A genome-wide atlas of recurrent repeat expansions in human cancer

Graham S. Erwin,[1,*,†] Gamze Gürsoy,[2,3,*] Rashid Al-Abri,[1] Ashwini Suriyaprakash,[1] Egor Dolzhenko,[4] Kevin Zhu,[1] Christian R. Hoerner,[5] Shannon M. White,[1] Lucia Ramirez,[1] Ananya Vadlakonda,[1] Alekhya Vadlakonda,[1] Konor von Kraut,[1] Julia Park,[1] Charlotte M. Brannon,[1] Daniel A. Sumano,[1] Raushun A. Kirtikar,[1] Alicia A. Erwin,[6] Thomas J. Metzner,[5] Ryan K. C. Yuen,[7,8] Alice C. Fan,[5,9] John T. Leppert,[9,10,11,12] Michael A. Eberle,[4] Mark Gerstein,[13,14,15†] Michael P. Snyder[1,†]

1. Department of Genetics, Stanford University, Stanford, California, 94305, USA.

2. Department of Biomedical Informatics, Columbia University, New York, NY, 10032, USA.

3. New York Genome Center, New York, NY, 10013, USA.

4. Illumina, Inc, San Diego, California, 92122, USA.

5. Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, California, 94305, USA.

6. Data Science Program, Northwestern University, Chicago, Illinois, 60611, USA.

7. Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada. 12. Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.

8. Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California, 94305, USA.

9. Department of Urology, Stanford University School of Medicine, Stanford, California, 94305, USA.

10. Veterans Affairs Palo Alto Health Care System, Palo Alto, California, 94304, USA.

11. Division of Nephrology, Department of Medicine, Stanford University School of Medicine, Stanford, California, 94305, USA.

12. Computational Biology and Bioinformatics Program, Yale University, New Haven, Connecticut, 06511, USA.

14. Molecular Biophysics and Biochemistry Department, Yale University, New Haven, Connecticut, 06511, USA.

15. Department of Computer Science, Yale University, New Haven, Connecticut, 06511, USA.

*These authors contributed equally to this work.

†To whom correspondence should be addressed: mpsnyder@stanford.edu, gerwin@stanford.edu, mark@gersteinlab.org

1   **Abstract**

2   **Expansion of a single repetitive DNA sequence, termed a tandem repeat (TR), is**

3   **known to cause more than 50 diseases. However, repeat expansions are often not explored**

4   **beyond neurological and neurodegenerative disorders. In some cancers, mutations**

5   **accumulate in short tracts of TRs (STRs), a phenomenon termed microsatellite instability**

6   **(MSI); however larger repeat expansions have not been systematically analyzed in cancer.**

7   **Here, we identified TR expansions in 2,622 cancer genomes, spanning 29 cancer types. In 7**

8   **cancer types, we found 160 recurrent repeat expansions (rREs); most of these (155/160)**

9   **were subtype specific. We found that rREs were non-uniformly distributed in the genome**

10   **with an enrichment near candidate cis-regulatory elements, suggesting a role in gene**

11   **regulation. One rRE located near a regulatory element in the first intron of *UGT2B7* was**

12   **detected in 34% of renal cell carcinoma samples and was validated by long-read DNA**

13   **sequencing. Moreover, targeting cells harboring this rRE with a rationally designed,**

14   **sequence-specific DNA binder led to a dose-dependent decrease in cell proliferation.**

15   **Overall, our results demonstrate that rREs are an important but unexplored source of**

16   **genetic variation in human cancers, and we provide a comprehensive catalog for further**

17   **study.**

18

19   **Introduction**

20   Expansions of tandem DNA repeats (TRs) are known to cause more than 50 devastating

21   human diseases including Huntington's disease and Fragile X syndrome[1,2]. TR tracts that cause

22   human disease are typically large (more than 100 base pairs)[1]. However, identifying large TRs

23   with short-read DNA sequencing methods is difficult because they are ubiquitous in the genome,

24   and many are too large—larger than the typical sequencing read length—to uniquely map to the

25   reference genome[3]. Thus, many large TRs go undetected with current genomic technologies, and

26   despite their importance to monogenic disease, the frequency and function of recurrent repeat

27   expansions (rREs) are unknown in complex human genetic diseases, such as cancer[4].

28   Previous studies have profiled the landscape of alterations in STRs in cancer genomes[5–7].

29   In particular, microsatellite instability (MSI)[8–10], defined by an alteration in the lengths of short

30   TRs, is prevalent in various types of cancer including endometrial (30%), stomach (20%), and

31   colorectal cancers (15%)[5,6,11–12]. However, the systematic analysis of the frequency of genome-

32   wide large TR expansions has not been studied in cancer even though it was posited more than

33   25 years ago[13].

34   Recently, new bioinformatic tools to identify repeat expansions in short-read whole-

35   genome sequencing (WGS) datasets[14–17] have led to the identification of both known and novel

36   repeat expansions in human disease, primarily in the area of neurological disorders where repeat

37   expansions have historically been studied[14–22]. Here, we analyzed 2,622 human cancer genomes

38   with matching normal samples for the presence of somatic repeat expansions. We identified 160

39   recurrent repeat expansions (rREs) in seven types of cancer, including many rREs located in or

40   near known regulatory elements. One of these rREs is observed in 34% of kidney cancers, and

41   targeting this repeat expansion with sequence-specific DNA binders led to a dose-dependent

42   decrease in cellular proliferation. Overall, our approach reveals a new class of recurrent changes

43   in cancer genomes and provides an initial resource of these changes. Our results also suggest an

44   opportunity for a new class of oncologic TR-targeted therapeutics[23,24].

45

Results

Recurrent repeat expansions in cancer

48   We collected uniformly-processed alignments of whole-genome sequencing (WGS) of

49   tumor-normal pairs in the International Cancer Genome Consortium (ICGC), The Cancer

50   Genome Atlas (TCGA), both a part of the pan-cancer analysis of whole genomes (PCAWG)

51   datasets[25]. After filtering, these data consist of 2,622 cancer genomes from 2509 patients across

52   29 different cancer types (**Extended Data Figure 1**). Each cancer type was treated as its own

53   cohort and analyzed independent of other cancer types. We called somatic recurrent repeat

54   expansions (rREs) with ExpansionHunter Denovo (EHdn) (see **Methods**), which measures TRs

55   whose length exceeds the sequencing read length in short-read sequencing datasets[26,27]. That is,

56   EHdn performs case-control comparisons using a non-parametric statistical test to determine

57   whether repeat length is longer in tumor genomes compared to matching normal genomes. This

58   approach is analogous to joint population-level genotyping.

59   We first confirmed the accuracy of EHdn by performing whole-genome short- and long-

60   read sequencing on 786-O and Caki-1 cancer lines. We found that EHdn captured 72% of the

61   repeat expansions observed in long-read sequencing (**Extended Data Figure 2**). We also tested

62   the effect of sequencing coverage on the detection of rREs, and found that EHdn was robust

63   down to 30x coverage (**Extended Data Figure 2**). We then analyzed 2,622 matching tumor and

64   normal genomes with EHdn (285,363 TRs). We identified 578 candidate rREs (locus-level false

65   discovery rate (FDR) < 10%).

