and context-dependent regulatory elements confound deeper understanding of
387 cancer susceptibility. To accelerate functional characterization at such loci, it is
388 imperative that repeat polymorphisms are included into GWAS imputation reference
389 panels given recent advances in long-read sequencing technologies.

390 While this work provides a comprehensive map of cross-cancer regulatory elements at
391 5p15.33, our approach is not without limitations. While this locus has been implicated in

392 numerous common cancers, we specifically fine-mapped and functionally interrogated
393 four representative cancer types. While we assessed 11 pleiotropic signals, functional
394 screening of only four cancer types may have limited our sensitivity to identify MCFVs
395 for some signals. Furthermore, all experiments were performed in cancer cell lines,
396 which do not fully recapitulate the physiological and epigenetic context in which
397 germline variants act. Future studies should aim to investigate the effects of MCFVs in
398 primary cells and 3D cell culture models to better understand the mechanisms of
399 pleiotropy at 5p15.33. In addition, lung cancer GWAS across ancestries have shown
400 that the signal tagged by rs2736100 is specific to lung adenocarcinoma (EAS), whereas
401 the signal tagged by rs6554758 is more strongly associated with squamous cell
402 carcinoma[49]. MCFVs assigned to each signal were supported by evidence from both
403 adenocarcinoma (A549) and squamous cell carcinoma (NCI-H520) models; however,
404 subsequent studies should examine how these functional alleles act across lung cancer
405 subtypes. More broadly, future work should also endeavor to test functional alleles in
406 contexts that model both genetic and environmental stressors. Because MCFV and
407 VNTR characterization assays were conducted in *TERT* promoter wild type PANC-1
408 and A549 cells[7,92], it would also be appropriate to assay the effect of these
409 polymorphisms in the context of somatically-mutated *TERT* promoters, such as in
410 UACC903 or UACC1113 melanoma cell lines. Lastly, our CRISPRi screens interrogated
411 regulatory element function without directly resolving allelic effects; lenti-MPRA or
412 CRISPR base-editing approaches could quantify allele-specific transcriptional
413 consequences within native chromatin contexts in models not amenable to transient
414 transfection.

415 In summary, our model identifies multiple potential causal variants across the region,
416 including several linked to the same signals. In addition, this work implicates rs421629
417 as part of a regulatory element with seemingly dual functions on target gene expression,
418 highlighting *cis*-regulatory mechanisms as a potential mediatory of antagonistic
419 pleiotropy at 5p15. Finally, we demonstrate that VNTR polymorphisms contribute to the
420 *CLPTM1L* signal marked by rs465498, suggesting contributions from both simple and
421 complex polymorphisms, as well as demonstrating a potentially broader role of repeat
422 polymorphism at complex susceptibility loci.

423 **Table 1: Credible sets identified via fine-mapping at 5p15.33 across four cancer types**

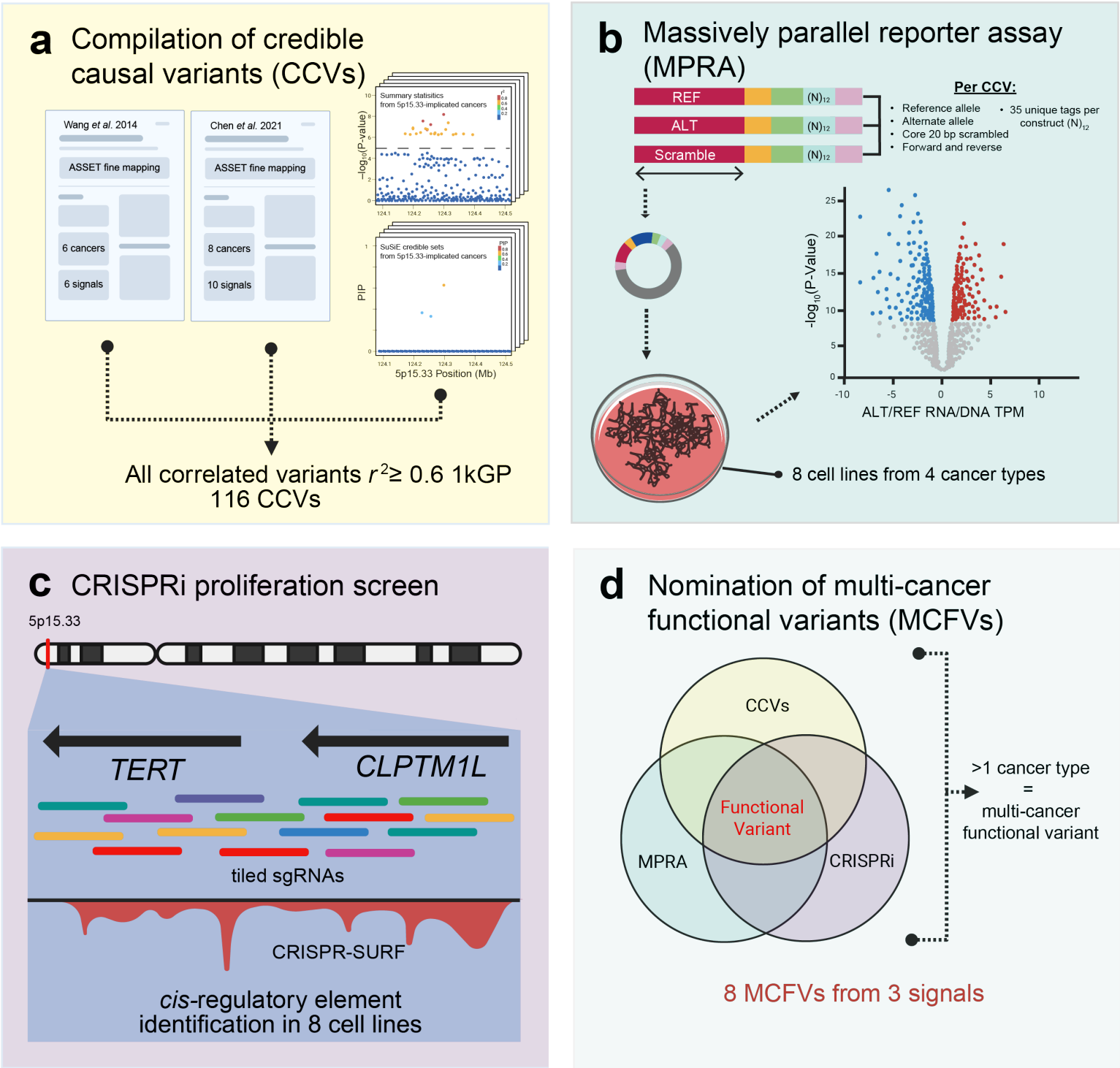
	Most significant 5p15.33 SNP							SuSiE credible sets		
GWAS Dataset	**rsID**	**chr5 Position (GRCh37)**	**chr5 Position (GRCh38)**	**Effect Allele**	**Non-Effect Allele**	**Odds Ratio**	**P-Value**	**CS-tagging CCV**	**nSNPs**	**Chen et al. 2021 SNP (r^2 1kGP EUR)**
PDAC[46]	rs31490	1344458	1344343	G	A	0.84	6.02×10^{-17}	rs31490	19	rs465498 (0.96)
								rs2735940	2	rs2853677 (0.62)
Melanoma[47]	rs13178866	1323212	1323097	C	T	0.88	1.10×10^{-43}	rs13178866	11	rs465498 (0.98)
								rs2736100	1	rs7705526 (0.51)
								rs2853667	1	-
Lung EUR[49]	rs7705526	1285974	1285859	G	A	1.25	3.80×10^{-35}	rs7705526	1	rs7705526
								rs6554758	2	rs465498 (0.77)
								rs2853677	1	rs2853677
								rs35033501	2	rs35033501
Lung EAS[50]	rs2736100	1286516	1286401	C	A	0.75	7.92×10^{-58}	rs2736100	1	rs7705526 (0.52, 0.859 EAS)
Bladder[48]	rs2242652	1280028	1279913	A	G	1.08	4.06×10^{-15}	rs10069690	1	rs10069690

424 **rsID** corresponds to the most significant 5p15.33 variant for each of the fine-mapped GWAS datasets. SuSiE fine-mapping of 5p15.33-implicated
425 cancer GWAS (PDAC (9,040 cases, 12,496 controls), melanoma (36,760 cases, 375,188 controls), lung cancer (EUR; 29,266 cases, 56,450
426 controls, EAS; 21,658 cases, 150,676 controls) and bladder cancer (13,790 cases, 343,502 controls).) resulted in 11 credible sets. All but
427 rs2853667 (melanoma) were confidently assigned to a signal-tagging SNP from Chen and colleagues' 2021 ASSET analysis[5]. Considerable
428 overlap between signals was observed (**Supplementary Fig. 3**) with GWAS of varying sample sizes and imputation strategies yielding credible
429 sets of varying sizes. Full fine-mapping results are presented in **Supplementary Table 1**.

Table 2: 5p15.33 MPRA results and functional annotations

MPRA-Hit CCVs			Functional Annotations			†MPRA Results				
rsID	chr5 Position (GRCh38)	REF>ALT	‡Regulome	‡FORGE	ƒmotifbreakR	PDAC	Melanoma	Lung	Bladder	*Number of MPRA positive cell lines
rs10054203	1279849	G>C	4	5	1	↑	↑	↑↑	-	5
rs4449583	1284020	C>T	5	5	7	↑	↑	↑↑	-	5
rs74682426	1289860	A>C	4	3	10	↑	↑	↑↑	-	5
rs62332591	1290204	T>G	1f	6	1	↓↓	↓	↓↓	↓	8
rs56023411	1291625	G>T	4	6	12	↑	↑	↑↑	-	5
rs2736102	1302029	T>C	1f	5	1	-	↓	-	↓	3
rs13170453	1317366	A>G	1b	7	0	-	↓	↓	↓	4
rs421629	1320021	G>A	1b	7	2	↓↓	↓↓	↓	↓	7
rs11133729	1326986	C>G	2b	8	5	↓	-	↓↓	↓	5
rs452384	1330725	T>C	1d	9	1	-	↑	-	↓	2
rs31487	1340986	G>C	1b	5	0	↓↓	↓↓	↓↓	↓↓	7
rs31490	1344343	G>A	1f	9	0	-	-	↓	-	1
rs37005	1356335	C>T	1b	8	10	↓↓	↓	↓↓	-	6

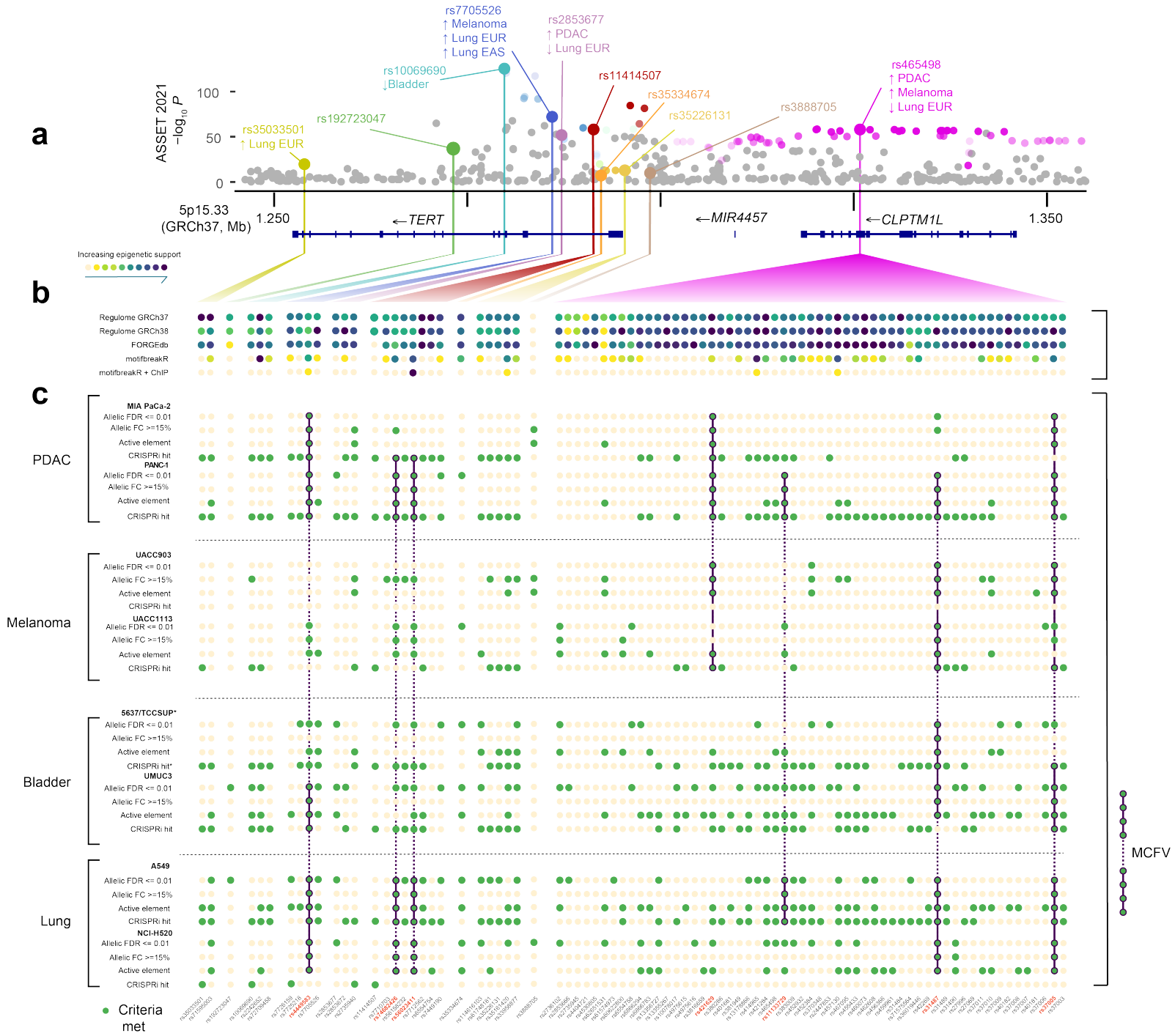
†↑↑ = ALT allele vs REF strong effect (both cell lines for each cancer type meet criteria: FDR ≤0.01, FC ≥ 15%, Active element), ↑ = Moderate ALT effect (1 cell line meets criteria/cancer), ↓↓ = REF allele strong effect (both cell lines meet criteria), ↓ = REF allele moderate effect (1 cell line meets criteria). ‡CCV functional scores from RegulomeDB (GRCh38) and FORGEdb. **RegulomeDB**: integrates eQTL/QTL, chromatin accessibility, footprinting, TF ChIP-seq and motif evidence to assign an **inverse rank (1a strongest → 7 weakest))**; **FORGEdb**: **0–10 additive evidence scores** summing support from eQTL, ABC 3D contacts, DNase hotspots, histone marks, TF motifs, and allele-specific TF occupancy (**higher = stronger regulatory support**). ƒnumber of 'strong' transcription factor binding motif events implicated with CCVs as determined by motifbreakR. *Number of cell lines across all cancer CCV where variant was an MPRA-hit.


a Compilation of credible causal variants (CCVs)
Wang et al. 2014
ASSET fine mapping
6 cancers
6 signals
Chen et al. 2021
ASSET fine mapping
8 cancers
10 signals
Summary statisitics from 5p15.33-implicated cancers
SuSiE credible sets from 5p15.33-implicated cancers
5p15.33 Position (Mb)
All correlated variants r²≥ 0.6 1kGP
116 CCVs
b Massively parallel reporter assay (MPRA)
REF
ALT
Scramble
Per CCV:
Reference allele
Alternate allele
Core 20 bp scrambled
Forward and reverse
35 unique tags per construct
-log10(P-Value)
ALT/REF RNA/DNA TPM
8 cell lines from 4 cancer types
c CRISPRi proliferation screen
5p15.33
TERT
CLPTM1L
tiled sgRNAs
CRISPR-SURF
cis-regulatory element identification in 8 cell lines
d Nomination of multi-cancer functional variants (MCFVs)
CCVs
Functional Variant
MPRA
CRISPRi
>1 cancer type = multi-cancer functional variant
8 MCFVs from 3 signals

438 ***Fig. 1***

439 ***Fig. 2: Integrative screening strategy for identification of 5p15.33 multi-cancer***
440 ***functional variants.***

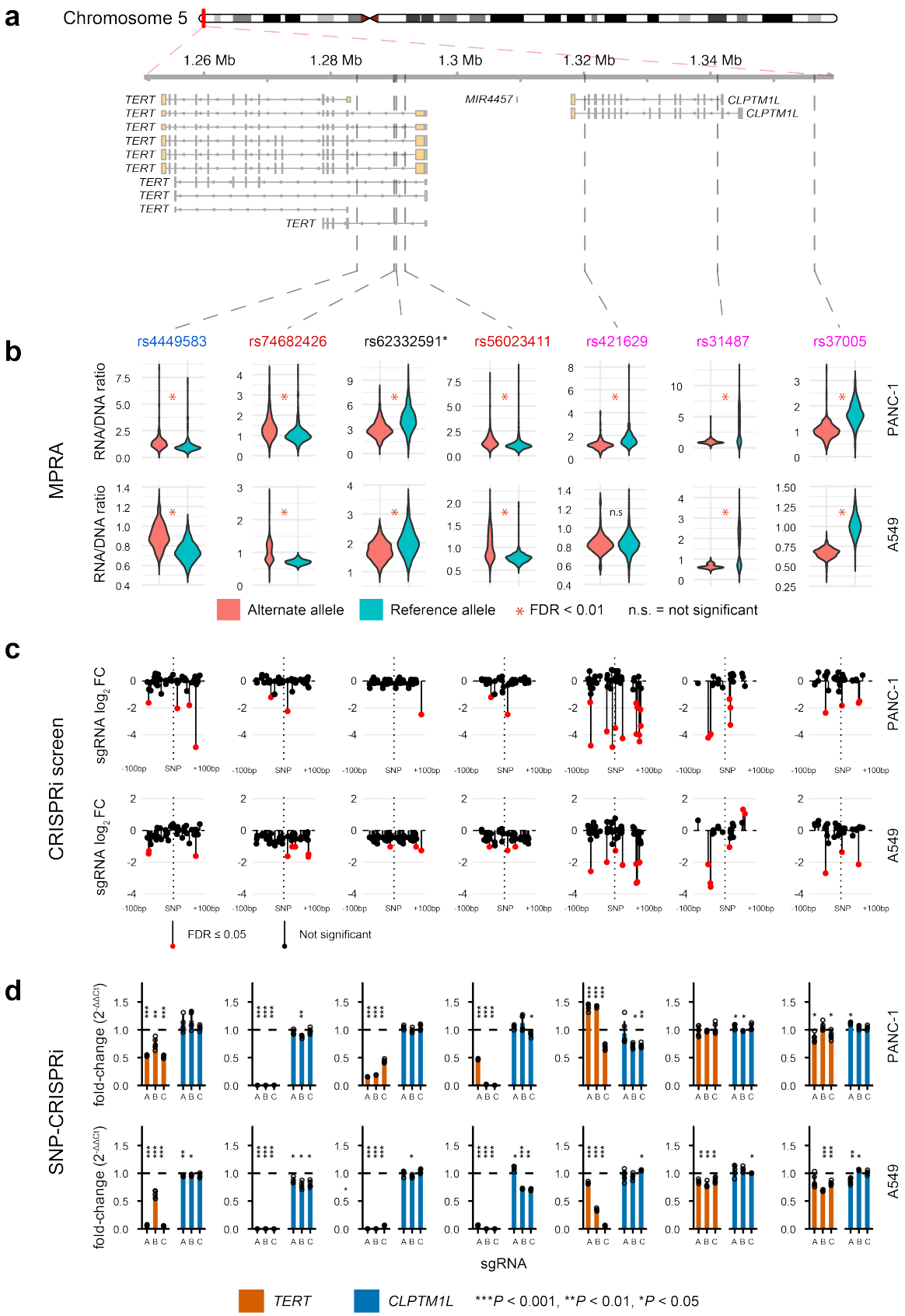
441 **a**, To collate credible causal variants (CCVs) at 5p15.33 for MPRA, we initially compiled
442 signal-tagging lead variants from two multi-cancer fine-mapping studies (Chen et al.
443 2021[5]; Wang et al. 2014[4]) along with credible set-tagging variants resulting from SuSiE
444 fine-mapping of GWAS summary statistics from four 5p15.33-implicated cancers. We
445 carried forward lead/tagging variants, additional SuSiE credible-set CCVs, and
446 ancestry-matched LD-correlated variants (1kGP Phase 3; $r^2 \geq 0.6$ for downstream
447 analyses, total 116 CCVs). **b,** Fine mapped CCVs were subject to MPRA to
448 characterize allelic effects on transcriptional output across eight cell lines representative
449 of the four fine mapped cancer types. **c,** Additionally, we assayed across the larger
450 locus to nominate *cis*-regulatory elements using a tiled CRISPRi proliferation screen
451 across eight cell lines spanning four cancer types. **d,** We then integrated results across
452 cancer types to nominate eight multi-cancer functional variants (MCFVs) across three
453 distinct signals.