66   EHdn is expected to be sensitive to the copy number variations observed in cancer

67   genomes. We therefore devised and implemented a local read depth filtering method to account

68   for copy number variants, which normalizes the signal originating from repeat reads using the

69   read depth in the vicinity of the TR (see **Methods** and **Extended Data Figure 3**). We

70   benchmarked the local read depth normalization approach with simulated chromosomal

71   amplifications ranging from two (diploid) to 10 copies. We found that this filter accounts for

72   changes in chromosomal copy number in a manner superior to the standard global read depth

73   normalization (**Fig. S5**). Overall, we conclude that local read depth normalization is valuable to

74   identify *bona fide* rREs in cancer genomes and that many of the rREs that pass the filter are

75   expanded in cancer. For example, without local read depth normalization, we could only detect

76   31% of candidate rREs in independent cohorts of matching tumor-normal tissue samples for

77   breast, prostate, and kidney cancer (15, 18, and 12 patients, respectively). Our local read depth

78   filtering approach removed >75% (418/578) false-positive candidate rREs (**Extended Data**

79   **Figure 3**). Importantly, several rRE candidates that were removed are situated in hotspots for

80   chromosomal amplification, such as chromosomal 8q amplifications that

81   increase *MYC* production in breast cancer (**Extended Data Figure 3**)[28]. Our analysis suggests

82   that the standalone EHdn method may have selected these loci due to amplification rather than

83   repeat expansions, and thus their removal is important.

84   After implementing our local read depth filtering strategy, we increased our detection rate

85   to 57% (8/14) in independent cohorts (**Extended Data Figure 3**). Importantly, the loci we could

86   not validate had lower expansion frequencies (5–12%). These rREs may be real but more

87   difficult to validate in the small validation cohorts (**Table S7**). Thus, we believe this number is

88   likely an underestimate of the independent detection rate. Of the 14 candidate rREs that failed

89   our local read depth filter, 29% (4/14) were detected in independent cohorts of samples

90   indicating that the filtering removes most loci that cannot be validated (**Extended Data Figure**

91   **3**), but also removes some true positives as well.

92   After accounting for local read depth, we detected 160 rREs in 7 human cancers (**Fig.**

93   **1b**). We expected high concordance with ExpansionHunter given that this tool is related to

94   EHdn, and indeed we observed a 91% confirmation rate with ExpansionHunter (**Extended Data**

95   **Figure 4**). We found that most (80%) of these loci are rarely expanded in the general population

96   (<5% of the time, n = 6,514 genomes, **Extended Data Figure 2**). rREs were primarily observed

97   in prostate and liver cancer, but we also detected rREs in ovarian, pilocytic astrocytoma, renal

98   cell carcinoma (RCC), chromophobe RCC, and squamous cell lung carcinoma. Thus, rREs are

99   found in tissues derived from each of the three primary germ layers (ectoderm, mesoderm, and

100   endoderm), suggesting these expansions are a phenomenon inherent to the human genome rather

101   than any tissue-specific process. In prostate and liver cancer, most cancer genomes (93% and

102   95%, respectively) contain at least one rRE, with some genomes harboring several rREs (**Fig.**

103   **1c**). For some pathogenic repeats, a larger TR length at birth predisposes an individual to somatic

104   repeat expansions later in life[1,2], but we did not generally observe that with rREs (**Table S8**).

105   Overall, rREs are found in 7 of 29 human cancers examined and are largely cancer subtype-

106   specific.

107   We next examined whether rREs correlate with changes in MSI[5,6]. We determined

108   whether samples harboring an rRE had a higher mutation rate in STRs, which is a hallmark of

109   MSI[5,29]. We did not observe any significant difference in STR mutation rate for genomes with an

110   rRE compared to those lacking an rRE (two-tailed Wilcoxon rank sum test, P = 0.27, **Fig. 1d**).

111   We also compared cancer genomes harboring rREs with cancer genomes previously identified as

112   MSI, using recent results from the PCAWG consortium[29]. We did not observe any enrichment in

113   MSI for samples harboring an rRE, and instead found a weak but significant preference for rREs

114   in microsatellite stable (MSS) samples, not MSI samples (Two-tailed Wilcoxon rank-sum test, *P*

115   = 0.04, **Fig. 1e**, see also **Extended Data Figure 5**). Thus, our findings might suggest a model

116   where rREs are formed by a process that is distinct from MSI.

117   In addition to MSI, different mutational processes lead to a signature of somatic

118   mutations. We tested whether rREs are associated with known mutational signatures by

119   comparing them to 49 signatures of single base substitutions (SBS) and 11 doublet base

120   substitutions (DBS)[30]. We performed a multiple linear regression to predict the number of rREs

121   in a sample based on SBS and DBS signatures, respectively. Only one DBS signature, DBS2,

122   showed a very weak association with rREs (r^2 = 0.12) (**Extended Data Figure 5**).

123

124   **rREs overlap regulatory elements**

125   Among the 160 rREs, we observed a variety of different motifs (**Table S1**) whose repeat

126   unit length follows a bimodal distribution, consistent with REs identified in other diseases (**Fig.**

127   **2a**, **Extended Data Figures 6 and 7**)[27]. rREs are distributed across a range of GC content;

128   approximately half (76/160) have GC content less than 50% (**Table S1**). Six rREs contained a

129   known pathogenic motif, all of which were GAA[31]. We examined if any motifs were enriched in

130   the rRE catalog compared to the tandem repeat finder (TRF) catalog. Although this enrichment

131   could arise from a biological and/or technical process, we found that one of the three enriched

132   motifs was GAA (**Fig. 2b**). As an example, Friedreich's ataxia is caused by a repeat expansion of

133   a GAA motif in the intron of the frataxin gene. This expansion leads to DNA methylation and the

134   deposition of repressive chromatin marks, leading to robust repression of the gene and

135   development of disease[31]. Because of this, we suspect some of the rREs found in cancer might

136   alter the epigenome and affect gene regulatory networks.

137   rREs were distributed non-uniformly across the genome, with a bias towards the ends of

138   chromosome arms (**Fig. 2c**, **Extended Data Figure 6**). This observation is consistent with

139   previous reports of TRs and structural variants[15,32]. We also examined the distribution of rREs

140   relative to gene features with annotatr (**Fig. 2d**)[33]. The 7% of rREs labeled as exonic appeared

141   proximal to, but not within, exons, but others were in introns, untranslated regions (UTRs), and

142   splice sites. These results suggest rREs may play different functional roles in the regulation of

143   gene expression.

144   We measured the distance between rREs and candidate cis-regulatory elements

145   (cCREs)[34]; cCREs comprise approximately one million functional elements including promoters,

146   enhancers, DNase-accessible regions, and insulators bound by CCCTC-binding factor (CTCF).

147   An rRE near a regulatory element could alter the function of that regulatory element, as is

148   observed in Fragile X syndrome and Friedreich's ataxia[1]. Interestingly, rREs are located closer to

149   cCREs than expected by chance, and we find that 47 of 160 rREs directly overlap with a known

150   cCRE (Welch's t-test, $P = 6.00e\text{-}45$, **Fig. 2e** & **Extended Data Figure 7**). Thus, rREs are often

151   found in or near functional regions of the genome.

152

153   **rREs with links to cancer**

154   We mapped each rRE to the nearest genes and found that nine rREs map to Tier 1 genes

155   present in the census of somatic mutations in cancer (COSMIC) database (**Fig. 3**, **Table S1**). We

156   also observed a strong correlation with cancer-related genes (Jensen disease-gene associations[35]).

157   That is, four of the top five diseases associated with the collection of 160 rRE are cancers (**Fig.**

158   **3b**, **Table S4**).

159   To examine whether some rREs play a role in oncogenesis, we looked at their association

160   with previously-identified cancer risk loci. Many rREs were identified in prostate cancer, and 63

161   loci have previously been associated with susceptibility to prostate cancer from available

162   genome-wide association studies[36]. When we examined the co-localization of rREs and cancer

163   risk loci in prostate cancer, we found that rREs are located closer to prostate cancer susceptibility

164   loci than standard STRs or by chance (Student's *t*-test, FDR q = 0.08, **Fig. 3c** & **Extended Data**

165   **Figure 7**).

166   We next studied the relationship between the occurrence of the census of somatic

167   mutations in cancer (COSMIC) genes to the occurrence of rREs (**Fig. 3d**). Interestingly, we

168   found that there are five COSMIC genes whose somatic mutations are found to occur

169   significantly more in patients' genomes with no rREs, after correcting for multiple hypothesis

170   testing. Among them, *TP53* was particularly striking, as wildtype *TP53* is critical for mediating

171   the pathogenic effects of repeat expansions in both Amyotrophic Lateral Sclerosis (ALS) and

172   Huntington's Disease[37,38]. Consistent with these findings, a DNA damage repair gene in yeast,

173   *Rad53*, is phosphorylated and activated in the presence of an expanded repeat[39].

174   MSI-high cancers are often correlated with higher levels of immune cell infiltration[40]. We

175   hypothesized that some rREs might also be associated with higher immune cell infiltration, but

176   we did not observe a correlation between cytotoxic activity[41] and the presence of an rRE

177   (**Extended Data Figure 9**). Because there are matching RNA-seq data for only 4 of 160 rREs,

178   this analysis warrants further investigation as more matching WGS and RNA-seq datasets

179   become available.

180

181   **An rRE in the *UGT2B7* gene observed in RCC**

182   A GAAA expansion located in the intron of *UGT2B7* was observed in 34% of RCC

183   samples. *UGT2B7* is a glucuronidase that clears small molecules—including

184   chemotherapeutics—from the body and is selectively expressed in the kidney and liver[42].

185   With gel electrophoresis, we identified the expected TR size of ~26 GAAA repeats in the

186   normal kidney cell line, HK-2, corresponding closely to the length observed in the reference

187   genome (**Fig. 4a**). In contrast, we identified an expansion between ~63 and ~160 GAAA repeat

188   units in 5 of 8 clear cell RCC cell lines. Most expansions were heterozygous (**Fig. 4a**). Long-

189   read DNA sequencing with highly-accurate PacBio HiFi reads confirmed the PCR results and

190   revealed the precise structure of this repeat expansion at single base pair resolution for both 786-

191   O and Caki-1 (**Fig. 4b**). We also detected this repeat expansion in five out of 12 primary kidney

192   tumor tissue samples from patients with clear cell RCC (**Extended Data Figure 8**), which

193   showed more heterogeneity than the RCC cell lines; more heterogeneity might be expected for

194   human tumor samples compared to the clonal cell lines.

195   Given that *UGT2B7* is selectively expressed in the liver and kidney, and that it plays a

196   role in clearing small molecules from the body, we examined whether this rRE may be located

197   near any functional elements that could regulate its expression. Analysis of the chromatin

198   environment surrounding the rRE in *UGT2B7* revealed a nearby enhancer, raising the possibility

199   that this rRE alters the expression of *UGT2B7* (**Fig. 4c**). The repeat motif of this rRE, GAAA,

200   appears similar to the pathogenic repeat motif found in Friedreich's ataxia, which is GAA. The

201   pathogenic GAA repeat expansion blocks *FXN* expression[31]. We therefore hypothesized that the

202   intronic GAAA repeat expansion might repress the expression of *UGT2B;* we found a modest

203   decrease in expression that was not statistically significant (**Extended Data Figure 8**). While

204   this rRE is also not associated with a difference in survival (**Extended Data Figure 8**), it is

205   associated with a significant decrease in a transcript isoform in *UGT2B7* (Wald test with FDR

206   correction, $P = 0.0048$) (**Fig. 4e**). Interestingly, a shift in isoform usage of *UGT2B7* has been

207   noted in cancer[43].

208

209   **Repeat-targeting anti-proliferative agents**

210   Do GAAA repeat expansions contribute to cell proliferation? We previously showed that

211   targeting a related TR motif, GAA, with synthetic transcription elongation factors (Syn-TEF1)

212   reverses pathogenesis in several models of Friedreich's ataxia[23]. Therefore, if the GAAA rRE in

213   RCC behaves similarly, then a Syn-TEF targeting GAAA may have anti-proliferative activity.

214   We rationally designed Syn-TEF3, which contains a GAAA-targeting polyamide (PA), and a

215   bromodomain ligand, JQ1, designed to recruit part of the transcriptional machinery (**Fig. 5a** and

216   **Fig. S2**). We also included a control molecule, Syn-TEF4, which targets G**G**AA TRs, as well as

217   polyamides (PAs) PA3 and PA4 that lack the JQ1 domain. We have previously shown that Syn-

218   TEFs and PAs localize to repetitive TRs in living cells[23,44].

219   The effect of Syn-TEFs on cell proliferation was examined (**Fig. 5b**). Caki-1 and 786-O were

220   selected because they have the largest (161) and smallest (24) GAAA tracts within the first

221   intron of *UGT2B7*, respectively. In a dose-dependent manner, we observed that Syn-TEF3 led to

222   a significant decrease in the proliferation of Caki-1 cells but had little effect on 786-O cells. Syn-

223   TEF4, which does not target a GAAA TR, did not significantly decrease proliferation in either of

224   the cell lines tested, demonstrating the requirement for GAAA-specific targeting (**Fig. 5b**). Two

225   additional GAAA repeat expansion cell lines as well as two additional control non-expanded

226   lines showed a similar association between Syn-TEF sensitivity and the presence of the repeat

227   expansion (**Extended Data Figure 10**). Consistent with this finding, Caki-1 cells treated with

228   Syn-TEF3 exhibited a significant increase in cell death compared to DMSO control, as measured

229   by propidium iodide staining (**Fig. 5c,d** and **Extended Data Figure 10**). In contrast, 786-O cells

230   treated with Syn-TEF3 showed no significant difference in propidium iodide-positive cells

231   compared to DMSO (**Fig. 5c,d** and **Extended Data Figure 10**). Importantly, the Syn-TEF4,

232   PA3, and PA4 control agents exhibited no significant effect on cell death in either cell line

233   compared to vehicle control (**Fig. 5c,d** and **Extended Data Figure 10**). These results suggest

234   that GAAA repeat expansions may represent a genetic vulnerability in RCC and provide a proof-

235   of-principle study for the functional role of rREs in cancer.

236

Discussion

238   Here, for the first time, we conduct a genome-wide survey of recurrent repeat expansions

239   (rREs) across cancer genomes, distinct from MSI. Our data identified (i) 160 rREs in 7 human

240   cancers and revealed that (ii) most (155 of 160) rREs are cancer subtype-specific; (iii) amongst

241   diseases, rREs are enriched in human cancer loci and tended to occur near regulatory elements;

242   (iv) recurrent repeat expansions do not correlate with MSI status; and (v) targeting a GAAA

243   repeat expansion in RCC with a small molecule leads to cancer cell killing. Taken together, our

244   results uncover an unexplored genetic alteration in cancer genomes with important mechanistic

245   and therapeutic implications.