a
b
c
ASSET 2021
−log10 P
100
50
0
rs35033501
↑ Lung EUR
rs192723047
rs10069690
↓Bladder
rs7705526
↑ Melanoma
↑ Lung EUR
↑ Lung EAS
rs2853677
↑ PDAC
↓ Lung EUR
rs11414507
rs35334674
rs35226131
rs3888705
rs465498
↑ PDAC
↑ Melanoma
↓ Lung EUR
5p15.33
(GRCh37, Mb)
1.250
←TERT
←MIR4457
←CLPTM1L
1.350
Increasing epigenetic support
Regulome GRCh37
Regulome GRCh38
FORGEdb
motifbreakR
motifbreakR + ChIP
PDAC
MIA PaCa-2
Allelic FDR <= 0.01
Allelic FC >=15%
Active element
CRISPRi hit
PANC-1
Melanoma
UACC903
UACC1113
Bladder
5637/TCCSUP*
CRISPRi hit*
UMUC3
Lung
A549
NCI-H520
MCFV
Criteria met

454 ***Fig. 3***

455 ***Fig. 4: Functional screening reveals multi-cancer functional variants (MCFVs)***
456 ***across 5p15.33.***

457 **a,** CCVs were stratified by pairwise LD $r^2 \geq 0.6$ relative to signal-tagging variants
458 reported by Chen *et al.* 2021[5]. ASSET signals were annotated with a cancer type if
459 SuSiE fine-mapping yielded a credible set for a given cancer corresponding to that
460 signal using the same LD threshold. Arrows assigned to each credible set indicate the
461 relative direction of effect for each fine-mapped cancer (↑ risk increasing, ↓ protective)
462 (top panel). **b,** CCVs were annotated with RegulomeDB and FORGEdb scores and
463 motifbreakR predictions of transcription factor motif disruption. MotifbreakR results were
464 further annotated for concordant TF-ChIP-seq data from the ReMap2022 database[93]
465 (motifbreakR + ChIP) (middle panel, full annotations in **Supplementary Table 3**). **c,** For
466 each CCV, binary indicators were assigned per cell line based on MPRA allelic activity
467 (FDR-significant effect, ≥ 15% absolute fold change, and active-element status) and
468 overlap with proliferation-essential elements from CRISPRi. CCVs were nominated as
469 screening hits when MPRA-functional status was supported by CRISPRi within cell lines
470 from the same cancer type (green filled circles). CCVs were further categorized as
471 multi-cancer functional variants (MCFVs) when screening hit status was observed in
472 more than one cancer type (green filled connected circles with black outline; rsIDs
473 highlighted in red, bottom panels). Functional support was strongest for CCVs mapping
474 to the rs465498 signal (four MCFVs) over *CLPTM1L* and the rs11414507 signals (two
475 MCFVs) over *TERT*. rs62332591 was also nominated as a MCFV but was not
476 correlated at $r^2 \geq 0.6$ to an ASSET signal (and not shown in this figure).


a
Chromosome 5
1.26 Mb
1.28 Mb
1.3 Mb
1.32 Mb
1.34 Mb
TERT
MIR4457
CLPTM1L
b
rs4449583
rs74682426
rs62332591*
rs56023411
rs421629
rs31487
rs37005
MPRA
RNA/DNA ratio
PANC-1
A549
n.s
Alternate allele
Reference allele
FDR < 0.01
n.s. = not significant
c
CRISPRi screen
sgRNA log2FC
-100bp
SNP
+100bp
FDR ≤ 0.05
Not significant
d
SNP-CRISPRi
fold-change (2-ΔΔCt)
A B C
sgRNA
TERT
CLPTM1L
***P < 0.001, **P < 0.01, *P < 0.05

477 ***Fig. 5***

478 ***Fig. 6: MCFV-targeted CRISPRi nominates TERT and CLPTM1L as effector genes.***

479 **a,** We identified eight MCFVs (rs11133729 assayed as part of subsequent VNTR-
480 focused analysis) across the four cancer types as MCFVs corresponding to three 2021
481 ASSET signals (LD $r^2 \geq 0.6$ EUR to ASSET signal lead SNP rs7705526 in purple,
482 rs35226131 in yellow and rs465498 in red, MCFV rs62332591 did not meet LD
483 correlation threshold to ASSET signal-tagging SNP*****). We chose to highlight screening
484 results and perform subsequent experiments for seven MCFVs in PANC-1 PDAC cells
485 and A549 lung cancer cells due to antagonistic pleiotropy across signals between these
486 two cancer types. **b,** MPRA results for MCFVs revealed significant allelic effects on
487 transcriptional output. Significance indicators reflect P-Value derived from linear model
488 comparing aggregate normalized RNA/DNA TPM ratios from 35 unique tag sequences
489 assigned to both reference (blue) and alternate alleles (red). **c,** While our CRISPRi
490 screening strategy aimed to nominate proliferation-essential elements agnostic to the
491 location of 5p15.33 CCVs, CCV-focused analysis of MCFV-harboring loci (± 100 bp
492 centered on each CCV) revealed significant depletion of sgRNAs targeting respective
493 loci in PANC-1 and A549 cells (red dots indicate FDR ≥ 0.05 depletion of sgRNA with y-
494 axis showing $\log_2$ fold change as determined by CRISPR-SURF following 25 cumulative
495 screen doublings). **d,** MCFV-targeted CRISPRi using three distinct sgRNAs (x-axis;
496 sgRNAs A, B and C) reduces *TERT* and *CLPTM1L* mRNA expression relative to a non-
497 targeting control sgRNA (y-axis; experimental sgRNA $2^{-\Delta\Delta Ct}$ normalized to $2^{-\Delta\Delta Ct}$ of non-
498 targeting sgRNA, * indicates significance under two-sided t-test). Variants attributable to
499 the rs465498 ASSET signal across *CLPTM1L* demonstrated more modest effects
500 relative to *TERT*, with rs421629 showing a striking effect on *TERT* expression across
501 cell lines, with significant upregulation observed in PANC-1 cells upon epigenetic
502 silencing contrasted to downregulation when using the same sgRNAs in A549,
503 suggesting the presence of a context-dependent regulatory element. Color of text of
504 SNPs above panel A indicates the signal each SNP is part of (in accordance with color
505 scheme in **Fig. 2**)

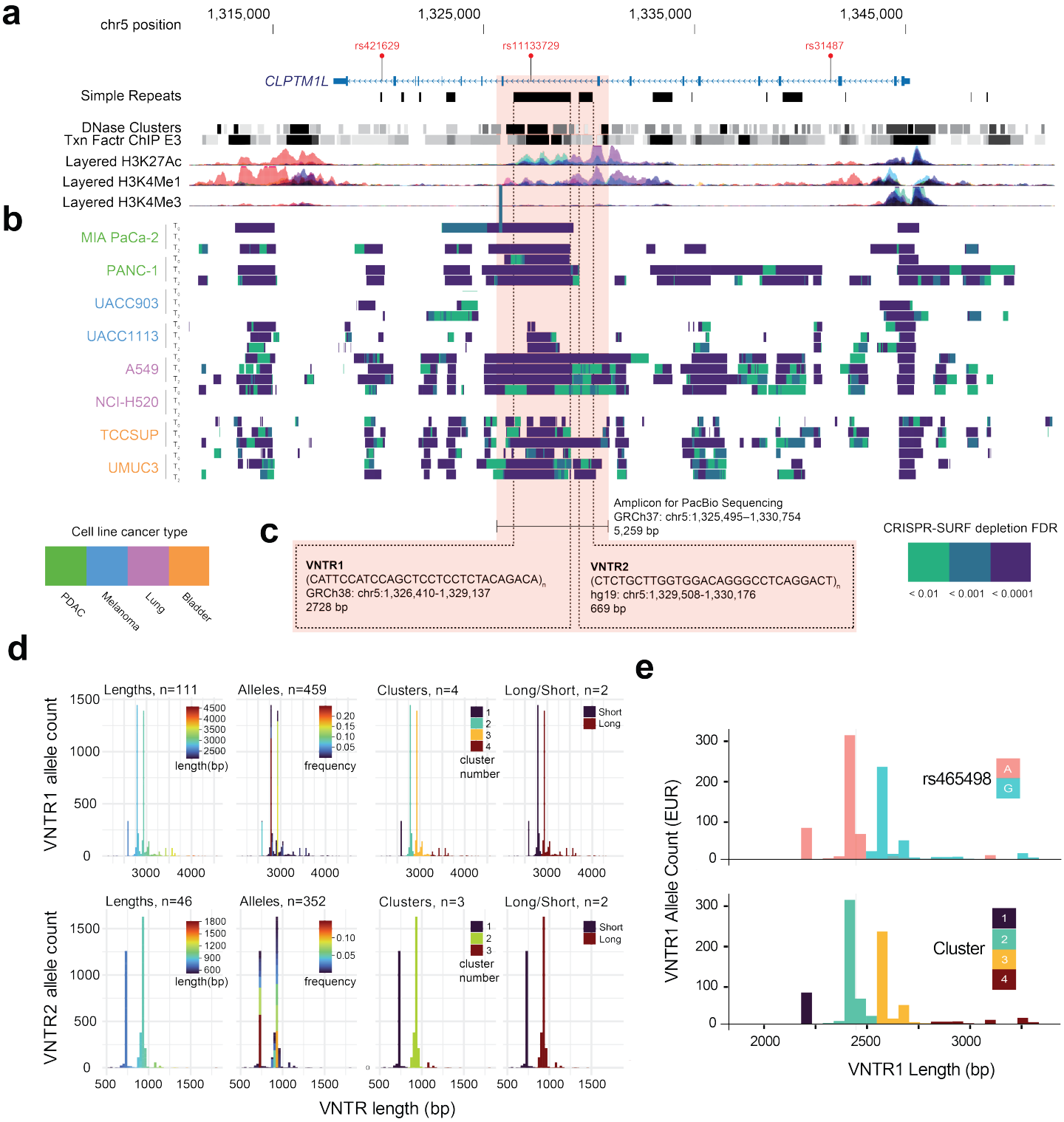
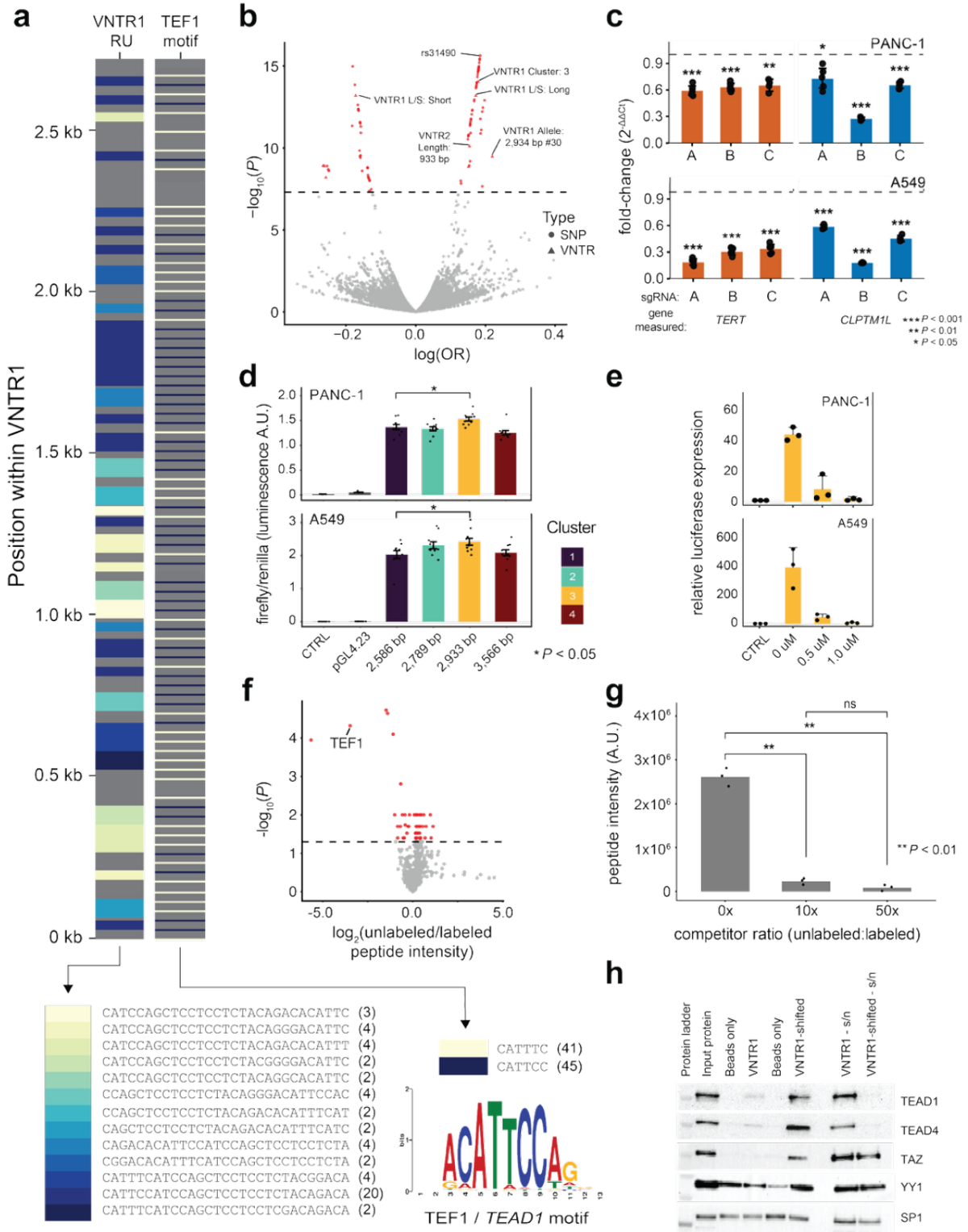

a
chr5 position
1,315,000
1,325,000
1,335,000
1,345,000
rs421629
rs11133729
rs31487
CLPTM1L
Simple Repeats
DNase Clusters
Txn Factr ChIP E3
Layered H3K27Ac
Layered H3K4Me1
Layered H3K4Me3
b
MIA PaCa-2
PANC-1
UACC903
UACC1113
A549
NCI-H520
TCCSUP
UMUC3
Cell line cancer type
PDAC
Melanoma
Lung
Bladder
c
Amplicon for PacBio Sequencing
GRCh37: chr5:1,325,495–1,330,754
5,259 bp
VNTR1
(CATTCCATCCAGCTCCTCCTCTACAGACA)n
GRCh38: chr5:1,326,410-1,329,137
2728 bp
VNTR2
(CTCTGCTTGGTGGACAGGGCCTCAGGACT)n
hg19: chr5:1,329,508-1,330,176
669 bp
CRISPR-SURF depletion FDR
< 0.01
< 0.001
< 0.0001
d
Lengths, n=111
Alleles, n=459
Clusters, n=4
Long/Short, n=2
VNTR1 allele count
length(bp)
frequency
cluster number
Short
Long
Lengths, n=46
Alleles, n=352
Clusters, n=3
Long/Short, n=2
VNTR2 allele count
VNTR length (bp)
e
rs465498
A
G
Cluster
VNTR1 Allele Count (EUR)
VNTR1 Length (bp)

Fig. 7

508 ***Fig. 8: Tiled CRISPRi proliferation screen nominated VNTRs represent novel***
509 ***CCVs for PDAC.***

510 **a,** Functional screening of the 5p15.33 locus nominated three MCFVs intronic to
511 *CLPTM1L*. One such MCFV, rs11133729, resides in intron nine of *CLPTM1L*, a region
512 enriched for enhancer-associated histone marks, DNAse hypersensitivity marks and
513 transcription factor ChIP-Seq peaks across seven ENCODE cell lines as demonstrated
514 from the UCSC genome browser's ENCODE regulation track (GRCh37). **b**, Analysis of
515 tiled CRISPRi screening data of the 5p15.33 locus using CRISPR-SURF nominated a
516 considerable proportion of *CLPTM1L* as proliferation-essential, including prominent
517 depletion of sgRNAs within intron 9. Aligned tracks display proliferation-essential
518 elements from CRISPR-SURF colored by magnitude of CRISPR-SURF depletion FDR.
519 For each cell line (colored by representative cancer type), screening data was analyzed
520 across three timepoints (T_0: immediately following puromycin selection and recovery, T_1:
521 ~ 12.5 cumulative screen doublings, T_2: ~ 25 cumulative doublings).**c,** Intron 9 is
522 notable for the presence of two variable number tandem repeat (VNTR) elements.
523 CRISPRi screening across tested cell lines nominated both VNTRs as proliferation-
524 essential elements. As VNTRs are known to be highly polymorphic, we utilized PacBio
525 long-read amplicon sequencing to genotype both VNTRs in 1kGP and PDAC GWAS
526 samples. **d,** VNTR genotyping in 1,810 PDAC cases and controls revealed trimodal
527 (VNTR1) and bimodal (VNTR2) allele length distributions. (VNTR allele count plotted on
528 y-axis with corresponding VNTR length on x-axis) Bar plots colored as follows;
529 aggregate VNTR lengths colored by corresponding length, distinct alleles by observed
530 frequency, clusters by k-means cluster index, long/short by binary indicator). **e,** Read-
531 based phasing of VNTR and SNP alleles in 484 1kGP EUR samples demonstrated co-
532 segregation of the rs465498 PDAC risk haplotype (top panel) with longer VNTR1 alleles
533 (clusters, bottom panel) .