246   Cancer cells evolve and adapt in response to environmental or pharmacological

247   perturbations, but the mechanisms supporting these changes are still being uncovered. One

248   source of genetic variation that may enable genetic adaptations is TR DNA sequences. Mutations

249   in repeat length of TRs can occur up to 10,000 times more frequently than single nucleotide

250   variants (SNVs) or insertions and deletions (INDELs)[1]. Repeat expansions may provide a source

251   of genetic variation to enable cancer cells to adapt to changes in the environment[45]. Indeed,

252   colorectal cancers acquire mutations in STRs in response to targeted therapy just 24 hours

253   following treatment, suggesting that mutations in these regions may associate with rapid

254   evolution[46]. In future studies, it will be particularly valuable to study repeat expansions in the

255   genomes of cancer cells that face changing environments, including metastasis and

256   chemotherapy.

257   Historically, MSI has been the focus of efforts to profile changes in STRs in cancer

258   genomes because specific cancer-causing genetic alterations in repair genes can promote

259   widespread STR alterations. Interestingly, we find little to no correlation between rREs and MSI.

260   These results are consistent with previous findings in which the correlation between MSI and

261   repeat instability at larger TRs is not definitive[47]. MSI may contribute to a subtype of rREs that

262   we have not yet uncovered, or rREs may arise from a mutation process that is distinct from that

263   of MSI. There are several different DNA cellular repair systems, and presumably the rREs that

264   we observed are due to very specific loci-associated mechanisms or activities. Some of these

265   repeat expansions may be due to cis-regions with interesting DNA or chromatin configurations

266   that are prone to expansion at distinct loci, rather than gene mutations that cause global trans

267   effects, as occurs in MSI.

268   There are numerous mechanisms by which a repeat expansion can alter cellular function.

269   Known pathogenic repeat expansions can alter the coding sequence of a protein, such as in the

270   case of Huntington's disease. However, there are several repeat expansions that occur in non-

271   coding regions that alter gene expression[1]. In other instances, the repeat expansion can lead to a

272   pathogenic RNA molecule (myotonic dystrophy) or protein (ALS)[1]. Finally, repeat expansions in

273   MSI-associated cancers, which are too small to detect by EHdn, can disrupt DNA replication[48].

274   Thus, our catalog represents a powerful resource to explore the mechanisms by which rREs alter

275   cellular function in cancer.

276   The identification of repeat expansions would benefit from improved sequencing

277   coverage and increased cohort sizes. Like other tools that identify repeat expansions, we cannot

278   distinguish zygosity from sample heterogeneity or obtain precise lengths of repeats. Our

279   independent experimental validation showed that some repeat expansions are heterogeneous

280   (**Extended Data Figure 8**). We suspect that tumor heterogeneity may lead to an underreporting

281   of rREs. Furthermore, this study focuses on somatic mutations, but repeat expansions that occur

282   in the context of normal development will be another important area of study[4]. Furthermore,

283   germline events that predispose an individual to cancer would also be worthwhile to study; there

284   is evidence that a TR in the androgen receptor gene is associated with prostate cancer onset,

285   tumor stage, and tumor grade[49]. Finally, we only detected changes in repeat length that were

286   greater than sequencing read length. In future studies, it will be important to explore recurrent

287   changes that are smaller in length. Finally, it is important to acknowledge that rREs could be

288   mediators of phenotypes or passengers that result from genetic instability and clonal selection. In

289   the one instance where we targeted the rRE in RCC, cell proliferation was reduced, consistent

290   with a mediator role for this rRE. Distinguishing between these two possibilities for each rRE is

291   an important line of work in the future.

292   To our knowledge, this is the first genome-wide survey of repeat expansions beyond a

293   neurological or neurodegenerative disorder. Thousands of high-quality whole genome sequences

294   exist for many diseases, and our data provide evidence that repeat expansions should be explored

295   beyond the classical bounds of neurodegenerative diseases where they have been most

296   investigated.

297   A careful dissection of repeat expansions in human disease may reveal their role as

298   causative or contributory. We show here that repeat expansions can be targeted by tandem

299   repeat-targeting precision molecules[23]. Thus, our results set the stage for a new class of

300   therapeutics to be deployed in cancer and other diseases.

301

302   **References**

303   1.   Hannan, A. J. Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev.*

304   *Genet.* **19**, 286–298 (2018).

305   2.   Gall-Duncan, T., Sato, N., Yuen, R. K. C. & Pearson, C. E. Advancing genomic

306   technologies and clinical awareness accelerates discovery of disease-associated tandem

307   repeat sequences. *Genome Res.* **32**, 1–27 (2022).

308   3.   Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev.*

309   *Genet.* (2019) doi:10.1038/s41576-019-0180-9.

310   4.   Hannan, A. J. Tandem repeat polymorphisms: modulators of disease susceptibility and

311   candidates for 'missing heritability'. *Trends Genet.* **26**, 59–65 (2010).

312   5.   Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and

313   characterization of microsatellite instability across 18 cancer types. *Nat. Med.* **22**, 1342–

314   1350 (2016).

315   6.   Cortes-Ciriano, I., Lee, S., Park, W. Y., Kim, T. M. & Park, P. J. A molecular portrait of

316   microsatellite instability across multiple cancers. *Nat. Commun.* **8**, 1–12 (2017).

317   7.   Grünewald, T. G. P. *et al.* Chimeric EWSR1-FLI1 regulates the Ewing sarcoma

318   susceptibility gene EGR2 via a GGAA microsatellite. *Nat. Genet.* **47**, 1073–1078 (2015).

319   8.   Aaltonen, L. A. *et al.* Clues to the pathogenesis of familial colorectal cancer. *Science (80-.*

320   *).* **260**, 812 LP – 816 (1993).

321   9.   Thibodeau, S. N., Bren, G. & Schaid, D. Microsatellite instability in cancer of the

322   proximal colon. *Science (80-.).* **260**, 816 LP – 819 (1993).

323   10.   Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D. & Perucho, M. Ubiquitous

324   somatic mutations in simple repeated sequences reveal a new mechanism for colonic

325   carcinogenesis. *Nature* **363**, 558–561 (1993).

326   11.   Wooster, R. *et al.* Instability of short tandem repeats (microsatellites) in human cancers.

327   *Nat. Genet.* **6**, 152–156 (1994).

328   12.   Risinger, J. I. *et al.* Genetic Instability of Microsatellites in Endometrial Carcinoma.

329   *Cancer Res.* **53**, 5100 LP – 5103 (1993).

330   13.   Panzer, S., Kuhl, D. P. A. & Caskey, C. T. Unstable triplet repeat sequences: A source of

331   cancer mutations? *Stem Cells* **13**, 146–157 (1995).

332   14.   Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome

333   sequence data. *Genome Res.* **27**, 1895–1903 (2017).

334   15.   Dolzhenko, E. *et al.* ExpansionHunter Denovo: a computational method for locating

335   known and novel repeat expansions in short-read sequencing data. *Genome Biol.* **21**, 102

336   (2020).

337   16.   Dashnow, H. *et al.* STRetch: detecting and discovering pathogenic short tandem repeat

338   expansions. *Genome Biol.* **19**, 121 (2018).

339   17.   Mousavi, N., Shleizer-Burko, S., Yanicky, R. & Gymrek, M. Profiling the genome-wide

340   landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 (2019).

341   18.   Kristmundsdottir, S., Eggertsson, H. P., Arnadottir, G. A. & Halldorsson, B. V. popSTR2

342   enables clinical and population-scale genotyping of microsatellites. *Bioinformatics* (2019)

343   doi:10.1093/bioinformatics/btz913.

344   19.   Tang, H. *et al.* Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human

345   Whole Genomes. *Am. J. Hum. Genet.* **101**, 700–715 (2017).

346   20.   Rafehi, H. *et al.* Bioinformatics-Based Identification of Expanded Repeats: A Non-

347   reference Intronic Pentamer Expansion in RFC1 Causes CANVAS. *Am. J. Hum. Genet.*

348   **105**, 151–165 (2019).

349   21.   Hannan, A. J. Repeat DNA expands our understanding of autism spectrum disorder.

350   *Nature* **589**, 200–202 (2021).

351   22.   Mitra, I. *et al.* Patterns of de novo tandem repeat mutations and their role in autism.

352   *Nature* **589**, 246–250 (2021).

353   23.   Erwin, G. S. *et al.* Synthetic transcription elongation factors license transcription across

354   repressive chromatin. *Science (80-.).* **358**, 1617–1622 (2017).

355   24.   Nakamori, M. *et al.* A slipped-CAG DNA-binding small molecule induces trinucleotide-

356   repeat contractions in vivo. *Nat. Genet.* **52**, (2020).

357   25.   The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer

358   analysis of whole genomes. *Nature* **578**, 82–93 (2020).

359   26.   Tankard, R. M. *et al.* Detecting Expansions of Tandem Repeats in Cohorts Sequenced

360   with Short-Read Sequencing Data. *Am. J. Hum. Genet.* **103**, 858–873 (2018).

361   27.   Trost, B. *et al.* Genome-wide detection of tandem DNA repeats that are expanded in

362   autism. *Nature* **589**, 80–86 (2020).

363   28.   Tirkkonen, M. *et al.* Molecular cytogenetics of primary breast cancer by CGH. *Genes,*

364   *Chromosom. Cancer* **21**, 177–184 (1998).

365   29.   Fujimoto, A. *et al.* Comprehensive analysis of indels in whole-genome microsatellite

366   regions and microsatellite instability across 21 cancer types. *Genome Res.* **30**, 334–346

367   (2020).

368   30.   Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature*

369   **578**, 94–101 (2020).

370   31.   Sandi, C., Al-Mahdawi, S. & Pook, M. A. Epigenetics in Friedreich's Ataxia: Challenges

371   and Opportunities for Therapy. *Genet. Res. Int.* **2013**, 852080 (2013).

372   32.   Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human

373   Genome. *Cell* **176**, 663-675.e19 (2019).

374   33.   Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics*

375   **33**, 2381–2383 (2017).

376   34.   Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse

377   genomes. *Nature* **583**, 699–710 (2020).

378   35.   Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X. & Jensen, L. J. DISEASES:

379   Text mining and data integration of disease–gene associations. *Methods* **74**, 83–89 (2015).

380   36.   Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new

381   prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).

382   37.   Maor-Nof, M. *et al.* p53 is a central regulator driving neurodegeneration caused by

383   C9orf72 poly(PR). *Cell* **184**, 689-708.e20 (2021).

384   38.   Bae, B.-I. *et al.* p53 Mediates Cellular Dysfunction and Behavioral Abnormalities in

385   Huntington's Disease. *Neuron* **47**, 29–41 (2005).

386 39. Sundararajan, R. & Freudenreich, C. H. Expanded CAG/CTG Repeat DNA Induces a
387   Checkpoint Response That Impacts Cell Proliferation in Saccharomyces cerevisiae. *PLOS*
388   *Genet.* **7**, e1001339 (2011).

389 40. Lin, A., Zhang, J. & Luo, P. Crosstalk Between the MSI Status and Tumor
390   Microenvironment in Colorectal Cancer   . *Frontiers in Immunology*   vol. 11 2039 (2020).

391 41. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and Genetic
392   Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell* **160**, 48–61
393   (2015).

394 42. Barre, L. *et al.* Substrate specificity of the human UDP-glucuronosyltransferase UGT2B4
395   and UGT2B7. *FEBS J.* **274**, 1256–1264 (2007).

396 43. Rouleau, M. *et al.* Divergent Expression and Metabolic Functions of Human
397   Glucuronosyltransferases through Alternative Splicing. *Cell Rep.* **17**, 114–124 (2016).

398 44. Erwin, G. S. G. S. *et al.* Synthetic genome readers target clustered binding sites across
399   diverse chromatin states. *Proc. Natl. Acad. Sci.* **113**, E7418–E7427 (2016).

400 45. Kim, J. C. & Mirkin, S. M. The balancing act of DNA repeat expansions. *Curr. Opin.*
401   *Genet. Dev.* **23**, 280–288 (2013).

402 46. Russo, M. *et al.* Adaptive mutability of colorectal cancers in response to targeted
403   therapies. *Science (80-.).* **366**, 1473 LP – 1480 (2019).

404 47. Persi, E. *et al.* Proteomic and genomic signatures of repeat instability in cancer and
405   adjacent normal tissues. *Proc. Natl. Acad. Sci.* **116**, 16987 LP – 16996 (2019).

406 48. van Wietmarschen, N. *et al.* Repeat expansions confer WRN dependence in microsatellite-

407   unstable cancers. *Nature* **586**, 292–298 (2020).

408   49.   Edward, G. *et al.* The CAG repeat within the androgen receptor gene and its relationship

409   to prostate cancer. *Proc. Natl. Acad. Sci.* **94**, 3320–3323 (1997).

410

411 **Figure 1. Genome-wide detection of recurrent repeat expansions (rREs) in cancer genomes.**

412 a) Scheme of method to identify rREs in 2,509 patients across 29 human cancers. 1, head and

413 neck squamous cell carcinoma (Head−SCC); 2, Skin−Melanoma; 3, glioblastoma (CNS−GBM);

414 4 medulloblastoma (CNS−Medullo); 5, pilocytic astrocytoma (CNS−PiloAstro); 6, esophageal

415 adenocarcinoma (Eso−AdenoCA), 7, osteosarcoma (Bone−Osteosarc); 8, leiomyosarcoma

416 (Bone−Leiomyo); 9, thyroid adenocarcinoma (Thy−AdenoCA); 10, lung adenocarcinoma

417 (Lung−AdenoCA); 11, lung squamous cell carcinoma (Lung−SCC); 12, mammary gland

418 adenocarcinoma (Breast−AdenoCA); 13, B-cell non-Hodgkin lymphoma (Lymph−BNHL); 14,

419 chronic lymphocytic leukemia (Lymph−CLL); 15, acute myeloid leukemia (Myeloid−AML); 16,

420 myeloproliferative neoplasm (Myeloid−MPN); 17, biliary adenocarcinoma (Biliary-AdenoCA);

421 18, hepatocellular carcinoma (Liver−HCC); 19, stomach adenocarcinoma (Stomach−AdenoCA);

422 20, pancreatic adenocarcinoma (Panc−AdenoCA), 21, pancreatic neuroendocrine tumor