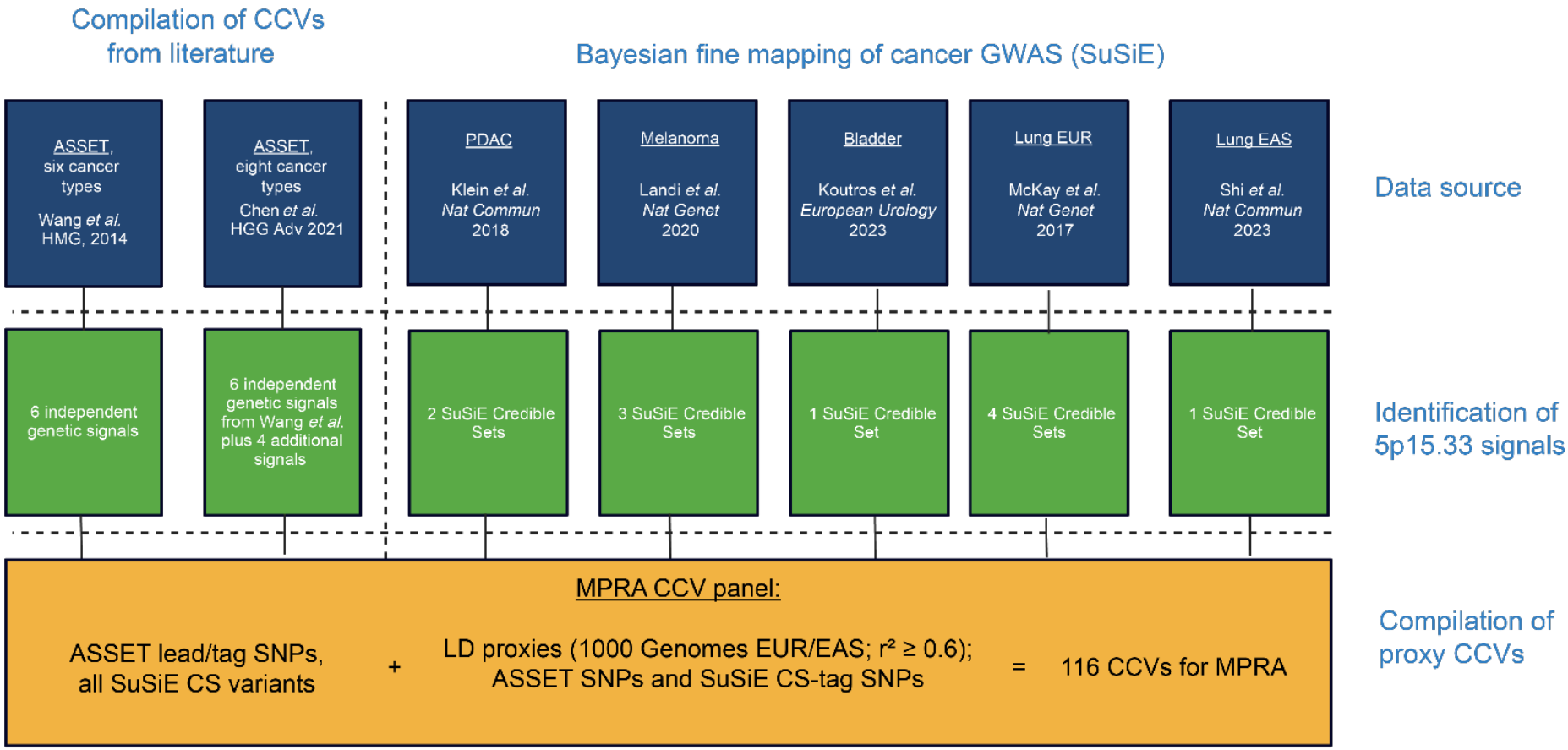

a
VNTR1 RU
TEF1 motif
Position within VNTR1
2.5 kb
2.0 kb
1.5 kb
1.0 kb
0.5 kb
0 kb
CATCCAGCTCCTCCTCTACAGACACATTC (3)
CATCCAGCTCCTCCTCTACAGGGACATTC (4)
CATCCAGCTCCTCCTCTACAGACACATTT (4)
CATCCAGCTCCTCCTCTACGGGGACATTC (2)
CATCCAGCTCCTCCTCTACAGGCACATTC (2)
CCAGCTCCTCCTCTACAGGGACATTCCAC (4)
CCAGCTCCTCCTCTACAGACACATTTCAT (2)
CAGCTCCTCCTCTACAGACACATTTCATC (2)
CAGACACATTCCATCCAGCTCCTCCTCTA (4)
CGGACACATTTCATCCAGCTCCTCCTCTA (2)
CATTTCATCCAGCTCCTCCTCTACGGACA (4)
CATTCCATCCAGCTCCTCCTCTACAGACA (20)
CATTTCATCCAGCTCCTCCTCGACAGACA (2)
CATTTC (41)
CATTCC (45)
TEF1 / TEAD1 motif
b
rs31490
VNTR1 Cluster: 3
VNTR1 L/S: Long
VNTR1 L/S: Short
VNTR2 Length: 933 bp
VNTR1 Allele: 2,934 bp #30
Type
SNP
VNTR
−log10(P)
log(OR)
c
PANC-1
A549
fold-change (2^-ΔΔCt)
sgRNA: A B C A B C
gene measured: TERT CLPTM1L
***P < 0.001
**P < 0.01
*P < 0.05
d
PANC-1
A549
firefly/renilla (luminescence A.U.)
Cluster
CTRL
pGL4.23
2,586 bp
2,789 bp
2,933 bp
3,566 bp
*P < 0.05
e
PANC-1
A549
relative luciferase expression
CTRL
0 uM
0.5 uM
1.0 uM
f
TEF1
-log10(P)
log2(unlabeled/labeled peptide intensity)
g
peptide intensity (A.U.)
ns
**P < 0.01
0x
10x
50x
competitor ratio (unlabeled:labeled)
h
Protein ladder
Input protein
Beads only
VNTR1
Beads only
VNTR1-shifted
VNTR1 - s/n
VNTR1-shifted - s/n
TEAD1
TEAD4
TAZ
YY1
SP1

534

535 ***Fig. 9***

536 ***Fig. 10: VNTR1 is a Hippo-dependent regulator of TERT and CLPTM1L.***

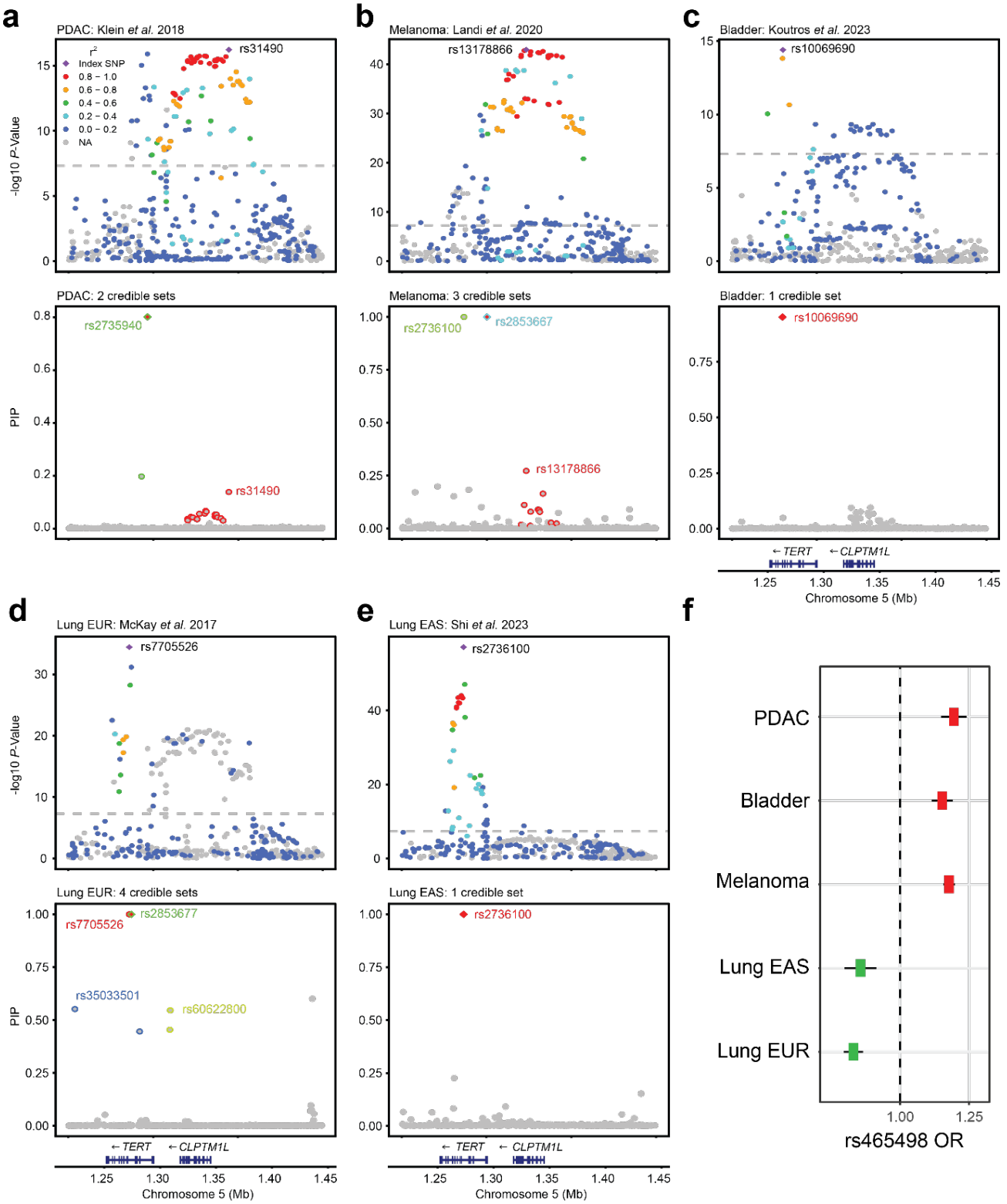
537 **a,** Motifscope plot of VNTR1 as comprised of tandemly repeating 29bp repeats (most
538 frequently observed repeat units are color-coded and shown at bottom left) which
539 themselves are enriched for TEF1 binding sites (canonical TEF1 binding motifs shown
540 in blue and cream, bottom right). **b,** Imputation of VNTR alleles into full GWAS (9,040
541 PDACs, 12,496 controls) revealed strong association for VNTR alleles (y-axis; -log10 *P*,
542 red: $P \leq 5\times10^{-8}$; logistic regression of SNP/VNTR dosage adjusted for age, sex sample
543 site and principle components) **c,** VNTR1-targeted CRISPRi using three distinct
544 sgRNAs (x-axis; sgRNAs A, B and C) reduces *TERT* and *CLPTM1L* expression (y-axis;
545 experimental sgRNA $2^{-\Delta\Delta Ct}$ normalized to $2^{-\Delta\Delta Ct}$ of non-targeting sgRNA, * indicates
546 significance under two-sided t-test) **d,** Dual luciferase reporter assays incorporating
547 cluster-representative VNTR1 sequences demonstrated strong VNTR1 enhancer
548 activity in PANC-1 and A549 cells; cluster 3 alleles show significantly higher luciferase
549 activity as compared to cluster 1 (Welch's t-test, raw firefly/renilla shown due to intensity
550 of signal (**Methods**)). **e,** Small molecule inhibition of TEF1/YAP binding negates VNTR1
551 enhancer activity, indicating Hippo pathway dependence (y-axis; firefly/renilla
552 luminescence normalized to negative control plasmid, **Methods**). **f,** DNA pulldown/mass
553 spectrometry comparing peptide intensity following incubation of Mia PaCa-2 nuclear
554 lysate with biotinylated DNA probes containing consensus VNTR1 repeat unit and
555 addition of 50-fold excess untagged competitor probe reveals VNTR1-specific TEF1
556 binding (x-axis; $\log_2$ unlabeled/labelled peptide intensity, y-axis: $-\log_{10}$ *P,* one-sided *t*-
557 test); **g**, Bar plot of TEF1-specific specific mass-spectrometry results, 10 and 50-fold
558 excess unlabeled competitor negate TEF1 intensity (independent-samples t-test). **h,**
559 Western blot following DNA oligo pulldown experiments confirms VNTR1-specific TEF1
560 binding. Binding is only seen for VNTR1-shifted probe (sequence of RU offset by 15 bp)
561 to accommodate TF motif in RU junction. Similar binding was observed for Hippo
562 pathway factors TEF-3 (gene name: *TEAD4*) and WW domain-containing transcription
563 regulator protein 1 (gene name: *WWTR1*; also named *TAZ*). Nonspecific binding to
564 beads and probes was seen for YY1 and SP1. “s/n” indicates flowthrough protein levels
565 after the pulldown



Supplementary Fig. 1

568 ***Supplementary Fig. 2: Credible causal variant (CCV) inclusion criteria for MPRA***
569 ***testing.***

570 Two studies have comprehensively identified common risk variants at 5p15.33, which
571 are associated with the risk of several cancer types. Wang et al. in 2014 used the
572 Association analysis based on SubSETs (ASSET) method of conditional and joint meta-
573 analysis of GWAS summary statistics to identify six distinct signals at the locus, which
574 was expanded upon by Chen et al. in 2021, identifying an additional four independent
575 signals. These papers served as a starting point for CCV compilation. These signals
576 were orthogonally supplemented by the application of the Bayesian fine-mapping tool
577 Sum of Single Effects (SuSiE) to individually identify independent signals from GWAS
578 summary statistics for PDAC, lung cancer, bladder cancer and melanoma. GWAS for
579 lung adenocarcinoma were separately analyzed for European and East Asian
580 ancestries. We supplemented ASSET-derived variants and variants assigned to
581 credible sets by identifying signal-tagging ASSET variants along with those variants with
582 the highest PIP for a given credible set and subsequently including correlated proxies
583 with LD $r^2 \geq 0.6$ to derive a final panel of 116 CCVs for inclusion in the MPRA pool.

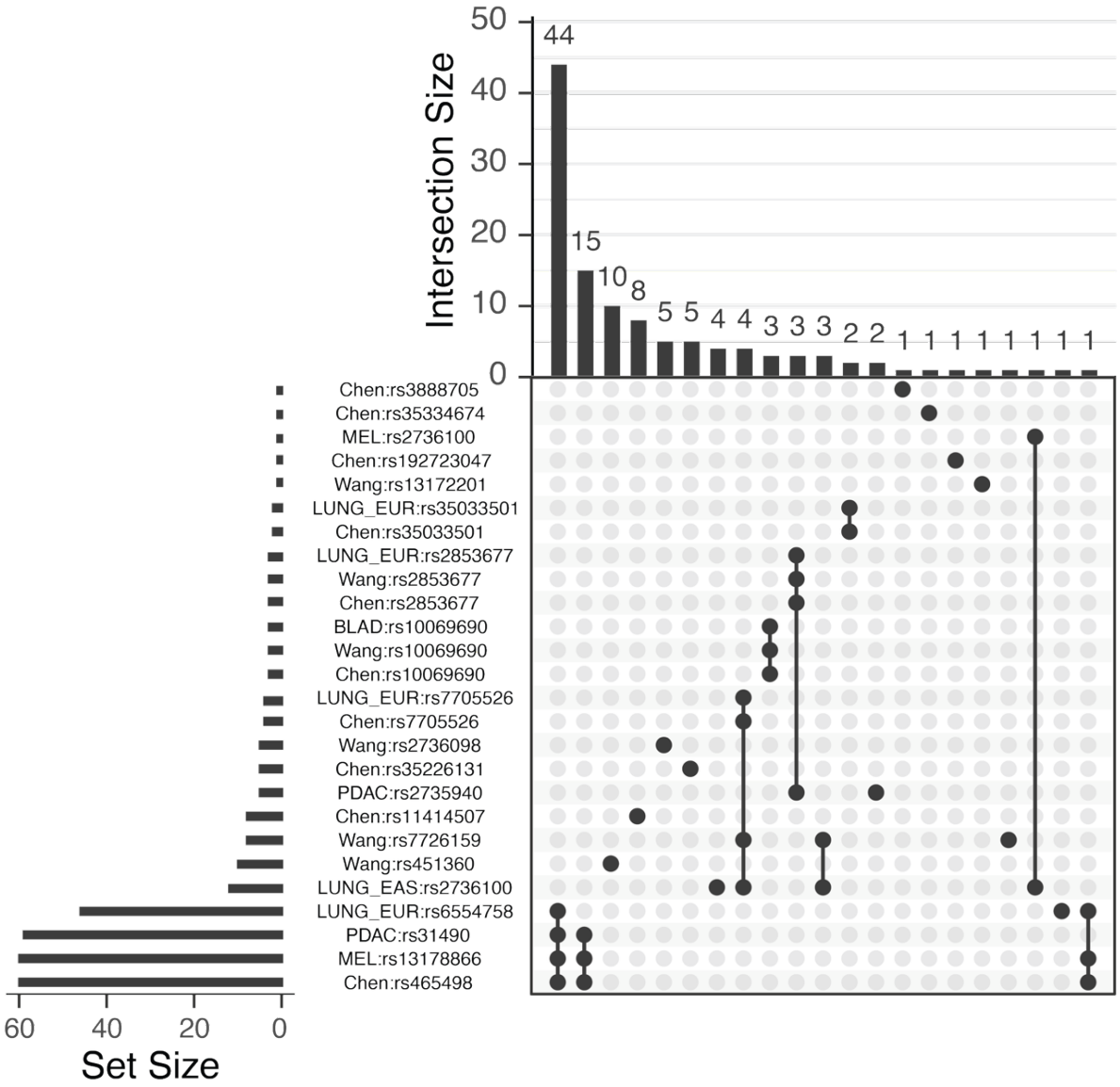


584

585 ***Supplementary Fig. 3***

586 ***Supplementary Fig. 4: Fine-mapping of 5p15.33 implicated cancer GWAS*.**

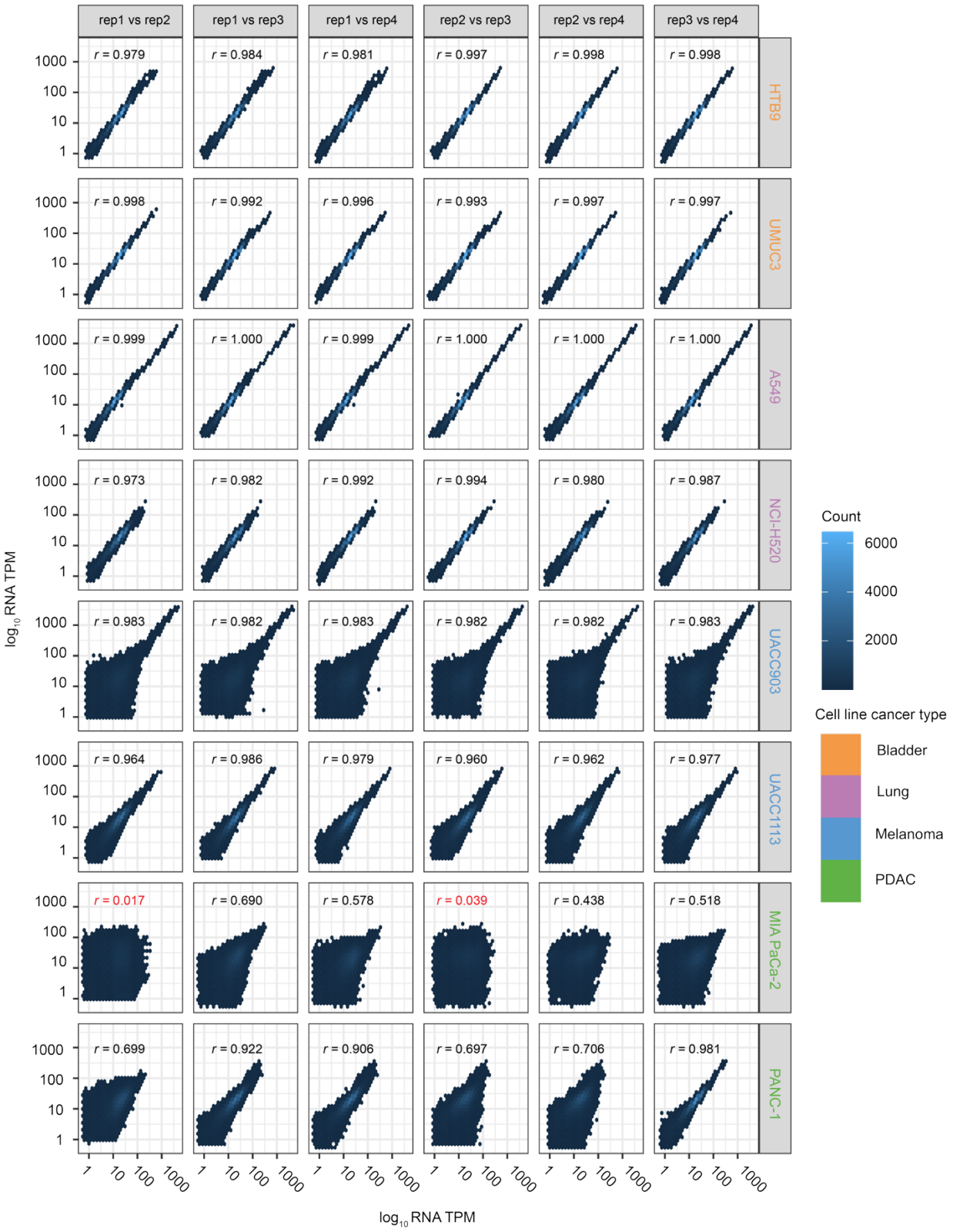
587 GWAS signals at the 5p15.33 locus in fine mapped cancers with resultant SuSiE
588 credible sets are shown with posterior inclusion probabilities (PIPs) for each variant
589 shown below for; **a,** PDAC, **b,** melanoma, **c,** bladder cancer, **d,** European (EUR) lung
590 cancer, and **e,** East-Asian (EAS) lung cancer. **f,** Odds ratios and 95% confidence
591 intervals across all fine-mapped cancers, showing antagonistic pleiotropy between
592 PDAC, melanoma, bladder cancer and lung cancer EUR/EAS GWAS for one of the
593 signals at chr5p15.33 (ASSET signal marked by rs465498).