423 (Panc−Endocrine); 22, colorectal adenocarcinoma (ColoRect−AdenoCA); 23, prostatic

424 adenocarcinoma (Prost−AdenoCA); 24, chromophobe renal cell carcinoma (Kidney−ChRCC);

425 25, renal cell carcinoma (Kidney−RCC); 26, papillary renal cell carcinoma (Kidney−pRCC); 27,

426 uterine adenocarcinoma (Uterus−AdenoCA); 28, ovarian adenocarcinoma (Ovary−AdenoCA);

427 29, transitional cell carcinoma of the bladder (Bladder−TCC). b) Distribution of rREs across

428 cancer types. c) Proportion of cancer genomes with rREs. d) STR mutation rate for cancer

429 genomes with and without a rRE. Two-tailed Wilcoxon rank sum test. e) Distribution of rREs

430 across microsatellite stable (MSS) and microsatellite instability high (MSI-high) cancers. Chi-

431 square test with Yates' correction.
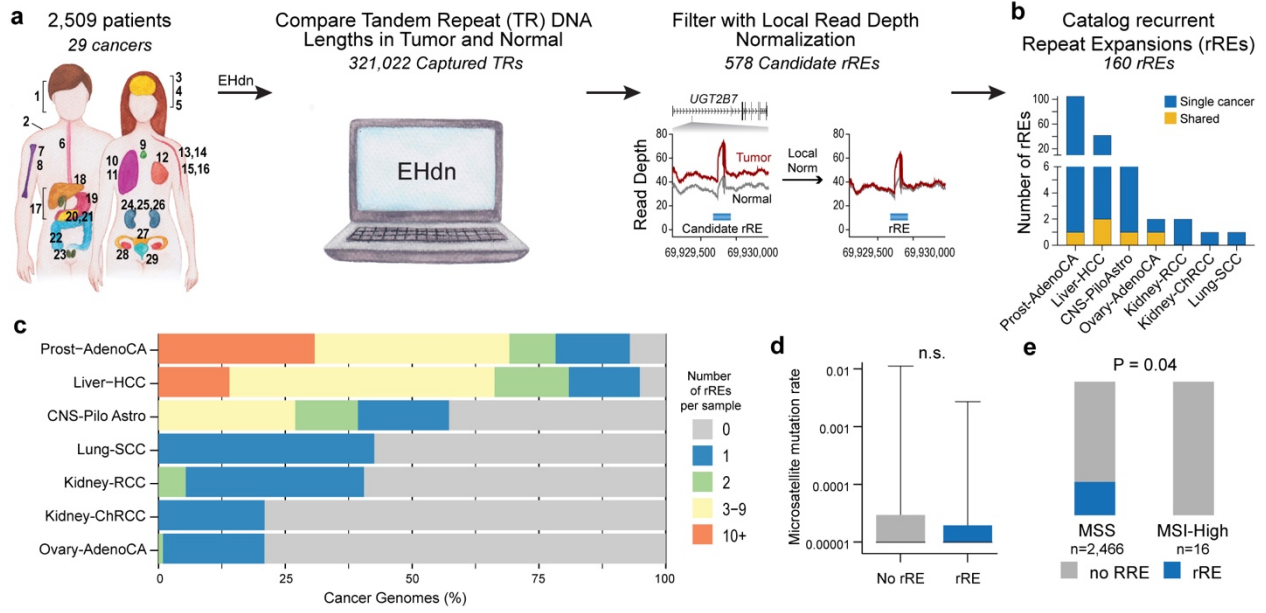
432

433

434

435

436   **Figure 2. Features of rREs.** a) Distribution of the repeat unit (motif) for rREs. b) Motifs

437   enriched in the catalog of rREs. c) Distance of rREs to the end of the chromosome arm. d)

438   Proportion of genic features that overlap with rREs. UTR, untranslated region. e) Distance of

439   simple repeats and rREs to the nearest ENCODE candidate cis-regulatory element (cCRE).

440   Center values represent the median. Welch's *t*-test.
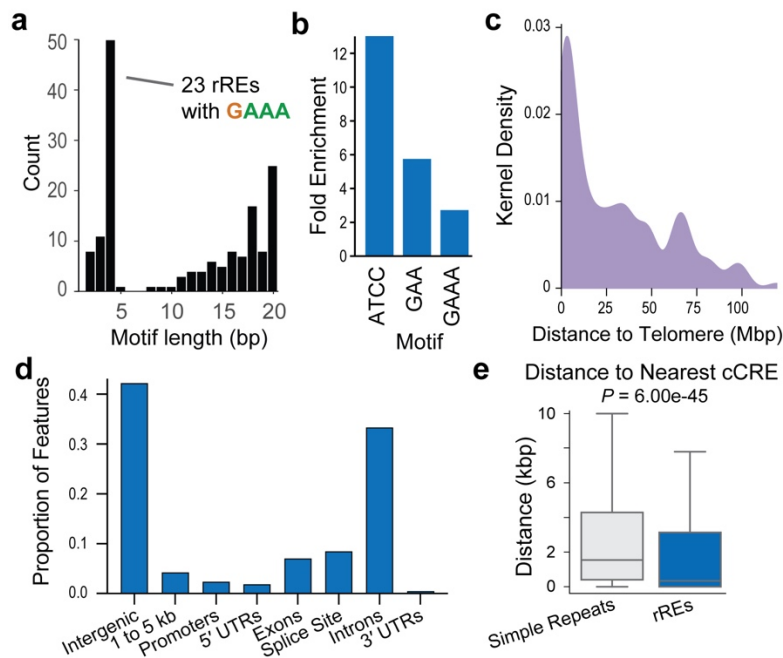
441

442

443

444

445 **Figure 3. Association of rREs with cancer features.** a) Association of rREs with human

446 diseases. b) Frequency of rREs in genes of interest, including the nine COSMIC genes, are

447 highlighted. c) Distance of simple repeats, non-prostate cancer rREs, and prostate-cancer rREs to

448 the nearest prostate cancer risk locus. Center values represent the median. Statistical significance

449 was measured with Welch's *t*-test (* q < 0.10). d) Association between SNVs in genes in the

450 census of somatic mutations in cancer (COSMIC) Tier 1 genes and the presence of rREs.