Supplementary Fig. 5

596 ***Supplementary Fig. 6: Cross-correlation of ASSET and SuSiE-nominated 5p15.33***
597 ***signals.***

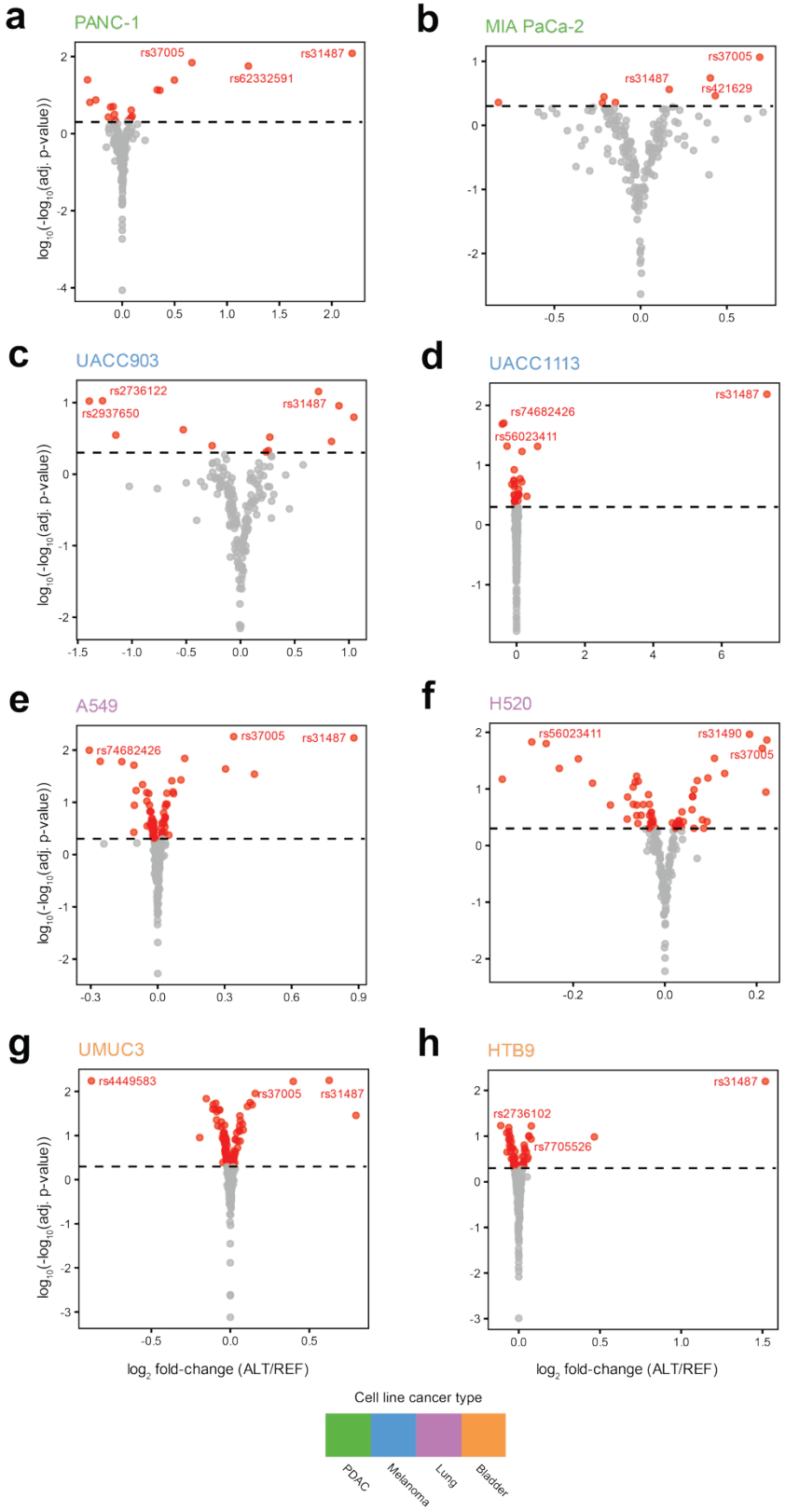
598 UpSet plot showing intersections between SuSiE fine-mapping credible sets and
599 ASSET signal-tagging SNPs identified across five cancer GWAS. Variants were
600 grouped based on linkage disequilibrium (LD $r^2 \geq 0.6$ 1kGP EUR). Horizontal bars (left)
601 indicate the total number of variants in each credible set or reported signal. Vertical bars
602 (top) show the size of intersections between sets, with connected dots (matrix below)
603 indicating which credible sets share variants. The most common intersection (n=44
604 variants) represents a core set of variants shared across multiple cancer types. Cancer
605 type abbreviations: PDAC, pancreatic ductal adenocarcinoma; MEL, melanoma; BLAD,
606 bladder cancer; LUNG_EUR, lung cancer in European ancestry; LUNG_EAS, lung
607 cancer in East Asian ancestry; Chen (Chen et al. ASSET 2021); Wang (Wang et al.
608 ASSET 2014). Results demonstrate extensive sharing of credible causal variants across
609 cancer types, with some cancer-specific signals, consistent with pleiotropic regulatory
610 mechanisms at this locus.



Supplementary Fig. 7

613 ***Supplementary Fig. 8: Inter-replicate correlation of MPRA experiments.***

614 Pairwise correlation matrices showing inter-replicate concordance for MPRA expression
615 measurements (log_{10} RNA TPM vs. log_{10} RNA TPM) in four biological replicates across
616 eight cancer cell lines. Each panel displays a hexbin scatter plot with Pearson
617 correlation coefficient (r) for comparisons between replicates (rep1 -rep4). Cell lines
618 tested: 5637 (bladder cancer), UMUC3 (bladder cancer), A549 (lung adenocarcinoma),
619 H520 (lung squamous cell carcinoma), UACC903 and UACC1113 (melanoma), MIA
620 PaCa-2 (PDAC), and PANC-1 (PDAC). Point density is indicated by color intensity (blue
621 scale). Most pairwise comparisons showed strong correlations ($r > 0.96$), except for MIA
622 PaCa-2 replicate 2 ($r = 0.017$ and $r = 0.039$, red text) and suggesting technical issues
623 with MIA PaCa-2 replicate 2 which was removed from downstream analyses. High
624 correlations across remaining replicates demonstrate robust and reproducible MPRA
625 measurements for functional variant characterization.

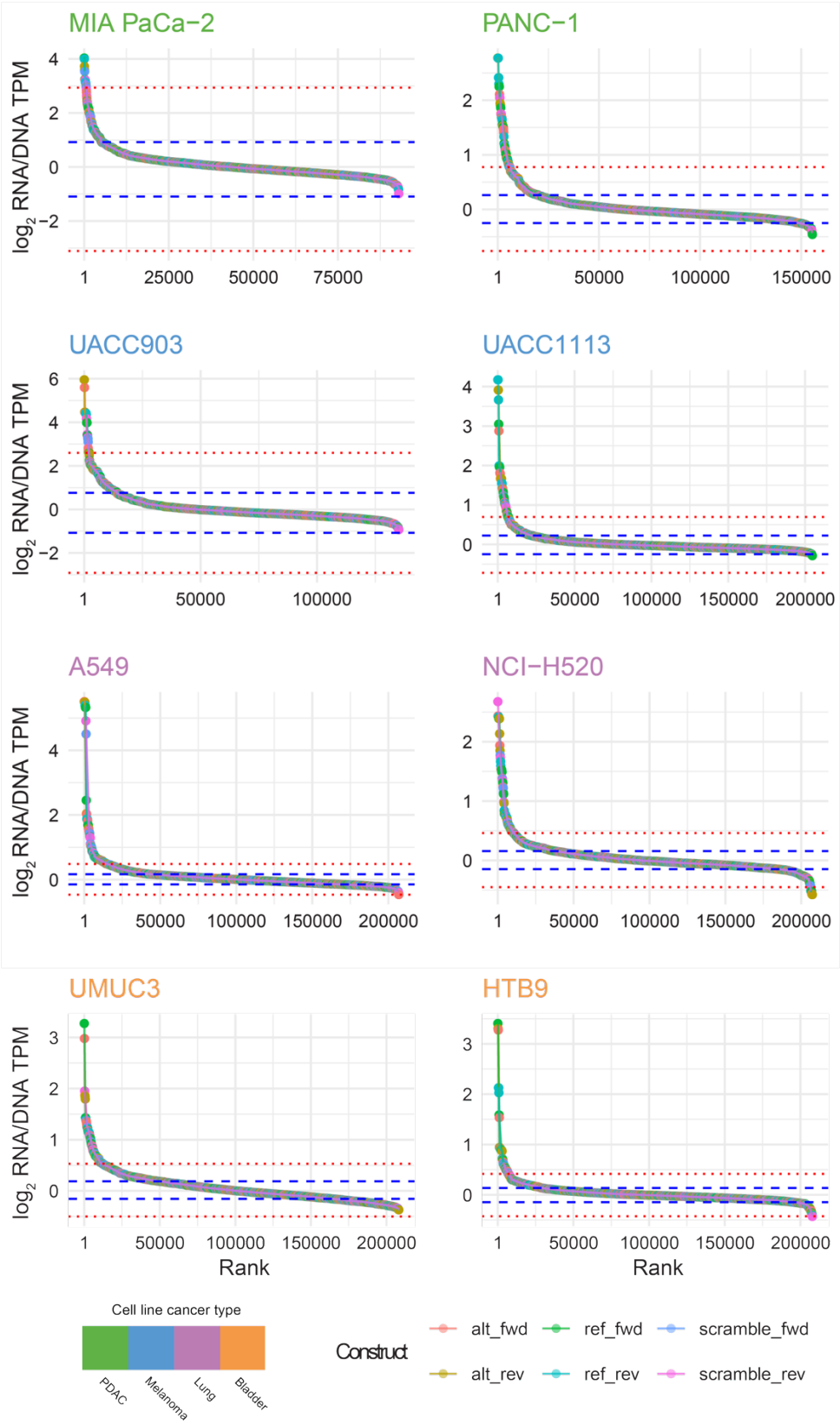


626

627 ***Supplementary Fig. 9***

628 ***Supplementary Fig. 10: MPRA results for 5p15.33 CCVs revealed variants***
629 ***exhibiting allele preferential activity across eight cell lines.***

630 Volcano plots showing allelic effects of MPRA-tested variants in; **a,** PANC-1, **b,** MIA
631 PaCa-2, **c,** UACC903, **d,** UACC1113, **e,** A549, **f,** NCI-H520, **g,** UMUC3, and **h,** 5637
632 cell lines. X-axis shows $\log_2$ fold-change in expression between alternative (ALT) and
633 reference (REF) alleles; y-axis shows statistical significance ($-\log_{10}$ adjusted p-value).
634 Red points indicate variants with significant allelic activity differences (MPRA-functional
635 variants); gray points and dashed horizontal line indicates non-significant variants (FDR
636 < 0.01). Selected variants of interest are labeled with their rsID numbers. Multiple
637 variants demonstrate significant allelic effects across cancer types, with some showing
638 cell-type-specific regulatory activity. The magnitude and direction of allelic effects vary
639 across cell lines, reflecting tissue-specific regulatory mechanisms.



Supplementary Fig. 11

642 ***Supplementary Fig. 12: Scrambled control constructs assign MPRA-functional***
643 ***CCVs to active elements***

644 Ranked distribution of log_2 RNA/DNA TPM values for all tested constructs across eight
645 cancer cell lines: MIA PaCa-2, PANC-1, UACC903, UACC1113, A549, NCI-H520,
646 UMUC3, and 5637. Constructs are ranked from highest to lowest expression and
647 colored by type and strand orientation: alternative allele forward strand (alt_fwd, pink),
648 alternative allele reverse strand (alt_rev, orange), reference allele forward strand
649 (ref_fwd, green), reference allele reverse strand (ref_rev, cyan), scrambled control
650 forward strand (scramble_fwd, blue), and scrambled control reverse strand
651 (scramble_rev, purple). Dashed blue lines indicate the first quartile (Q1, lower) and third
652 quartile (Q3, upper) of scrambled control expression within each cell line; dotted red
653 lines indicate Q1-IQR (lower suppressor threshold) and Q3+IQR (upper activator
654 threshold). Alleles with mean expression above Q3 were classified as putative
655 activators; those below Q1 were classified as putative suppressors. CCVs meeting
656 criteria for MPRA-functionality (FDR < 0.01, |fold-change| ≥ 15%, and activator or
657 suppressor classification) demonstrate activity substantially divergent from
658 transcriptionally inert scrambled controls, indicating genuine regulatory function.

Methods

Cell culture

UACC903 and UACC1113 melanoma cell lines were obtained from the University of Arizona Cancer Center (Phoenix, AZ). All other cells were obtained from American Type Culture Collection (ATCC; Manassas, VA) and maintained in standard culture conditions at 37 °C in a humidified atmosphere with 5% CO_2. All cell culture media were obtained from Thermo Fisher Scientific, Waltham, MA. **MIA PaCa-2** (CRM-CRL-1420) cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) high glucose supplemented with 10% fetal bovine serum (FBS) and 2.5% horse serum. **PANC-1** (CRL-1469), **UMUC3** and **A549** cells were maintained in DMEM high glucose with 10% FBS. **NCI-H520 (HTB-182), UACC903, UACC1113**, **5637 (**HBT9) and **TCCSUP (HTB-5)** cells were cultured in RPMI-1640 with 10% FBS. **HEK293T (CRL-3216)** cells, used for lentiviral production, were grown in high-glucose DMEM supplemented with 10% FBS, 1% glutamine, and 1% sodium pyruvate. All cell lines were routinely monitored for expected morphology and growth rates in addition to regularly being screened for mycoplasma contamination (Lonza MycoAlert Mycoplasma Detection Kit: LT07-318) which consistently tested negative. Authentication was performed using a short tandem repeat (STR) profiling panel (Identifiler kit, Life Technologies, Carlsbad, CA), and resulting profiles were compared against publicly available reference datasets from ATCC and DSMZ (https://www.dsmz.de/). All cell lines demonstrated correct STR identity based on database records.

Fine-mapping and MPRA variant selection

To derive credible sets, we fine mapped previously published GWAS for PDAC[46] (9,040 cases, 12,496 controls), melanoma[47] (36,760 cases, 375,188 controls), lung[49,50] (EUR; 29,266 cases, 56,450 controls, EAS; 21,658 cases, 150,676 controls) and bladder cancer[48] (13,790 cases, 343,502 controls). We filtered summary statistics[46–50] to include all available variants within ±500kb of the most significant 5p15.33 SNP for each cancer

686 type. We then used MungeSummStats[94] (v1.18.0) to lift genomic coordinates over to
687 GRCh38 and standardized effect alleles to the minor allele (**Supplementary Table 1**).
688 We then utilized the Sum of Single Effects (SuSiE) method[9] (v0.14.2) susie_rss()
689 function under default parameters to derive credible sets (CSs) with L = 10 (up to 10
690 possible causal variants per credible set). LD matrices for SuSiE were generated using
691 PLINK[95] (v2.0.0-a.7) from the 1000 Genomes Project Phase 3 (1kGP v3) dataset[51],
692 using genotypes from unrelated East Asian ancestry samples (EAS) for East Asian lung
693 cancer summary statistics and unrelated European ancestry (EUR) samples for the
694 remaining four GWAS. For our MPRA, we included CCVs from each of the identified
695 SuSiE credible sets in addition to the following correlated variants (1kGP v3 EUR for
696 PDAC, melanoma, bladder and lung EUR, EAS for lung EAS), using an LD threshold r^2
697 ≥ 0.6 as a compromise between inclusivity and specificity:

698 1. LD r^2 ≥ 0.6 relative to the SNP with the highest posterior inclusion probability
699 (PIP) for an identified SuSiE credible set.
700 2. LD r^2 ≥ 0.6 relative to any signal-tagging SNP identified by Chen and
701 colleagues in 2021.[5]
702 3. LD r^2 ≥ 0.6 relative to any signal-tagging SNP identified by Wang and
703 colleagues in 2014.[4]

704 This approach resulted in 116 5p15.33 CCVs which were included in the MPRA panel
705 (**Supplementary Figure 1**). In addition to these variants, we also included 44 additional
706 variants from previous MPRAs and luciferase reporter assays as positive controls
707 (**Supplementary Table 3**).

708 Epigenetic annotation of CCVs

709 CCVs were annotated with functional scoring metrics from RegulomeDB[57] (v2.2) and
710 FORGEdb[58] using the LDlink online tool[96]. CCVs were further annotated for predicted
711 disruptions or creations of transcription factor binding motifs using motifbreakR[59].
712 motifbreakR utilized TF models from HOCOMOCO v11. We used the information

713 content method and required motif matches to meet a maximum p-value of 10^{-4}. Strong
714 motifbreakR effects were defined as a difference between reference and alternate
715 alleles scores greater than 0.7. We further annotated motifbreakR results with
716 colocalization of matching TF ChIP-Seq data from the ReMap22[93] database.

717 MPRA library design

718 For each CCV, we designed oligonucleotides comprised of 145 bases of contextual
719 sequence (hg38) centered on either the effect or non-effect GWAS alleles, in both
720 forward and reverse orientations with a unique 12-mer barcode separated by
721 contiguous KpnI and XbaI restriction sites (structured oligo sequence: 5'-
722 ACTGGCCGCTTCACTG-145 bases-GGTACCTCTAGA-12 base tag-
723 AGATCGGAAGAGCGTCG-3') (Agilent Technologies, Santa Clara, CA). Barcode
724 sequences were preselected from the publicly available FREE barcodes database[12] for
725 balanced GC content, minimal homopolymer runs, and reduced hairpin propensity.
726 Additionally, we filtered barcodes to exclude any known mammalian miRNA seed
727 sequences using miRbase[13]. For each variant, we designed six constructs (reference,
728 alternate, and core 20 bp randomized in both forward and reverse orientations), with 35
729 unique barcodes per construct. Seven variants (rs11278847, rs11414507, rs3215401,
730 rs34218850, rs28379291, rs3030832, rs57516180) failed sequence design due to the
731 presence of KpnI and XbaI restriction sites within their contextual sequence.

732 MPRA plasmid pool construction, transfection and sequencing

733 To clone MPRA oligonucleotide pools into the pMPRA2 delivery vector (Addgene,
734 Watertown, MA, plasmid #49350), 10 fmol of reconstituted oligonucleotides were first
735 amplified by emulsion PCR (ePCR) using Herculase II Fusion Polymerase (Agilent) with
736 primers containing SfiI restriction sites. ePCR was performed with the Micellula DNA
737 Emulsion & Purification Kit (EURx/CHIMERx, Milwaukee, WI). The amplified
738 oligonucleotide pools were then ligated to Illumina adapter sequences using the KAPA
739 HyperPrep Kit (KAPA Biosystems, Wilmington, MA) and quantified with the KAPA qPCR
740 assay. Tag recovery rates were assessed via Illumina MiSeq v3 (150 bp, paired-end

741 reads; Illumina, San Diego, CA). ePCR products were digested with SfiI (New England
742 Biolabs, Ipswich, MA) and ligated into the complementary SfiI site of the pMPRA1
743 vector (Addgene plasmid #49349) using Electroligase (New England Biolabs, Ipswich,
744 MA) at a 1:25 vector-to-insert molar ratio. The ligation product was digested with KpnI
745 and XbaI (New England Biolabs, Ipswich, MA) between the 145 bp element and 12 bp
746 tag sequence to enable ligation of a *luc2* open reading frame (ORF) cloned from the
747 corresponding restriction sites of the pMPRA2 vector. The final ligation product was
748 electrotransformed into ONESHOT electrocompetent cells (Thermo Fisher Scientific),
749 ensuring that each construct was represented in at least ten independent
750 transformations. The cloned library intended for transfection was validated as a single
751 band on an agarose gel following KpnI digestion.

752 We performed MPRA transient transfections in eight cancer cell lines: PANC-1, MIA
753 PaCa-2, UACC903, UACC1113, A549, H520, UMUC3, and 5637 (We used 5637
754 bladder cancer cells for the MPRA (rather than TCCSUP for CRISPRi) because 5637
755 cells are amenable to transient transfection). Each cell line was transfected in biological
756 quadruplicate, with replicate cultures plated on separate days. For each transfection,
757 the number of transfected cells exceeded the library complexity by more than 100-fold
758 to ensure comprehensive representation of all sequences. 14 µg of MPRA plasmid was
759 transfected to cell cultures using Lipofectamine 3000 (Thermo Fisher Scientific)
760 according to the manufacturer's protocol. Cell densities were optimized for each line to
761 maximize transfection efficiency, which was verified by a parallel GFP transfection
762 assay and fluorescence microscopy (no less than ~ 50% transfection efficiency). Cells
763 were harvested 48 hours post-transfection for RNA isolation.

764 Total RNA was extracted with the RNeasy Kit (QIAGEN, Hilden, Germany) and
765 enriched for mRNA using the Poly(A)Purist MAG Kit (Thermo Fisher Scientific).
766 Complementary DNA (cDNA) was synthesized using SuperScript III Reverse
767 Transcriptase (Thermo Fisher Scientific). Unique 12-mer barcode sequences were
768 amplified from cDNA with Q5 High-Fidelity DNA Polymerase (New England Biolabs)
769 using primers containing Illumina TruSeq adapter sequences. Tag sequence libraries
770 were prepared from both input DNA and cDNA using identical workflows. Sequencing

771 was performed on a NovaSeq 6000 SP flow cell (Illumina, San Diego, CA) with 100 bp,
772 dual-indexed, single-end reads at NCI's Cancer Genomics Research Laboratory. For
773 each cell line, sequencing depth ranged from 150 to 300 million reads per sample,
774 providing sufficient coverage for downstream analyses (**Supplementary Table 17**).

775 MPRA data analyses

776 Using FASTQ files generated from transfected plasmid DNA and RNA transcript
777 sequencing, we quantified reads that perfectly matched each 12 bp tag sequence, along
778 with downstream regions encompassing the XbaI recognition site and 20 bp of the 3′
779 *luc2* sequence. For each transfection, tag counts per million (TPM) were calculated by
780 dividing the raw tag count by the total number of sequence-matching tag counts per
781 million reads. To infer the transcriptional activity of each element, a single pseudocount
782 was added to every TPM value and the RNA/DNA ratio for each tag was then computed
783 by dividing its RNA TPM by the corresponding DNA TPM. For each transfection, the
784 median ratio across all tags was calculated, and adjusted ratios were obtained by
785 dividing each tag's ratio by this median to ensure standardized scaling across
786 replicates. Adjusted ratios were subsequently $\log_2$-transformed to derive activity levels
787 associated with each tested tag sequence.

788 Across all eight cell lines, at least 97.05% of barcode sequences were detected in both
789 input DNA libraries and RNA samples. Reproducibility between transfections was
790 assessed using Pearson correlation. After removal of a single outlier replicate per cell
791 line, transcriptional activity levels showed high correlation across transfection replicates
792 (median Pearson's *r*; PANC-1 = 0.455, MIA PaCa-2 = 0.236, UACC903 = 0.852,
793 UACC1113 = 0.903, A549 = 0.999, NCI-H520 = 0.956, UMUC3 = 0.981, 5637 = 0.964;
794 n = 4 transfections per cell line;). We restricted analysis to barcodes with ≥100 reads in
795 either RNA or DNA libraries, retaining 89.65% of designed tags. To evaluate the
796 transcriptional impact of alternate (ALT) versus reference (REF) alleles, we applied a
797 standard linear regression model adjusting for strand orientation (forward or reverse) as
798 a binary covariate and biological replicate as a categorical covariate. Model significance
799 was assessed using the robust sandwich-type variance estimate (Wald test) to account

800 for potential heteroskedasticity in measurement error, as implemented in the R package
801 sandwich. P-values for each linear model were corrected for multiple testing using the
802 Benjamini–Hochberg procedure, applying a false discovery rate (FDR) threshold ≤ 0.01
803 to identify variants exhibiting significant transcriptional differences between alleles and
804 scrambled controls.

805 To further assess whether tested elements displayed divergent activity relative to
806 transcriptionally inert controls, the mean $\log_2$(RNA/DNA TPM) ratio for each allele was
807 compared to the distribution of $\log_2$(RNA/DNA TPM) ratios from all scrambled control
808 constructs within the same cell line. For each cell line, the interquartile range (IQR) of
809 scrambled control activity was determined: alleles with mean $\log_2$(RNA/DNA TPM)
810 ratios above the third quartile (Q3) were classified as putative activators, whereas those
811 below the first quartile (Q1) were classified as putative suppressors (**Supplementary**
812 **Fig. 6**). CCVs exhibiting an FDR <0.01 allelic effect, an absolute allelic fold change
813 ≥15% and classification as either activator or suppressor were defined as MPRA-
814 functional for a given cell line.

815 Tiled CRISPR interference (CRISPRi) proliferation screen design

816 sgRNAs targeting chr5:1,233,287–1,380,000 bp (hg19) were downloaded from the
817 UCSC Genome Browser "CRISPR Targets" track. We also incorporated 992 inert and
818 validated negative control sgRNAs in addition to sgRNAs targeting a panel of gene
819 promoters selected based on susceptibility gene status for a given cancer type or noted
820 as proliferation-essential across DepMap CRISPRi screening datasets[60,97]. CRISPRi
821 oligonucleotides were synthesized using the SurePrint Oligonucleotide Synthesis
822 Platform (Agilent) in the following configuration: 5′-AGGCACTTGCTCGTACGACG-
823 CGTCTCACACCG-[sgRNA20nt]-GTTTCGAGACG-ATGTGGGCCCGGCACCTTAA-3′.
824 A total of 27,340 sgRNAs were included in the final library, of which 25,778 tiled across
825 the 5p15.33 region. For the 116 CCVs identified in fine-mapping, there was a median of
826 14 sgRNAs mapping to within ±50bp of each CCV, with one CCV (rs62651531) failing
827 to have any within this window. rs2853672 had the most sgRNAs mapping within this
828 window, with 39 (**Supplementary Table 5**).

829 ## CRISPRi sgRNA pool generation

830 sgRNA oligonucleotides were PCR-amplified using flanking primer sites: Forward – 5′
831 AGGCACTTGCTCGTACGACG 3′; Reverse – 5′ ATGTGGGCCCGGCACCTTAA 3′.
832 Amplified products were cloned into the pXPR_050 vector (Addgene #96925) using the
833 same emulsion PCR strategy described above for MPRA. ePCR products were purified
834 with a QIAquick PCR Purification Kit (QIAGEN). The pXPR_050 backbone was digested
835 with Esp3I (NEB), gel-extracted to isolate the linearized vector, and ligated to the
836 sgRNA ePCR amplicons using Electroligase (NEB). The ligation product was
837 electroporated into ONESHOT electrocompetent cells in six independent reactions.
838 After a one-hour recovery at 37 °C in S.O.C. medium (Invitrogen; 300 rpm), serial
839 dilutions were plated to yield > 100× colony-forming units (CFUs) per sgRNA in the
840 pool. Bulk cultures were incubated overnight (37 °C, agitation at 300 rpm, 16 h) to
841 minimize selective expansion of individual clones. Plasmids were isolated via six
842 parallel endo-free maxi preps (QIAGEN), pooled, quantified by Qubit High Sensitivity
843 DNA Assay and NanoDrop 8000 (Invitrogen, Waltham, MA), and PCR-amplified.
844 Resultant amplicons were then sequenced for QC (Illumina MiSeq v3, 100 cycles,
845 single-end). FASTQ files were analyzed for sgRNA recovery and quality metrics. >99 %
846 of input sgRNAs were recovered, confirming library complexity sufficient for lentiviral
847 packaging.

848 ## Lentiviral Packaging

849 HEK293T cells were plated and grown to ~70 % confluency in six T175 flasks in virus
850 production medium (high-glucose DMEM media supplemented with 10% FBS, 1%
851 glutamine, and 1% sodium pyruvate). Transfection mixtures were prepared using
852 lipofectamine 3000 (Invitrogen) to the manufacturer's protocol, incubated 15 min at
853 room temperature, and added dropwise to cells. After 6–8 h at 37ºC, media was
854 replaced with nutrient-rich DMEM. At 48 hours post-transfection (Day 3), viral
855 supernatants were collected, centrifuged to remove debris, and filtered (0.45 µm). To
856 concentrate virus, 2× PEG precipitation buffer (Millipore Sigma, Burlington, MA) was
857 added and incubated overnight at 4 °C. On Day 4, viral particles were pelleted by

858 centrifugation, resuspended in 1 mL aliquots of 1× PBS (Quality Biological,
859 Gaithersburg, MD), and stored at −80 °C pending titration and transduction. Viral titer
860 was quantified using Lenti-X GoStix Plus (Takara Bio, Kusatsu, Japan). 20 μL of viral
861 sample were mixed with 80 μL of chase buffer, incubated 10 min at room temperature,
862 and assessed using the GoStix Plus iOS app. The control (C) and test (T) band intensity
863 ratio was mapped to batch-specific calibration curves via QR code, yielding an estimate
864 of p24 concentration and infectious titer.

865 CRISPRi screen transduction, cell culture and sequencing

866 CRISPRi screening experiments were performed across eight cell lines (MIA PaCa-2,
867 PANC-1, UACC903, UACC1113, A549, H520, UMUC3, and TCCSUP) which had been
868 transduced with lentivirus for dCas9-KRAB-ZIM3 and cultured from clonal subdivisions.
869 CRISPRi experiments were transduced at a multiplicity of infection (MOI) of 0.3 to
870 ensure single sgRNA integration per cell. Transductions were conducted at a scale
871 sufficient to maintain ~500× representation of each sgRNA throughout the screen.
872 Three biological replicates were performed with transductions performed over
873 consecutive days. For each replicate, counted cells were resuspended with lentiviral
874 supernatant and 8 μg/mL polybrene and plated at 1×10^7 cells per 15-cm dish. After 24
875 h, cultures were washed and selected with 10 μg/mL puromycin until all cells in no-virus
876 controls were eliminated (approximately 24–72 h). Post-selection, CRISPRi populations
877 were expanded to generate sufficient material for both the initial time point (T_0) and
878 continued passaging until subsequent harvests at T_1 (~12.5 doublings) and T_2 (25
879 doublings).

880 Genomic DNA (gDNA) was isolated using the Quick-DNA Genomic DNA Kit (Zymo,
881 Irvine, CA). For each biological replicate, gDNA was harvested at the initial post-
882 selection time point and again at approximately 12.5 and 25 population doublings. The
883 full quantity of gDNA from each sample was used as input for the first round of sgRNA
884 amplification (PCR1), distributing reactions such that each contained 1 μg gDNA to
885 preserve library complexity. PCR1 products were purified using a 1:1 Ampure XP bead
886 cleanup (Beckman Coulter, Brea, CA). A second amplification (PCR2) was performed

887 using the same number of reactions as PCR1 to incorporate sample-specific barcodes
888 and full Illumina adapter sequences. Each 20 µL PCR2 reaction contained 15 ng of
889 PCR1 product and PCR2-specific primers and was run for five cycles under the same
890 cycling parameters used for PCR1. Reactions were pooled and purified using the
891 QIAquick PCR Purification Kit (QIAGEN). Library size (324 bp) was verified using a
892 BioAnalyzer High Sensitivity DNA assay (Agilent). A 30 µL aliquot of each prepared
893 library was submitted to the NCI Cancer Genomics Research Laboratory (Rockville,
894 MD, USA) for sequencing on an Illumina NovaSeq SP flow cell.

895 CRISPRi screen quality control and data analyses

896 Sequencing data from the tiled CRISPRi proliferation screen spanning 5p15.33 (eight
897 cell lines; biological triplicates; T0 immediately after puromycin recovery, T1 at ~12.5
898 cumulative doublings, T2 at ~25 doublings) were processed from raw FASTQ files using
899 the standard MAGeCK-VISPR workflow (v0.5.6) to quantify sgRNA counts for the input
900 plasmid, baseline, and post-proliferation timepoints and to perform library and assay
901 QC. Across plasmid controls, we observed high read mapping efficiency (72.28–
902 73.25%), near-uniform sgRNA representation (median Gini index 0.052; range 0.0488–
903 0.055), and minimal dropout (mean reads/sgRNA 859.48; mean missing sgRNAs
904 155.69; **Supplementary Table 6**), whereas cell-line libraries showed progressive
905 selection-associated skew (Gini increasing from T0 to T2; **Supplementary Table 6**).
906 For each sample, MAGeCK was used to compute sgRNA log2 fold changes (LFCs) and
907 gene-level statistics with multiple-testing correction (FDR), including positive-control
908 evaluation using 95 proliferation-essential genes (**Supplementary Tables 7 and 9**) and
909 targeted aggregation of sgRNAs assigned to the *TERT* and *CLPTM1L* promoter regions
910 (*TERT;* chr5:1,294,628–1,298,988: GH05J001293 and *CLPTM1L*: chr5:1,343,491–
911 1,346,621: GH05J001343; **Supplementary Table 5**). Downstream analysis was
912 restricted to comparisons between individual timepoints and plasmid counts. Fine-
913 mapped CCVs were linked to perturbations by intersecting sgRNAs within ±50 bp of
914 each CCV (median 14 sgRNAs/CCV; **Supplementary Table 5**) and testing CCV-
915 assigned sgRNA sets for depletion or enrichment (**Supplementary Tables 18&19**).
916 Finally, proliferation-essential regulatory elements were inferred from sgRNA LFCs

917 using CRISPR-SURF[62](v1.0) to deconvolve localized peaks of depletion at FDR ≤ 0.05;
918 datasets in which deconvolution failed to converge (specified in **Supplementary**
919 **Results**) were excluded, and locus-level patterns were visually confirmed in IGV by
920 concordant read coverage and dropout profiles across replicates and cell lines.

921 Epigenetic silencing of SNP and VNTR loci by targeted CRISPRi

922 Individual sgRNAs corresponding targeting selected CCVs and both VNTRs were
923 cloned into the pXPR_050 CRISPRi expression vector. For each sgRNA,
924 complementary oligonucleotides containing the 20-bp spacer sequence and BsmBI-
925 compatible overhangs (Forward: 5′-CACCG[$N_{(20)}$]-3′; Reverse: 5′-AAAC[$N_{(20)}$]C-3′)
926 were synthesized (IDT) and annealed using a standard thermocycler-based annealing
927 protocol. pXPR_050 was linearized by BsmBI (NEB) digestion and gel-purified prior to
928 ligation. Annealed oligonucleotides were ligated into the digested backbone using T4
929 DNA ligase (NEB). Ligations were transformed into *E. coli* TOP10 chemically competent
930 cells (Thermo Fisher Scientific) and plated on LB-ampicillin agar plates (Quality
931 Biological). Single colonies were expanded overnight in selective LB medium, and
932 plasmid DNA was isolated using the ZymoPURE Plasmid Miniprep kit (Zymo). To
933 confirm correct sgRNA insertion, plasmid DNA from each clone was barcoded and
934 sequenced using the Oxford Nanopore Rapid Sequencing and Barcoding Kit on a single
935 Flongle flow cell (Oxford Nanopore, Oxford UK). Sequence-verified clones were
936 expanded in 50mL cultures and purified using the EndoFree Maxiprep kit (QIAGEN) for
937 downstream experiments.

938 Lentiviral particles encoding individual sgRNAs were generated by transient transfection
939 of HEK293T cells. Briefly, 4×10^6 HEK293T cells were plated in $10cm^2$ dishes in high-
940 glucose DMEM supplemented with 10% FBS, pyruvate, and glutamine (Quality
941 Biological). Cells were transfected the following day with the sgRNA expression
942 construct, psPAX2 packaging plasmid (Addgene #12260), and MD2.G envelope
943 plasmid (Addgene #12259) using Lipofectamine 3000 (Thermo Fisher Scientific) in
944 accordance with the manufacturer's recommendations. Media was replaced 24 h post-
945 transfection. Viral supernatant was collected at 48 h, clarified by low-speed

946 centrifugation, filtered (0.45 μm), and concentrated by overnight PEG precipitation at 4
947 °C. Viral pellets were resuspended in sterile PBS and stored at −80°C. For sgRNA
948 delivery, 1×10^5 PANC-1 or A549 cells stably expressing dCas9-KRAB-ZIM3 were
949 seeded in 6-well plates in selective media containing blasticidin and polybrene. 25μl of
950 concentrated virus was added to each well, and plates were centrifuged at 300 × g for 1
951 h to enhance transduction efficiency, followed by overnight incubation at 37 °C. Media
952 was replaced the next day with blasticidin- and puromycin-containing growth medium,
953 and cultured for 96 hours before harvesting total RNA.

954 Total RNA was extracted using a QIAcube connect instrument with the RNeasy mini kit
955 (QIAGEN). 1 μg of total RNA was converted to cDNA with SuperScript III reverse
956 transcriptase (Thermo Fisher Scientific). RT–qPCR was performed using a QuantStudio
957 7 Flex system. *TERT* and *CLPTM1L* expression relative to negative control sgRNAs
958 was assessed using TaqMan assays (Thermo Fisher Scientific) (*TERT:*
959 Hs00972656_m1 *CLPTM1L:* Hs00363947_m1) with Ct values normalized to *HPRT1*
960 values from the same well (Hs99999909_m1): An unpaired t-test with Welch's
961 correction was used to test for statistical significance.

962 Targeted PacBio Amplicon Sequencing of *CLPTM1L* Intron 9

963 Genomic DNA samples for HapMap trios (CEU panel for CEPH Utah residents with
964 ancestry from Northern and Western Europe, n = 72) and 1000 Genomes Project
965 Europeans (Toscani in Italia (TSI); n = 114, Finnish in Finland (FIN); n = 100, British
966 from England and Scotland (GBR); n = 100; Iberian Populations in Spain (IBS); n =100)
967 were purchased from the Coriell Institute of Medical Research Institute (Camden, NJ).
968 Genomic DNA samples for pancreatic cancer cases and controls were sourced from the
969 first two phases of the PanScan GWAS [20,21].

970 Amplicons spanning chr5:1,325,380–1,330,639 (GRCh38) were generated using
971 barcoded M13 primers and the PacBio SMRTbell prep kit v3.0 with minor protocol
972 modifications. PCR cycling consisted of 30 cycles of 98 °C denaturation (10 s) and
973 combined annealing/extension (6 min), followed by a 10 min final extension at 72 °C.

974 Amplicons were purified with AMPure beads, quantified by PicoGreen (Invitrogen),
975 barcoded with an additional 10 PCR cycles, pooled, and sequenced on a Sequel II
976 SMRT Cell. HiFi reads were processed using SMRT Analysis for circular consensus
977 sequence (CCS) generation and demultiplexed using Lima.