451 Student's *t*-test with FDR correction by Benjamini-Hochberg.
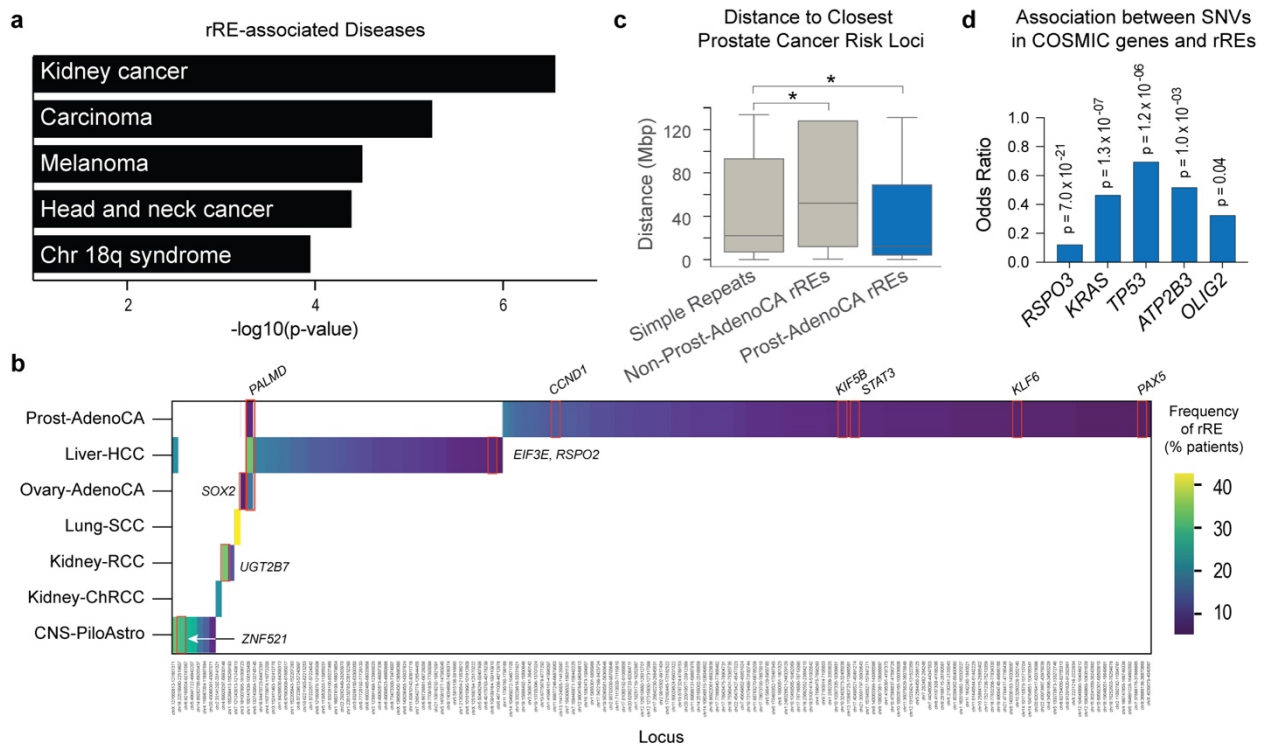

452

453

454

455   **Figure 4. An rRE in Renal Cell Carcinoma (RCC).** a) Gel electrophoresis of the GAAA

456   tandem repeat in RCC samples. This analysis was performed in duplicate and the gel is

457   representative of the results. For gel source data, see **Fig. S1**. b) Visualization of long-read

458   sequencing of GAAA rRE in the intron of *UGT2B7*. Data are from PacBio HiFi sequencing. c)

459   The locus surrounding the rRE detected in the intron of *UGT2B7*. Signal traces of Pol2,

460   H3K27ac, H3K4me1, and p300 in HepG2 cells are shown. Candidate cis-regulatory elements

461   (cCREs) and chromatin states (ChromHMM) are also depicted. d) Expression of *UGT2B7*

462   isoform ENST00000508661.1 in RCC samples as a function of the detection of the rRE in

463   *UGT2B7* (Normalized Expression, Counts). Center values represent the median. Significance

464   was measured by Wald test with FDR correction (Benjamini-Hochberg).
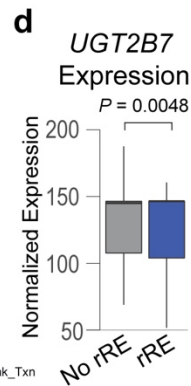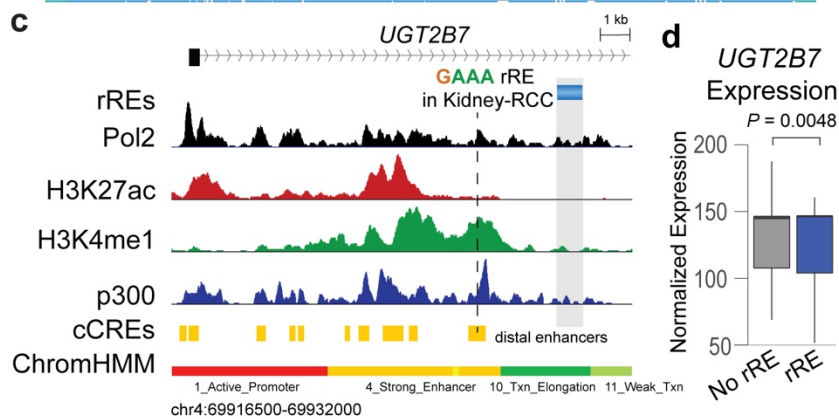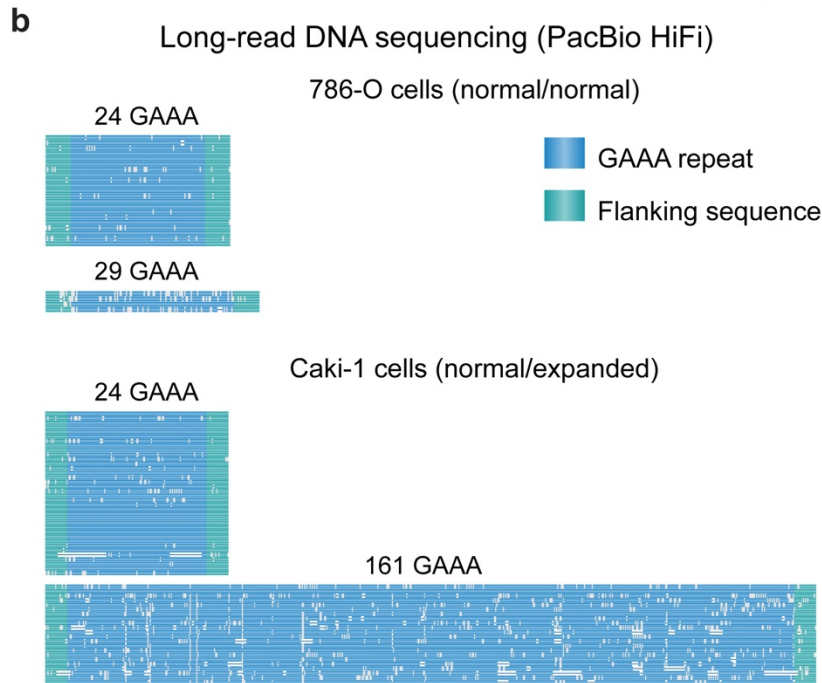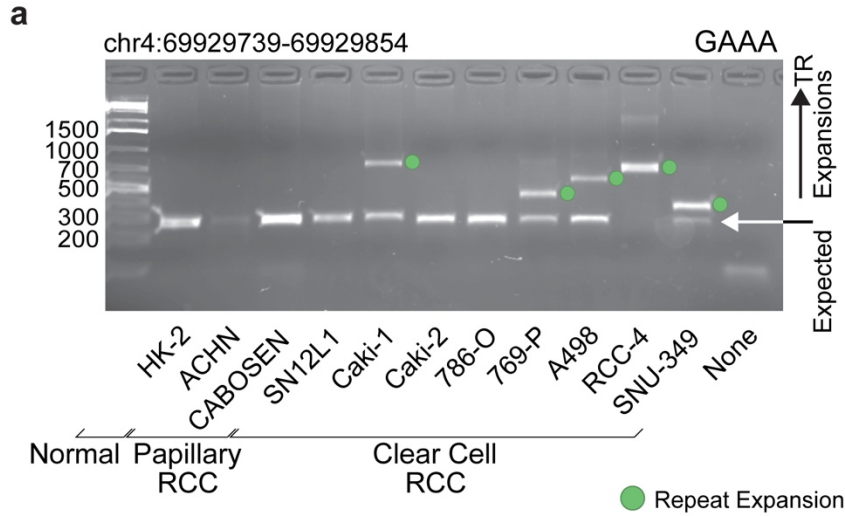
465