978 Genotyping of VNTRs and SNPs from HiFi reads

979 HiFi amplicon reads were aligned to GRCh38 using pbmm2 with a minimum read-length
980 threshold of 4,700 bp. Aligned HiFI reads were then processed with PacBio Amplicon
981 Analysis[98] (pbAA). Reads were first clustered to the target region (amplicon ±100 bp;
982 chr5:1325280–1330739) using pbAA cluster, and haplotype tags (HP) were assigned to
983 each read using pbAA bampaint. For each sample, the two modal clusters
984 (corresponding to diploid haplotype sequences) were retained, and reads from minor
985 artefactual clusters were excluded. Consensus sequences for each haplotype were
986 reconstructed, realigned to GRCh38, and used as input for tandem repeat calling with
987 TRGT[99] and SNP genotyping with DeepVariant/GLnexus[100] followed by read-based
988 phasing of VNTRs and SNPs using HiPhase[101].

989 We assessed genotyping concordance across 484 1kGP EUR samples between
990 PacBio-derived SNP calls and orthogonal short read sequence calls[51]. For 17 SNPs
991 passing quality thresholds ($MAF_{1kGP} \geq 0.01$; HWE $P_{1kGP} \geq 0.05$), minor allele frequencies
992 (MAF) derived from PacBio SNP calls were highly concordant with those from 1kGP
993 short-read WGS calls (Pearson's $r = 0.99$; **Supplementary Table 10**), with VNTR
994 genotypes also conforming to Mendelian segregation in 24 HapMap CEU trios
995 (**Supplementary Table 11**).

996 For TRGT VNTR genotypes, we derived four complementary allele-encoding
997 approaches:

998 1. Discrete VNTR alleles were maintained as distinct sequences denoted by
999 their length and assigned a unique index number (<LEN_n >).
1000 2. Distinct VNTR alleles described above were grouped together based on
1001 length irrespective of their sequence composition (<LEN>)

1002 3. VNTR lengths and their associated allele counts were used to assign lengths
1003 based on k-means clustering (4 clusters for VNTR1, 3 for VNTR2),
1004 (<Cluster_n>).
1005 4. Binary long/short VNTR genotypes were derived by combining (<L/S>).

1006 CEU trios and EUR samples were analyzed jointly. Of 503 sequenced samples, 481
1007 were matched to individuals in the 1000 Genomes Phase 3 high-coverage WGS
1008 dataset. Mendelian transmission was evaluated across 24 CEU trios using both VNTR1
1009 and VNTR2 consensus alleles with all trios demonstrating complete Mendelian
1010 concordance. To further validate genotyping accuracy, variant calls for 31 dbSNP 155-
1011 annotated variants were compared with orthogonal 1000 Genomes short-read
1012 genotypes[51]. PacBio SNP genotypes demonstrated negligible differences in MAF
1013 (Median ΔMAF= 0.015, range = 0.003-0.228) between PacBio and 1000G WGS
1014 genotypes **(Supplementary Table 11),** indicating that accompanying VNTR calls were
1015 reflective of true genotypes**.**

1016 VNTR imputation and association analysis

1017 To impute VNTR alleles into the broader PanScan GWAS[46], we generated a reference
1018 panel designed to only impute VNTR alleles whilst leveraging phase information from
1019 both PacBio-genotyped SNPs and accurately-imputed variants from the same samples.
1020 Using data from the PanScan 940 PDAC cases and 870 controls subjected to amplicon
1021 sequencing, we compiled a multisample VCF file comprised of all VNTR genotypes
1022 (each genotyping method represented as a distinct multiallelic row to be imputed as a
1023 separate variant) along with corresponding SNP calls that met MAF > 1%, call rate = 1,
1024 and Hardy–Weinberg equilibrium $P > 10^{-5}$ (11 SNPs total). Next, we took imputed
1025 dosages from the same sample set which had been imputed using the TOPMed
1026 imputation server, excluding any variants outside of chr5:0-6Mb, MAF < 1%, imputation
1027 R2 < 0.9 and any overlap with SNPs already genotyped by PacBio. VCFs were
1028 concatenated into a single file, sorted and then phased using SHAPEIT4.2[102],
1029 leveraging HiPhase-derived phase set tags (PS) as a haplotype scaffold to preserve
1030 physical phasing information. Recipient PanScan genotypes were prepared similarly to

1031 the 1,810-reference panel cohort, filtering for position, MAF and imputation R2. To
1032 accommodate downstream imputation and association frameworks, multiallelic VNTR
1033 sites were decomposed to biallelic variants using bcftools[103]. VNTR alleles were then
1034 imputed into three cohorts of recipient samples (PanScan I&II (additional 2,585 cases,
1035 2,823 controls)[66,67], PanScan III (1,597 cases, 4,799 controls)[52], and PanC4 (from the
1036 Pancreatic Cancer Case-Control Consortium[46], dbGaP Study Accession:
1037 phs000648.v2.p1; 3,884 cases, 3586 controls; approval to Laufey T. Amundadottir))
1038 using Beagle[104] (v5.4) whilst retaining hard called VNTR alleles in the 1,810 reference
1039 samples for downstream regression. Downstream analysis of linkage disequilibrium and
1040 logistic regression and conditional analyses were performed using PLINK[95] (v2.0.0-a.7).
1041 For logistic regression, we analyzed each cohort independently, including age, sex,
1042 sample collection site and eigenvectors from principal components analysis as fixed
1043 covariates as performed in the original meta-analysis[46]. We then meta-analyzed each
1044 cohort using PLINK (v1.9).

1045 Analysis of VNTR- transcription factor binding

1046 To assess VNTR-enriched transcription factor binding sites, we obtained VNTR1
1047 sequence in FASTA format from the UCSC genome browser simple repeats track
1048 (GRCh38). VNTR1 was then analyzed with the TRAP[69] (Transcription Factor Affinity
1049 Prediction) online tool. TRAP provides an enrichment measure indicative of whether a
1050 transcription factor binds the VNTR sequence more strongly than expected relative to
1051 human promoters curated by TRANSFAC (**Supplementary Table 15**). We
1052 subsequently visualized VNTR1 RU composition and TEAD motif enrichment using
1053 motifscope[105].

1054 To validate *in-silico* TF binding for VNTR1, we designed biotinylated double-stranded
1055 oligonucleotides corresponding to the VNTR1 reference repeat unit (RU) and a +15-bp
1056 shifted RU were synthesized (IDT; **Supplementary Table 20**). Protein–DNA pulldown
1057 and mass spectrometry analyses were performed at the National Cancer Institute's
1058 Protein Characterization Laboratory (Frederick, MD). Oligonucleotides were annealed in
1059 2× binding and wash buffer (10 mM Tris pH 7.5, 1 mM EDTA, 2 M NaCl) and incubated

1060 with Neutravidin beads for immobilization. For each pulldown, 200 pmol of duplex oligo
1061 was incubated with 400 µg of nuclear extract prepared from MIA PaCa-2 cells in IP
1062 buffer (50 mM Tris pH 8.0, 150 mM NaCl, 0.1% NP-40, 1 mM $MgCl_2$, 100 µM $ZnSO_4$, 1
1063 mM DTT, 5% glycerol, protease inhibitors). Competition controls included 50× and 100×
1064 excess of unlabeled duplex DNA to assess binding specificity. Following a 1hour
1065 incubation with oligo-conjugated beads, samples were washed in IP buffer,
1066 resuspended in SDS loading buffer, and submitted for mass spectrometry. Peptide
1067 identification and quantification followed established core facility pipelines. Relative
1068 peptide abundance across competitor and non-competitor conditions was compared
1069 using a one-sided *t*-test to determine significant displacement of DNA-associated
1070 proteins.

1071 VNTR repeat unit (RU) protein binding analysis by DNA pulldown

1072 Biotinylated double-stranded DNA probes were generated by annealing complementary
1073 single-stranded oligonucleotides (100 µM each strand, 3 µl per strand, 300 pmol total,
1074 IDT, **Supplementary Table 20**) comprising sequences representative of the 1 reference
1075 RU and 15bp shifted sequence as used for global proteomics. Annealed probes were
1076 immobilized on NeutrAvidin agarose beads (30 µl) and incubated with nuclear protein
1077 extracts (100 µg) prepared using the NE-PER Nuclear and Cytoplasmic Extraction Kit
1078 (Thermo Fisher Scientific, Waltham, MA). Following incubation, protein-DNA-bead
1079 complexes were pelleted by centrifugation, and both the supernatant and bead-bound
1080 fractions were collected for analysis.

1081 Proteins were eluted from the beads by heating in sample buffer. Input nuclear extract
1082 (20 µg), supernatant fractions (20 µg), and bead-bound proteins were resolved on
1083 Criterion XT 3-8% Tris-Acetate precast gels (Bio-Rad #3450129, Hercules, CA) using
1084 XT Tricine SDS running buffer (Bio-Rad #1610790) optimized for separation of 40-400
1085 kDa proteins. Proteins were transferred to 0.45 µm pore PVDF membranes (Thermo
1086 Fisher Scientific #88518) using standard wet transfer methods. Membranes were
1087 stained with 0.1% Ponceau S (G Biosciences #786-575, VWR #89167-800) to verify
1088 transfer efficiency.

1089 Membranes were blocked in 2% bovine serum albumin (BSA) in TBST (10 mM Tris-HCl
1090 pH 7.5, 150 mM NaCl, 0.05% Tween-20) and incubated overnight at 4°C with primary
1091 antibodies diluted 1:1,000 in 2% BSA/TBST: rabbit polyclonal anti-SP1 (Abcam
1092 ab13370), rabbit polyclonal anti-TAZ (Abcam ab110239), rabbit monoclonal anti-TEAD1
1093 (Abcam ab133533), mouse monoclonal anti-TEAD1 (BD Transduction 610923), rabbit
1094 polyclonal anti-TEAD4 (Thermo Fisher Scientific 12418-1-AP), mouse monoclonal anti-
1095 TEAD4 (Santa Cruz sc-101184), rabbit monoclonal anti-YAP1 (Abcam ab52771), and
1096 rabbit monoclonal anti-YY1 (Abcam ab109228). Following TBST washes, membranes
1097 were incubated for 1 h at room temperature with HRP-conjugated secondary antibodies
1098 diluted 1:2,000 in 2% BSA/TBST: donkey anti-rabbit HRP (Abcam ab205722) or donkey
1099 anti-mouse HRP (Abcam ab7061). Chemiluminescence was detected using
1100 SuperSignal West Femto Maximum Sensitivity Substrate (Thermo Fisher Scientific) and
1101 imaged on a ChemiDoc imaging system (Bio-Rad).

1102 VNTR luciferase assays

1103 To assess VNTR1 transcriptional activity, VNTR1 alleles from samples from the 1000
1104 Genomes European Ancestry Set were amplified using custom PCR primers
1105 (**Supplementary Table 20)**, allowing for Gateway cloning. HapMap samples HG00171
1106 (2,586 bp), GM20759 (2,789 bp), GM20818 (2,933 bp) and GM20787 (3,566 bp) were
1107 selected to represent both the extremes of allelic length distribution and the two modal
1108 lengths. A *negative* control construct was generated by amplifying a 2.7 kb region in
1109 intron 9 of *TERT* which was not annotated by any epigenetic marks in ENCODE or any
1110 CRISPR-SURF peaks from the screen. PCR, Gateway cloning and sequence
1111 verification were performed at the National Cancer Institute's Protein Expression
1112 Laboratory (Frederick, MD, USA) using a protocol described by Wall et al.[26]

1113 PANC-1 and A549 cell lines were seeded in 48-well plates at densities of 2×10^4 and 4
1114 $\times 10^4$ cells per well, respectively. Cells were transfected the following day with firefly
1115 luciferase reporter constructs using Lipofectamine 3000 (Invitrogen), together with 250
1116 ng of the corresponding control plasmid and 2.5 ng of a Renilla luciferase plasmid
1117 (pGL4.74). Each construct was assayed in biological triplicate with six technical

1118 replicates per experiment. Transfection media was replaced after 6 h with complete
1119 growth medium, and cells were lysed 24 h later in passive lysis buffer (Promega
1120 Madison, WI). Firefly and Renilla luciferase activities were quantified using the Dual-
1121 Luciferase Reporter Assay System on a GloMax Explorer plate reader (Promega)
1122 following the manufacturer's protocol. For Hippo pathway inhibition studies, reporter
1123 assays were performed in the presence of 0.5, or 1.0 μM verteporfin (Millipore-Sigma),
1124 alongside a DMSO vehicle control. Firefly luciferase values were normalized to Renilla
1125 luciferase for each well, averaged across technical replicates, and expressed relative to
1126 the activity of the negative control plasmid. Data were pooled across experiments, and
1127 mean values with standard deviations were plotted using ggplot2. Statistical
1128 comparisons between constructs comprised a pairwise two-sided unpaired Welch's t-
1129 test.

1130 Reagents and oligonucleotides

1131 All reagents, oligonucleotides, plasmids, antibodies, commercial kits, and DNA samples
1132 used in this study, including supplier information and catalog numbers, are listed in
1133 **Supplementary Table 20**.

Supplementary Results

MPRA functional score enrichment

Several databases based on publicly available epigenomics data exist for the functional annotation of GWAS SNPs purporting to aide in triaging of functional variants from correlated proxies. To investigate the predictive power of these tools, we assessed whether annotation scores were enriched for MPRA-validated functional variants. For each set of MPRA variants (CCVs and positive controls), we assayed SNPs stratified by cell line and in a combined model and performed Fisher's exact tests to determine whether specific RegulomeDB and FORGEdb scores were associated with MPRA hit status. We evaluated multiple definitions of functional significance: (i) statistical significance alone (FDR ≤ 0.01), (ii) statistical significance with meaningful effect size (FDR ≤ 0.01 and absolute fold change ≥ 15%), (iii) active regulatory element status (activator or suppressor designation in either allele), and (iv) the intersection of all criteria. Annotative scores were found to have limited predictive capacity in for MPRA status, with the only significant associations observed between active element status and FORGEdb scores of 8 in A549 (P_{adj} = 0.0257) regulomeDB scores of 1b in PANC-1 cells (P_{adj} = 0.0469). These findings suggest that annotative scores should be used on a cautiously across individual loci and be used as a complement to comprehensive functional screening.

CRISPRi Screening Validation

Sequencing data from the tiled CRISPRi proliferation screen of 5p15.33 across eight cell lines were processed using MAGeCK-VISPR[106] (v0.5.6) for sgRNA-level and positive control QC. FASTQ files corresponding to the plasmid library, baseline (T_0), and post-proliferation (T_1 and T_2) timepoints were analyzed using the standard MAGeCK-VISPR workflow to assess library quality, read mapping, and assay efficacy. Across all plasmid control libraries, high mapping efficiency (72.28%–73.25%) and uniform sgRNA representation were observed, with Gini indices ranging from 0.05 to 0.06, consistent

1161 with even distribution of sgRNA counts. In contrast, cell-line derived libraries exhibited
1162 selection-mediated variability through timepoints, reflected by progressive elevation of
1163 Gini indices—from 0.05–0.10 at T_0 to 0.13 at T_1 and up to 0.17 at T_2.

1164 To evaluate assay sensitivity and reproducibility, depletion and enrichment statistics
1165 were derived from the normalized sgRNA counts. Enrichment of small P-values was
1166 noted across positive-control proliferation genes (**Supplementary Table 7**), including
1167 *KRAS*, *MYC*, and *CCND1*, indicative of proliferation-mediated of associated sgRNAs.
1168 Consistent and significant depletion was observed for sgRNAs targeting the *TERT* and
1169 *CLPTM1L* promoter regions, defined as chr5:1,294,628–1,298,988 (GeneHancer ID:
1170 GH05J001293) and chr5:1,343,491–1,346,621 (GeneHancer ID: GH05J001343),
1171 respectively, encompassing 915 and 632 intersecting sgRNAs. These loci exhibited
1172 strong depletion signals across multiple cell lines, consistent with their established roles
1173 in supporting proliferative capacity (**Supplementary Table 5**).

1174 Inference of proliferation-essential regulatory elements from the tiling data was
1175 performed using CRISPR-SURF[62] (v1.0) which applies a deconvolution-based approach
1176 to aggregate sgRNA-level effects into distinct regions of proliferation-essentiality.
1177 CRISPR SURF deconvolution successfully resolved high-confidence peaks of depletion
1178 overlapping the *TERT* and *CLPTM1L* promoters in six of eight cell lines. Convolution
1179 failed to converge for MP2 (T_1), UACC903 (all timepoints), and H520 (T_1–T_2) datasets,
1180 which were excluded from downstream integration analyses. Visual inspection of
1181 aligned read coverage and sgRNA depletion profiles in Integrative Genomics Viewer
1182 (IGV) confirmed reproducible patterns of promoter-specific dropout across cell lines.
1183 Together, these analyses demonstrated that the CRISPRi proliferation screens
1184 achieved robust coverage, reproducible sgRNA representation, and consistent detection
1185 of proliferation-essential loci, including canonical oncogenes and core promoter
1186 elements within the 5p15.33 region.

1187 For the 116 CCVs identified in fine mapping, there was a median of 14 sgRNAs
1188 intersecting within ± 50bp of each CCV, with one CCV (rs62651531) failing to have any
1189 within this window. rs2853672 had the most intersecting sgRNAs, with 39

1190 (**Supplementary Table 5**). We annotated 915 and 632 of the tiling sgRNAs as
1191 intersecting the *TERT* and *CLPMT1L* promoters, respectively (Methods). We
1192 transduced lentiviral packaged sgRNAs into the same cell lines used for MPRA which
1193 were clonally engineered to stably express dCas9-KRAB-ZIM3 (**Supplementary Table**
1194 **4**). Screens were performed in biological triplicate for 25 cumulative cell doublings and
1195 harvested at an initial time point (T_0, immediately following puromycin selection and
1196 initial recovery), T_1 (~12.5 cumulative screen doublings), and T_2 (25 screen doublings).

1197 Following sequencing of sgRNAs from harvested time points and the primary plasmid,
1198 we assessed library quality, read mapping, and assay efficacy using the output of
1199 MAGeCK. Sequencing libraries from input plasmid and cell line demonstrated high
1200 mapping efficiency, even sgRNA read distribution and minimal sgRNA dropout (Average
1201 reads/sgRNA = 859.48, average missing sgRNAs = 155.69, **Supplementary Table 6**).
1202 Gini indices, a measure of sgRNA read distribution across a sample, increased
1203 progressively through plasmid (median Gini = 0.052, range = 0.0488-0.055) and cell line
1204 timepoints from (range T_0 = 0.065, 0.0524-0.175, T_1 = 0.079, 0.055-0.1548, and T_2 =
1205 0.085, 0.052-.1743), indicative of CRISPRi-mediated negative selection of sgRNAs over
1206 time.

1207 We used MAGeCK to calculate $\log_2$ fold changes for each timepoint relative to the
1208 plasmid in addition to its preceding timepoint and measured inter-replicate Pearson
1209 correlation to assess reproducibility (**Supplementary Fig. 6**). Across all comparisons
1210 (three replicate pairs per experiment), the inter-replicate Pearson correlation coefficients
1211 ranged from –0.42 to 1.00, with a median of 0.46 (IQR 0.20–0.81), indicating that while
1212 some comparisons show very strong agreement between replicates, others are
1213 considerably noisier. Reproducibility was highest for the plasmid control samples. For
1214 plasmid experiments the mean inter-replicate correlation was 0.78 (median 0.86), and
1215 ~76% of replicate pairs had $r > 0.7$. In contrast, cell line derived time-point contrasts
1216 were less consistent: T_1_vs_T_0 / T_2_vs_T_0 comparisons across all cell lines had a
1217 median r of 0.20. T_2_vs_T_1 comparisons were lower still (median 0.05), with several
1218 replicate pairs near zero or weakly negative. There was also a clear trend with time: at
1219 T0, replicate correlations were generally high (mean 0.68, median 0.82), decreasing at

1220 T_1, (median *r* = 0.54) and further at T_2, with a median *r* =0.27. By cell line, PANC1, TCC,
1221 UMUC3, A549 and MP2 showed the best overall replicate concordance (median *r* range
1222 across timepoints = 0.54–0.80), whereas UACC903, H520 and UACC1113 were more
1223 variable (median r = 0.15–0.29). With this in mind, we restricted downstream analysis
1224 only to comparisons between sgRNA counts derived from cell line time points and their
1225 corresponding count in the input plasmid.

1226 To assess the efficacy of CRISPRi systems, we assessed knockdown of 95 genes
1227 selected for proliferation-essentiality and cancer-type susceptibility status using
1228 MAGeCK. Across all timepoints and cell lines, we observed a median of 39 knocked
1229 down genes at FDR ≤ 0.05 (41.05%, range 5-60, **Supplementary Tables 7 and 9**).
1230 Depleted genes varied considerably between cell lines and cancer types, suggestive of
1231 divergent physiology. Ranking each gene within a timepoint in ascending order of
1232 depletion FDR, we found top ranked associations for canonical oncogenes such as
1233 *KRAS* and *CCND1* in most cell lines across time points (**Supplementary Table 7**).
1234 Depletion testing of both genes in UACC903 melanoma cell lines was consistent with
1235 subsequent confirmation of progressive reduction in dCas9-KRAB-ZIM3 expression
1236 (**Supplementary Fig. 7**). We noted no significant knockdown of the canonical tumor
1237 suppressor *TP53*, concordant with the depletion-based design of our screen.

1238 To further validate our experimental design, we assigned sgRNAs from the tiling
1239 component of the pool to either the *TERT* or *CLPTM1L* promoter and tested both genes
1240 using MAGeCK Analysis of *TERT* promoter targeting-sgRNAs demonstrated striking cell
1241 line-specific effects, suggestive of both physiological differences and experimental
1242 limitations. NCI-H520 lung cancer cells showed the most dramatic depletion of *TERT*
1243 sgRNAs, with $\log_2$ fold changes progressing from -0.028 at T_0 to -1.297 at T_2 (FDR =
1244 3.30×10^{-5}), indicating strong growth dependency on the *TERT* promoter. A549 cells
1245 exhibited more modest but consistent depletion of *TERT* promoter sgRNAs across
1246 timepoints ($\log_2$ fold change = -0.062, FDR = 2.55×10^{-4} at T_2). Among melanoma lines,
1247 UACC1113 demonstrated significant growth dependency on the *TERT* promoter ($\log_2$
1248 fold change= -0.217, FDR = 1.90×10^{-5} at T_2), while UACC903 showed minimal effects
1249 across all timepoints (maximum 7 total depleted positive control genes). Bladder cancer

line 5637 transitioned from modest enrichment at T_0 (LFC = 0.030) to significant depletion by T_2 (LFC = -0.082, FDR = 3.30×10^{-5}), whereas UMUC3 maintained consistent low-level enrichment (LFC ~0.04) with marginal significance. Pancreatic lines showed variable patterns: MIA PaCa-2 exhibited minimal fold changes but significant FDRs, while PANC-1 progressed from non-significant enrichment at T0 to moderate depletion at later timepoints (LFC = -0.080, FDR = 2.10×10^{-5} at T2). *CLPTM1L* promoter silencing generally paralleled *TERT* promoter effects but with attenuated magnitude. NCI-H520 again displayed the strongest dependency (LFC = -0.703, FDR = 4.32×10^{-4} at T_2), followed by UACC1113 (LFC = -0.167, FDR = 1.14×10^{-3} at T_2). Notably, several samples showed significant *CLPTM1L* promoter effects at early timepoints: UACC1113 at T_0 (FDR = 4.32×10^{-4}) and MIA PaCa-2 at T_0 (FDR = 4.50×10^{-5}), suggesting immediate growth consequences upon *CLPTM1L* promoter perturbation in these contexts. A549 showed modest depletion (LFC = -0.059 at T2), while UMUC3 and PANC-1 displayed weak or non-significant effects (**Supplementary Table 8**).

We then went on to apply MAGeCK to 5p15.33 CCVs, assigning sgRNAs to CCVs by intersecting them with ± 50 bp each fine mapped CCV. SNP-level analysis identified 5 to 54 significantly depleted variants per sample, with depletion counts generally tracking overall screen sensitivity (**Supplementary Table 9).** NCI-H520 yielded the highest SNP hits at T_2 (n=54), coinciding with maximal promoter effects and suggesting coordinated regulatory dependencies across the locus. 5637 and A549 maintained consistent SNP depletion across time points while melanoma line UACC903 showed minimal regulatory variant sensitivity (n=5 across all timepoints), consistent with its weak overall selection. PANC-1 demonstrated a notable increase in depleted SNPs from T_0 (n=10) to T_1/T_2 (n=35), suggesting that SNP-mediated *cis*-regulatory effects may take a longer period to manifest in phenotypic divergence. While our we primarily hypothesized that functional SNPs would occupy loci which would be epigenetically silenced upon CRISPRi inhibition, we observed six SNPs for which sgRNAs were significantly enriched (**Supplementary Table 10).**

1279 Next, we used CRISPR-SURF[8] to deconvolve sgRNA log_2 fold changes for each
1280 timepoint into discrete regions of putative functional regulatory activity. Unlike gene-
1281 centric or SNP-focused analyses, CRISPR-SURF operates in an unbiased, genome-
1282 agnostic manner, identifying functional regulatory regions based solely on local
1283 enrichment of depleted sgRNAs without prior assumptions about annotated features or
1284 genetic markers. This approach enabled discovery of both promoter-proximal and distal
1285 regulatory elements, as well as unannotated functional sequences across the tiled
1286 5p15.33 region (**Supplementary Fig. 9**). CRISPR-SURF deconvolution failed for MIA
1287 PaCa-2 T_1 and NCI-H520 T_1, and T_2.and UACC903 T_0. Of the remaining datapoints, we
1288 identified a total of 2,718 proliferation-essential elements at FDR ≤ 0.05, with a median
1289 length of 470 bp.

1290 Acknowledgments

1291 This research was supported in part by the Intramural Research Program of the
1292 National Institutes of Health (NIH). The contributions of the NIH authors are considered
1293 Works of the United States Government. The findings and conclusions presented in this
1294 paper are those of the authors and do not necessarily reflect the views of the NIH or the
1295 U.S. Department of Health and Human Services

1296 The authors acknowledge the research contributions of the NCI Cancer Genomics
1297 Research Laboratory for their expertise, execution, and support of this research in the
1298 areas of project planning, wet laboratory processing of specimens, and bioinformatics
1299 analysis of generated data. We thank Dr. Raj Chari, Director, Genome Modification
1300 Core, Frederick National Laboratory for Cancer Research, National Cancer Institute,
1301 Frederick, MD, for assistance with CRISPR guide design. We also thank Dr. Dominic
1302 Esposito, Director, Protein Expression Laboratory (PEL) Frederick National Laboratory
1303 for Cancer Research, National Cancer Institute, Frederick, MD, for assistance in cloning
1304 of VNTR luciferase constructs.

1305 This work utilized the computational resources of the NIH HPC Biowulf cluster
1306 (https://hpc.nih.gov).

SL was supported by CA297773 (NIH).

The Pancreatic Cancer Cohort Consortium and the Pancreatic Cancer Case-Control Consortium (PanScan/PanC4) Acknowledgements and Funding statements

American Cancer Society. The authors express sincere appreciation to all Cancer Prevention Study-II participants, and to each member of the study and biospecimen management group. The authors would like to acknowledge the contribution to this study from central cancer registries supported through the Centers for Disease Control and Prevention's National Program of Cancer Registries and cancer registries supported by the National Cancer Institute's Surveillance Epidemiology and End Results Program. The American Cancer Society funds the creation, maintenance, and updating of the Cancer Prevention Study-II cohort.

The **ATBC Study** was supported by the Intramural Research Program, Division of Cancer Epidemiology and Genetics of the U.S. National Cancer Institute (NCI), National Institutes of Health.

Cancer incidence data for **CLUE** were provided by the Maryland Cancer Registry, Center for Cancer Surveillance and Control, Department of Health and Mental Hygiene, 201 W. Preston Street, Room 400, Baltimore, MD 21201, http://phpa.dhmh.maryland.gov/cancer, 410-767-4055. We acknowledge the State of Maryland, the Maryland Cigarette Restitution Fund, and the National Program of Cancer Registries of the Centers for Disease Control and Prevention for the funds that support the collection and availability of the cancer registry data." We thank all the CLUE participants.

EPIC-Europe Funding Statement and Disclaimer:

EPIC Financial support: The coordination of EPIC-Europe is financially supported by International Agency for Research on Cancer (IARC) and also by the Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London which

has additional infrastructure support provided by the NIHR Imperial Biomedical Research Centre (BRC).

The national cohorts are supported by: Danish Cancer Society (Denmark);Ligue Nationale Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale (MGEN), Institut National de la Santé et de la Recherche Médicale (INSERM), French National Research Agency (ANR, reference ANR-10-COHO-0006), French Ministry for Higher Education (subsidy 2102918823, 2103236497, and 2103586016) (France); German Cancer Aid, German Cancer Research Center (DKFZ), German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Federal Ministry of Research, Technology and Space (BMFTR) (Germany); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy, Italian Ministry of Health, Italian Ministry of University and Research (MUR), Compagnia di San Paolo (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), the Netherlands Organisation for Health Research and Development (ZonMW), World Cancer Research Fund (WCRF), (The Netherlands); UiT The Arctic University of Norway; Health Research Fund (FIS) - Instituto de Salud Carlos III (ISCIII), Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, and the Catalan Institute of Oncology - ICO (Spain); Swedish Cancer Society, Swedish Research Council of Region Skåne and Region Västerbotten (Sweden); Cancer Research UK (C864/A14136 to EPIC-Norfolk; C8221/A29017 to EPIC-Oxford), Medical Research Council (MR/N003284/1, MC-UU_12015/1 and MC_UU_00006/1 to EPIC-Norfolk (DOI 10.22025/2019.10.105.00004); MR/Y013662/1 to EPIC-Oxford) (United Kingdom). Previous support has come from "Europe against Cancer" Programme of the European Commission (DG SANCO).

EPIC Disclaimer: Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

The Health Professionals Follow-up Study is supported by NIH grant UM1 CA167552. from the National Cancer Institute, Bethesda, MD USA.

The work at **Johns Hopkins University** was supported by the NCI Grants P50CA062924 and R01CA154823. Additional support was provided by the Lustgarten Foundation, Susan Wojcicki and Dennis Troper and the Sol Goldman Pancreas Cancer Research Center.

The **Multiethnic Cohort Study (MEC)** is supported by NCI grant U01CA164973.

The **MD Anderson case control study** was supported by NIH R01 CA98380.

The **Melbourne Collaborative Cohort Study (MCCS)** cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further augmented by Australian National Health and Medical Research Council grants 209057, 396414 and 1074383 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry and the Australian Institute of Health and Welfare, including the National Death Index and the Australian Cancer Database.

Work at **Memorial Sloan Kettering Cancer Center** was funded by the center grant P30CA008748 (Selwybn Vickers).

The **NYU-WHI** study was funded by NIH R01 CA098661, UM1 CA182934 and center grants P30 CA016087 and P30 ES000260.

The **Nurses' Health Study** is supported by NIH grants UM1 CA186107, P01 CA87969, and R01 CA49449 from the National Cancer Institute, Bethesda, MD USA.

The **Physicians' Health Study** was supported by research grants CA-097193, CA-34944, CA-40360, HL-26490, and HL-34595 from the National Institutes of Health, Bethesda, MD USA

The **Queensland Pancreatic Cancer Study** was supported by a grant from the National Health and Medical Research Council of Australia (NHMRC) (Grant number 442302). RE Neale is supported by a NHMRC Senior Research Fellowship (#1060183).

The **SELECT** study is supported by National Institutes of Health grant award number UG1CA189974 (CD Blanke), and U01 CA182883 (CM Tangen/IM Thompson). The authors thank the site investigators and staff and, most importantly, the participants who donated their time to this trial.

Data collection at the **University of Toronto** was supported by the CIHR Canada Research Chair and from the Canadian Cancer Society Research Institute.

The **WHI** program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005.The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: https://www-whi-org.s3.us-west-2.amazonaws.com/wp-content/uploads/WHI-Investigator-Short-List.pdf.

The **Harvard Women's Health Study** funding was as follows: CA047988, CA182913, HL043851, HL080467 and HL099355.

The **Yale (CT) pancreas cancer study** is supported by National Cancer Institute at the U.S. National Institutes of Health, grant 5R01CA098870. The cooperation of 30 Connecticut hospitals, including Stamford Hospital, in allowing patient access, is gratefully acknowledged. The Connecticut Pancreas Cancer Study was approved by the State of Connecticut Department of Public Health Human Investigation Committee. Certain data used in that study were obtained from the Connecticut Tumor Registry in the Connecticut Department of Public Health. The authors assume full responsibility for analyses and interpretation of these data.

Consortia authors from the Pancreatic Cancer Cohort Consortium and Pancreatic Cancer Case-Control Consortium (PanScan/PanC4)

Jun Zhong[1], Demetrius Albanes[2], Gabriella Andreotti[3], Samuel O. Antwi[4], Alan A Arslan[5], William R. Bamlet[6], Laura Beane-Freeman[3], Sonja I Berndt[3], Paige M. Bracci[7], Daniele Campa[8], Federico Canzian[9], Stephen J Chanock[10], Yu Chen[11], Charles C Chung[12], Mengmeng Du[13], A. Heather Eliassen[14,15,16], Steven Gallinger[17], J. Michael Gaziano[18,19], Jeanine Genkinger[20], Michael Goggins[21], Phyllis J Goodman[22], Christopher A Haiman[23], Manal Hassan[24], Belynda Hicks[12], Rayjean J. Hung[25], Amy Hutchinson[12], Verena Katzke[26], Manolis Kogevinas[27,28], Charles Kooperberg[29], Martijn P Kolijn[30,31], Peter Kraft[32], Robert C. Kurtz[33], I-Min Lee[34,14], Loic LeMarchand[35], Núria Malats[36], Roger Milne[37,38,39], Steven C Moore[2], Rachel E. Neale[40], Kimmie Ng[41], Ann L. Oberg[6], Irene Orlow[42], Alpa V Patel[43], Ulrike Peters[44], Miquel Porta[45], Kari G. Rabe[46], Francisco X Real[47,48], Harvey A. Risch[49], Nathaniel Rothman[10], Paul Scheet[50], Howard D Sesso[18,14], Veronica W Setiawan[51], Xiao-Ou Shu[52], Debra Silverman[10], Mingyang Song[14,53,54], Meir J Stampfer[14,15,16], Melissa C Southey[37,38,39], Connie Thurman[14], Geoffrey S Tobias[10], Therese Truong[55], Caroline Um[43], Kala Visvanathan[56,57], Nicolas Wentzensen[10], Jean Wactawski-Wende[58], Herbert Yu[59], Chen Yuan[41], Wei Zheng[52], Brian M Wolpin[41], Rachael Z Stolzenberg-Solomon[10], Alison P. Klein[60,61], Laufey T Amundadottir[1]

[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA, [2]Metabolic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA, [3]Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA, [4]Department of Quantitative Health Sciences, Mayo Clinic College of Medicine, Jacksonville, FL, USA, [5]Departments of Obstetrics and Gynecology and Population Health, NYU Grossman School of Medicine, NYU Perlmutter Comprehensive Cancer Center, New York, NY, USA, [6]Department of Quantitative Health Sciences, Mayo Clinic College of Medicine, Rochester, MN, USA, [7]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA, [8]Unit of Genetics., Department of Biology, University of Pisa, Pisa, Italy, [9]Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany, [10]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA, [11]Department of Population Health, NYU Grossman School of Medicine, NYU Perlmutter Comprehensive Cancer Center, New York, NY, USA, [12]Cancer Genomics Research Laboratory, Frederick National Lab for Cancer Research, Frederick, MD, USA, [13]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center,, New York, NY, USA, [14]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [15]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA, [16]Department of Nutrition, Harvard T. H. Chan School of Public Health, Boston, MA, USA, [17]Lunenfeld-Tanenbaum Research Institute, Sinai Health System and University of Toronto,, Toronto, Ontario, Canada, [18]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA, [19]Division of Aging, Brigham and Women's Hospital, Boston, MA, USA, [20]Department of Epidemiology, Columbia University, New York, NY, USA, [21]Department of Pathology, Johns Hopkins School of Medicine, Baltimore, MD, USA, [22]SWOG Statistical Center, Fred Hutchinson Cancer Center, Seattle, WA, USA, [23]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA, [24]Department of Epidemiology, Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, TX, USA, [25]Prosserman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada, [26]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany, [27]ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain, [28]Hospital del Mar Institute of Medical Research (IMIM), Universitat Autònoma de Barcelona, Barcelona, Spain, [29]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA, USA, [30]Division of Environmental Epidemiology and Veterinary Public Health, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands, [31]Julius Global Health, The Julius Center for Health Sciences and Primary Care, University Medical Center, [32]Trans-Divisional Research Program (TDRP), Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA, [33]Gastroenterology, Hepatology, and Nutrition Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA, [34]Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA, [35]Population Sciences in the Pacific Program, Cancer Epidemiology, University of Hawaii Cancer Center, Honolulu, HI, USA, [36]Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Center (CNIO), Madrid, Spain, [37]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, VIC, Australia, [38]Centre for Epidemiology and

Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC, Australia, [39]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, VIC, Australia, [40]Department of Population Health, QIMR Berghofer Medical Research Institute, Queensland, Australia, [41]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Harvard University, Boston, MA, USA, [42]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA, [43]Department of Population Science, American Cancer Society, Atlanta, GA, USA, [44]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA, USA, [45]Hospital del Mar Institute of Medical Research (IMIM), Centro de Investigacion Biomedica en Red de Epidemiologia y Salud Publica (CIBERESP), Barcelona, Spain, [46]Department of Quantitative Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN, USA, [47]Epithelial Carcinogenesis Group, Tumor Biology Programme, Spanish National Cancer Research Center (CNIO), Madrid, Spain, [48]Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain, [49]Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT, USA, [50]Dept of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA, [51]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA, [52]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA, [53]Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [54]Clinical and Translational Epidemiology Unit and Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA, [55]Paris-Saclay University, UVSQ, Inserm, Gustave Roussy, CESP, Villejuif, France, [56]Department of Epidemiology, Johns Hopkins School of Public Health, Baltimore, MD, USA, [57]Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD, USA, [58]Department of Epidemiology and Environmental Health, University of Buffalo, Buffalo, NY, USA, [59]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA, [60]Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD, USA, [61]Department of Medicine, Division of Gastroenterology & Hepatology, Johns Hopkins School of Medicine, Baltimore, MD, USA.

1495 # References

1496 1. Alsheikh, A. J. *et al.* The landscape of GWAS validation; systematic review
1497 identifying 309 validated non-coding variants across 130 human diseases. *BMC*
1498 *Med Genomics* **15**, 74 (2022).

1499 2. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional
1500 variation in humans. *Nature* **501**, 506–511 (2013).

1501 3. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature*
1502 **550**, 204–213 (2017).

1503 4. Wang, Z. *et al.* Imputation and subset-based association analysis across different
1504 cancer types identifies multiple independent risk loci in the TERT-CLPTM1L region
1505 on chromosome 5p15.33. *Human Molecular Genetics* **23**, 6616–6633 (2014).

1506 5. Chen, H. *et al.* Large-scale cross-cancer fine-mapping of the 5p15.33 region
1507 reveals multiple independent signals. *Human Genetics and Genomics Advances* **2**,
1508 100041 (2021).

1509 6. Bhattacharjee, S. *et al.* A Subset-Based Approach Improves Power and
1510 Interpretation for the Combined Analysis of Genetic Association Studies of
1511 Heterogeneous Traits. *The American Journal of Human Genetics* **90**, 821–835
1512 (2012).

1513 7. Fang, J. *et al.* Functional characterization of a multi-cancer risk locus on
1514 chr5p15.33 reveals regulation of TERT by ZNF148. *Nature Communications* **8**, 1–
1515 15 (2017).

1516 8. Kim, N. W. *et al.* Specific Association of Human Telomerase Activity with Immortal
1517 Cells and Cancer. *Science* **266**, 2011–2015 (1994).

1518 9. Castelo-Branco, P. *et al.* Methylation of the TERT promoter and risk stratification of
1519 childhood brain tumours: an integrative genomic and molecular study. *The Lancet*
1520 *Oncology* **14**, 534–542 (2013).

1521 10. Shay, J. W. & Bacchetti, S. A survey of telomerase activity in human cancer.
1522 *European Journal of Cancer* **33**, 787–791 (1997).

1523 11. Yoo, S. S. *et al. TERT* Polymorphism rs2853669 Influences on Lung Cancer Risk
1524 in the Korean Population. *J Korean Med Sci* **30**, 1423 (2015).

1525 12. Bojesen, S. E. *et al.* Multiple independent variants at the TERT locus are
1526 associated with telomere length and risks of breast and ovarian cancer. *Nat Genet*
1527 **45**, 371–384 (2013).

1528 13. Beesley, J. *et al.* Functional Polymorphisms in the TERT Promoter Are Associated
1529 with Risk of Serous Epithelial Ovarian and Breast Cancers. *PLOS ONE* **6**, e24987
1530 (2011).

1531 14. Walsh, K. M. *et al.* Variants near TERT and TERC influencing telomere length are
1532 associated with high-grade glioma risk. *Nat Genet* **46**, 731–735 (2014).

1533 15. Mosrati, M. A. *et al.* TERT promoter mutations and polymorphisms as prognostic
1534 factors in primary glioblastoma. *Oncotarget* **6**, 16663–16673 (2015).

1535 16. Harland, M. *et al.* Germline TERT promoter mutations are rare in familial
1536 melanoma. *Familial Cancer* **15**, 139–144 (2016).

1537 17. George, J. R. *et al.* Association of TERT Promoter Mutation, But Not BRAF
1538 Mutation, With Increased Mortality in PTC. *J Clin Endocrinol Metab* **100**, E1550–
1539 E1559 (2015).

1540 18. Rachakonda, P. S. *et al.* TERT promoter mutations in bladder cancer affect patient
1541 survival and disease recurrence through modification by a common polymorphism.
1542 *Proceedings of the National Academy of Sciences* **110**, 17426–17431 (2013).

1543 19. Pópulo, H. *et al.* TERT promoter mutations in skin cancer: the effects of sun
1544 exposure and X-irradiation. *J Invest Dermatol* **134**, 2251–2257 (2014).

1545 20. Griewank, K. G. *et al.* TERT Promoter Mutation Status as an Independent
1546 Prognostic Factor in Cutaneous Melanoma. *J Natl Cancer Inst* **106**, dju246 (2014).

1547 21. Spiegl-Kreinecker, S. *et al.* Prognostic quality of activating TERT promoter
1548 mutations in glioblastoma: interaction with the rs2853669 polymorphism and
1549 patient age at diagnosis. *Neuro Oncol* **17**, 1231–1240 (2015).

1550 22. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma.
1551 *Science* **339**, 959–961 (2013).

1552 23. Hosen, I. *et al.* Mutations in TERT promoter and FGFR3 and telomere length in
1553 bladder cancer. *International Journal of Cancer* **137**, 1621–1629 (2015).

1554 24. Labussière, M. *et al.* TERT promoter mutations in gliomas, genetic associations
1555 and clinico-pathological correlations. *Br J Cancer* **111**, 2024–2032 (2014).

1556 25. Hiyama, E. *et al.* Telomerase activity is detected in pancreatic cancer but not in
1557 benign tumors. *Cancer Res* **57**, 326–331 (1997).

1558 26. Bell, R. J. A. *et al.* Understanding TERT Promoter Mutations: a Common Path to
1559 Immortality. *Mol Cancer Res* **14**, 315–323 (2016).

1560 27. James, M. A., Vikis, H. G., Tate, E., Rymaszewski, A. L. & You, M.
1561 CRR9/CLPTM1L Regulates Cell Survival Signaling and Is Required for Ras
1562 Transformation and Lung Tumorigenesis. *Cancer Res* **74**, 1116–1127 (2014).

1563 28. Clarke, W. R., Amundadottir, L. & James, M. A. CLPTM1L/CRR9 ectodomain
1564 interaction with GRP78 at the cell surface signals for survival and chemoresistance
1565 upon ER stress in pancreatic adenocarcinoma cells. *International Journal of*
1566 *Cancer* **144**, 1367–1378 (2019).

1567 29. Jia, J. *et al.* CLPTM1L Promotes Growth and Enhances Aneuploidy in Pancreatic
1568 Cancer Cells. *Cancer Res* **74**, 2785–2795 (2014).

1569 30. Parashar, D. *et al.* Targeted biologic inhibition of both tumor cell-intrinsic and
1570 intercellular CLPTM1L/CRR9-mediated chemotherapeutic drug resistance. *npj*
1571 *Precis. Onc.* **5**, 16 (2021).

1572 31. James, M. A. *et al.* Functional Characterization of CLPTM1L as a Lung Cancer
1573 Risk Candidate Gene in the 5p15.33 Locus. *PLOS ONE* **7**, e36116 (2012).

1574 32. Derrien, A.-C. *et al.* Functional characterization of 5p15.33 risk locus in uveal
1575 melanoma reveals rs452384 as a functional variant and NKX2.4 as an allele-
1576 specific interactor. *Am J Hum Genet* **109**, 2196–2209 (2022).

1577 33. Chang, Y.-H. *et al.* Isoform-level analyses of 6 cancers uncover extensive genetic
1578 risk mechanisms undetected at the gene-level. *Br J Cancer* **133**, 874–885 (2025).

1579 34. Zhong, J. *et al.* A Transcriptome-Wide Association Study Identifies Novel
1580 Candidate Susceptibility Genes for Pancreatic Cancer. *JNCI: Journal of the*
1581 *National Cancer Institute* https://doi.org/10.1093/jnci/djz246 (2020)
1582 doi:10.1093/jnci/djz246.

1583 35. Yang, Y.-C. *et al.* rs401681 and rs402710 confer lung cancer susceptibility by
1584 regulating TERT expression instead of CLPTM1L in East Asian populations.
1585 *Carcinogenesis* **39**, 1216–1221 (2018).

1586 36. Wei, R. *et al.* TERT Polymorphism rs2736100-C Is Associated with EGFR
1587 Mutation-Positive Non-Small Cell Lung Cancer. *Clin Cancer Res* **21**, 5173–5180
1588 (2015).

1589 37. Helbig, S. *et al.* Functional dissection of breast cancer risk-associated TERT
1590 promoter variants. *Oncotarget* **8**, 67203–67217 (2017).

1591 38. Ge, M. *et al.* Functional evaluation of TERT-CLPTM1L genetic variants associated
1592 with susceptibility of papillary thyroid carcinoma. *Sci Rep* **6**, 26037 (2016).

1593 39. Zhou, L. *et al.* The identification of two regulatory ESCC susceptibility genetic
1594 variants in the TERT-CLPTM1L loci. *Oncotarget* **7**, 5495–5506 (2015).

1595 40. Hsiung, C. A. *et al.* The 5p15.33 Locus Is Associated with Risk of Lung
1596 Adenocarcinoma in Never-Smoking Females in Asia. *PLoS Genet* **6**, e1001051
1597 (2010).

1598 41. Florez-Vargas, O. *et al.* Genetic regulation of TERT splicing affects cancer risk by
1599 altering cellular longevity and replicative potential. *Nat Commun* **16**, 1676 (2025).

1600 42. Hofer, P. *et al.* MNS16A tandem repeats minisatellite of human telomerase gene: a
1601 risk factor for colorectal cancer. *Carcinogenesis* **32**, 866–871 (2011).

1602 43. Wang, L. *et al.* Association of a functional tandem repeats in the downstream of
1603 human telomerase gene and lung cancer. *Oncogene* **22**, 7123–7129 (2003).

1604 44. Yuan, H. *et al.* MNS16A polymorphism of the TERT gene on cancer risk: a
1605 systematic review and meta-analysis. *Nucleosides, Nucleotides & Nucleic Acids*
1606 **42**, 1004–1018 (2023).

1607 45. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary
1608 data with the “Sum of Single Effects” model. *PLOS Genetics* **18**, e1010299 (2022).

1609 46. Klein, A. P. *et al.* Genome-wide meta-analysis identifies five new susceptibility loci
1610 for pancreatic cancer. *Nat Commun* **9**, 556 (2018).

1611 47. Landi, M. T. *et al.* Genome-wide association meta-analyses combining multiple risk
1612 phenotypes provide insights into the genetic architecture of cutaneous melanoma
1613 susceptibility. *Nat Genet* **52**, 494–504 (2020).

1614 48. Koutros, S. *et al.* Genome-wide Association Study of Bladder Cancer Reveals New
1615 Biological and Translational Insights. *Eur Urol* **84**, 127–137 (2023).

1616 49. McKay, J. D. *et al.* Large-scale association analysis identifies new lung cancer
1617 susceptibility loci and heterogeneity in genetic susceptibility across histological
1618 subtypes. *Nat Genet* **49**, 1126–1132 (2017).

1619 50. Shi, J. *et al.* Genome-wide association study of lung adenocarcinoma in East Asia
1620 and comparison with a European population. *Nat Commun* **14**, 3043 (2023).

1621 51. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the
1622 expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-
1623 3440.e19 (2022).

1624 52. Wolpin, B. M. *et al.* Genome-wide association study identifies multiple susceptibility
1625 loci for pancreatic cancer. *Nat Genet* **46**, 994–1000 (2014).

1626 53. Zhong, J. *et al.* Large-scale multiomic analysis identifies non-coding somatic driver
1627 mutations and nominates ZFP36L2 as a driver gene for pancreatic ductal
1628 adenocarcinoma. *Gut* https://doi.org/10.1136/gutjnl-2025-335152 (2025)
1629 doi:10.1136/gutjnl-2025-335152.

1630 54. Long, E. *et al.* High-throughput characterization of functional variants highlights
1631 heterogeneity and polygenicity underlying lung cancer susceptibility. *Am J Hum*
1632 *Genet* **111**, 1405–1419 (2024).

1633 55. Choi, J. *et al.* Massively parallel reporter assays of melanoma risk variants identify
1634 MX2 as a gene promoting melanoma. *Nat Commun* **11**, 2718 (2020).

1635 56. Colli, L. M. *et al.* Altered regulation of DPF3, a member of the SWI/SNF complexes,
1636 underlies the 14q24 renal cancer susceptibility locus. *The American Journal of*
1637 *Human Genetics* **108**, 1590–1610 (2021).

1638 57. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using
1639 RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).

1640 58. Breeze, C. E. *et al.* FORGEdb: a tool for identifying candidate functional variants
1641 and uncovering target genes and mechanisms for complex diseases. *Genome Biol*
1642 **25**, 3 (2024).

1643 59. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor
1644 package for predicting variant effects at transcription factor binding sites.
1645 *Bioinformatics* **31**, 3847–3849 (2015).

1646 60. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16
1647 (2017).

1648 61. Li, W. *et al.* MAGeCK enables robust identification of essential genes from
1649 genome-scale CRISPR/Cas9 knockout screens. *Genome Biol* **15**, 554 (2014).

1650 62. Hsu, J. Y. *et al.* CRISPR-SURF: discovering regulatory elements by deconvolution
1651 of CRISPR tiling screen data. *Nat Methods* **15**, 992–993 (2018).

1652 63. Mukamel, R. E. *et al.* Repeat polymorphisms underlie top genetic risk loci for
1653 glaucoma and colorectal cancer. *Cell* **186**, 3659-3673.e23 (2023).

1654 64. Lee, O. W. *et al.* Targeted long-read sequencing of the Ewing sarcoma 6p25.1
1655 susceptibility locus identifies germline-somatic interactions with EWSR1-FLI1
1656 binding. *The American Journal of Human Genetics* **110**, 427–441 (2023).

1657 65. Grünewald, T. G. P. *et al.* Chimeric EWSR1-FLI1 regulates the Ewing sarcoma
1658 susceptibility gene EGR2 via a GGAA microsatellite. *Nat Genet* **47**, 1073–1078
1659 (2015).

1660 66. Amundadottir, L. *et al.* Genome-wide association study identifies variants in the
1661 ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet* **41**, 986–
1662 990 (2009).

1663 67. Petersen, G. M. *et al.* A genome-wide association study identifies pancreatic
1664 cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat*
1665 *Genet* **42**, 224–228 (2010).

1666 68. Zagouri, F. *et al.* HTERT MNS16A polymorphism in breast cancer: a case–control
1667 study. *Mol Biol Rep* **39**, 10859–10863 (2012).

1668 69. Thomas-Chollier, M. *et al.* Transcription factor binding predictions using TRAP for
1669 the analysis of ChIP-seq data and regulatory SNPs. *Nat Protoc* **6**, 1860–1869
1670 (2011).

1671 70. Yu, F.-X., Zhao, B. & Guan, K.-L. Hippo Pathway in Organ Size Control, Tissue
1672 Homeostasis, and Cancer. *Cell* **163**, 811–828 (2015).

1673 71. Landin Malt, A. *et al.* Alteration of TEAD1 Expression Levels Confers Apoptotic
1674 Resistance through the Transcriptional Up-Regulation of Livin. *PLoS ONE* **7**,
1675 e45498 (2012).

1676 72. Zhao, B., Li, L., Lei, Q. & Guan, K.-L. The Hippo–YAP pathway in organ size
1677 control and tumorigenesis: an updated version. *Genes Dev.* **24**, 862–874 (2010).

1678 73. Landin Malt, A., Georges, A., Silber, J., Zider, A. & Flagiello, D. Interaction with the
1679 Yes-associated protein (YAP) allows TEAD1 to positively regulate NAIP
1680 expression. *FEBS Letters* **587**, 3216–3223 (2013).

1681 74. Liu-Chittenden, Y. *et al.* Genetic and pharmacological disruption of the TEAD–YAP
1682 complex suppresses the oncogenic activity of YAP. *Genes Dev.* **26**, 1300–1305
1683 (2012).

1684 75. Szeto, S. G. *et al.* YAP/TAZ Are Mechanoregulators of TGF-β-Smad Signaling and
1685 Renal Fibrogenesis. *J Am Soc Nephrol* **27**, 3117–3128 (2016).

1686 76. Weischer, M. *et al.* Short Telomere Length, Cancer Survival, and Cancer Risk in
1687 47102 Individuals. *J Natl Cancer Inst* **105**, 459–468 (2013).

1688 77. Pooley, K. A. *et al.* A genome-wide association scan (GWAS) for mean telomere
1689 length within the COGS project: identified loci show little association with hormone-
1690 related cancer risk. *Hum Mol Genet* **22**, 5056–5064 (2013).

1691 78. Wentzensen, I. M., Mirabello, L., Pfeiffer, R. M. & Savage, S. A. The association of
1692 telomere length and cancer: a meta-analysis. *Cancer Epidemiol Biomarkers Prev*
1693 **20**, 1238–1250 (2011).

1694 79. Campa, D. *et al.* Leukocyte Telomere Length in Relation to Pancreatic Cancer
1695 Risk: A Prospective Study. *Cancer Epidemiol Biomarkers Prev* **23**, 2447–2454
1696 (2014).

1697 80. Lynch, S. M. *et al.* A prospective analysis of telomere length and pancreatic cancer
1698 in the alpha-tocopherol beta-carotene cancer (ATBC) prevention study.
1699 *International Journal of Cancer* **133**, 2672–2680 (2013).

1700 81. Nan, H. *et al.* Shorter Telomeres Associate with a Reduced Risk of Melanoma
1701 Development. *Cancer Res* **71**, 6758–6763 (2011).

1702 82. Iles, M. M. *et al.* The Effect on Melanoma Risk of Genes Previously Associated
1703 With Telomere Length. *J Natl Cancer Inst* **106**, dju267 (2014).

1704 83. Seow, W. J. *et al.* Telomere Length in White Blood Cell DNA and Lung Cancer: A
1705 Pooled Analysis of Three Prospective Cohorts. *Cancer Res* **74**, 4090–4098 (2014).

1706 84. Antwi, S. O. *et al.* Leukocyte Telomere Length and Its Interaction with Germline
1707 Variation in Telomere-Related Genes in Relation to Pancreatic Adenocarcinoma
1708 Risk. *Cancer Epidemiol Biomarkers Prev* **29**, 1492–1500 (2020).

1709 85. Pearson, J. D. *et al.* Binary pan-cancer classes with distinct vulnerabilities defined
1710 by pro- or anti-cancer YAP/TEAD activity. *Cancer Cell* **39**, 1115-1134.e12 (2021).

1711 86. Jakubosky, D. *et al.* Properties of structural variants and short tandem repeats
1712 associated with gene expression and complex traits. *Nat Commun* **11**, 2927
1713 (2020).

1714 87. Bhati, M., Mapel, X. M., Lloret-Villas, A. & Pausch, H. Structural variants and short
1715 tandem repeats impact gene expression and splicing in bovine testis tissue.
1716 *Genetics* **225**, iyad161 (2023).

1717 88. Jermusyk, A. *et al.* A 584 bp deletion in CTRB2 inhibits chymotrypsin B2 activity
1718 and secretion and confers risk of pancreatic cancer. *Am J Hum Genet* **108**, 1852–
1719 1865 (2021).

1720 89. Bakhtiari, M. *et al.* Variable number tandem repeats mediate the expression of
1721 proximal genes. *Nat Commun* **12**, 2075 (2021).

1722 90. Garg, P. *et al.* Pervasive *cis* effects of variation in copy number of large tandem
1723 repeats on local DNA methylation and gene expression. *The American Journal of*
1724 *Human Genetics* **108**, 809–824 (2021).

1725 91. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene
1726 expression variation in humans. *Nat Genet* **48**, 22–29 (2016).

1727 92. Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nat*
1728 *Commun* **4**, 2185 (2013).

1729 93. Coetzee, S. G. & Hazelett, D. J. motifbreakR v2: expanded variant analysis
1730 including indels and integrated evidence from transcription factor binding
1731 databases. *Bioinformatics Advances* **4**, vbae162 (2024).

1732 94. Murphy, A. E., Schilder, B. M. & Skene, N. G. MungeSumstats: a Bioconductor
1733 package for the standardization and quality control of many GWAS summary
1734 statistics. *Bioinformatics* **37**, 4593–4596 (2021).

1735 95. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and
1736 richer datasets. *Gigascience* **4**, s13742-015-0047–8 (2015).

1737 96. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring
1738 population-specific haplotype structure and linking correlated alleles of possible
1739 functional variants: Fig. 1. *Bioinformatics* **31**, 3555–3557 (2015).

1740 97. Sanson, K. R. *et al.* Optimized libraries for CRISPR-Cas9 genetic screens with
1741 multiple modalities. *Nat Commun* **9**, 5416 (2018).

1742 98. PacificBiosciences/pbAA. PacBio (2025).

1743 99. Dolzhenko, E. *et al.* Characterization and visualization of tandem repeats at
1744 genome scale. *Nat Biotechnol* **42**, 1606–1614 (2024).

1745 100. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural
1746 networks. *Nat Biotechnol* **36**, 983–987 (2018).

1747 101. Holt, J. M. *et al.* HiPhase: jointly phasing small, structural, and tandem repeat
1748 variants from HiFi sequencing. *Bioinformatics* **40**, btae042 (2024).

1749 102. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T.
1750 Accurate, scalable and integrative haplotype estimation. *Nat Commun* **10**, 5436
1751 (2019).

1752 103. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**,
1753 giab008 (2021).

1754 104. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from
1755 Next-Generation Reference Panels. *The American Journal of Human Genetics*
1756 **103**, 338–348 (2018).

1757 105. Zhang, Y. *et al.* Multisample motif discovery and visualization for tandem repeats.
1758 *Genome Res.* **35**, 850–862 (2025).

1759 106. Li, W. *et al.* Quality control, modeling, and visualization of CRISPR screens with
1760 MAGeCK-VISPR. *Genome Biology* **16**, 281 (2015).

1761