466

467 **Figure 5**. **The design and characterization of GAAA-targeting molecules in RCC.** a)

468 Chemical structures of Syn-TEF3, PA3, Syn-TEF4 and PA4. Syn-TEF3, and PA3 target 5′-

469 AAGAAAGAA-3′. Syn-TEF4* and PA4 target 5′-AAGGAAGG-3′. The structures of *N*-

470 methylpyrrole (open circles), *N*-methylimidazole (filled circles), and β-alanine (diamonds) are

471 shown. N-methylimidazole is bolded for clarity. The structure of JQ1 linked to polyethylene

472 glycol (PEG$_6$) is represented as a blue circle. The structure of isophthalic acid and linker is

473 represented as IPA. Complete chemical structures are depicted in **Fig. S2**. The asterisk indicates

474 the site where the R group attaches to the polyamide. Mismatches formed with Syn-TEF4 and

475 PA4 are indicated with orange. b) Relative cell density of RCC cell lines Caki-1 and 786-O

476 following treatment (72 h) with compounds, as indicated. Relative cell density was measured

477 with CCK-8 assay (see **Methods**). Results are mean ± SEM ($n = 4$). c) Quantitation of the

478 percentage of propidium iodide-positive cells. Whiskers are minimum and maximum values. * p

479 < 0.05. P values are from a one-way ANOVA with multiple comparisons. d) Live cell

480 microscopy of Caki-1 and 786-O cells stained with propidium iodide (red) and Hoechst 33342

481 (blue). Scale bars, 100 μm. See also **Extended Data Figure 10**.
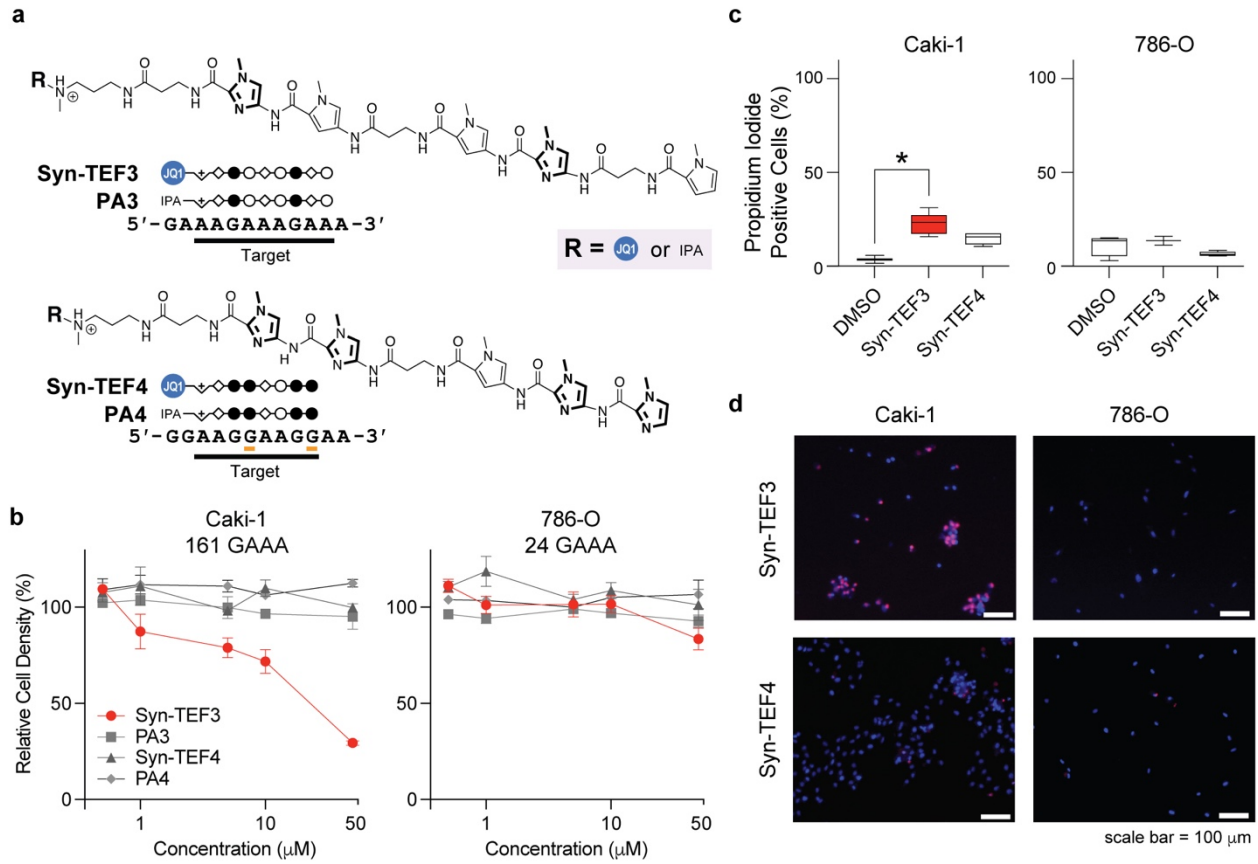
482

483

484   **Methods**

485   **Data curation**

486   We obtained white-listed data from the International Cancer Genome Consortium (ICGC) and

487   The Cancer Genome Atlas (TCGA) pan-cancer analysis of whole genomes (PCAWG) dataset.

488   Data were accessed through the Cancer Genome Collaboratory. We used the aligned reads

489   (BAM files), which were aligned to GRCh37 as described previously[25]. These data are available

490   through the PCAWG data portal (https://docs.icgc.org/pcawg). A list of samples included in the

491   analysis is available in **Table S2**.

492

493   **Identification of somatic recurrent repeat expansions**

494   We analyzed tumor and matching normal samples for each cancer type independently.

495   We executed ExpansionHunter Denovo (EHdn) (v0.9.0)[15] with the following parameters: --min-

496   anchor-mapq 50 --max-irr-mapq 40. To prioritize loci, we developed a workflow termed Tandem

497   Repeat Locus Prioritization in Cancer (TROPIC). We included loci from chr1-22, X, and Y for

498   downstream analysis. We removed loci where >10% of Anchored in-repeat read (IRR) values

499   were >40, which is the theoretical maximum value. The p-value (a non-parametric one-sided

500   Wilcoxon rank sum test) for each locus was used to calculate a false discovery rate (FDR) q-

501   value. Loci with FDR < 0.10 are reported. We selected loci where >5% of samples had an

502   Anchored in-repeat read (IRR) Quotient > 2.5. For a repeat expansion to be detected by EHdn,

503   the tandem repeat must be larger than the sequencing read length. A somatic repeat expansion

504   was defined as having an FDR q-value < 0.05 between tumor and normal samples. To call repeat

505   expansions in individual cancer samples, we analyzed the distribution of tumor and normal

506   Anchored IRR values and selected a conservative threshold for the Anchored IRR Quotient

507   ((Tumor Anchored IRR – Normal Anchored IRR)/(Normal Anchored IRR + 1)) > 2.5 (**Extended**

508   **Data Figure 4**).

509

510   **Local read depth normalization**

511   EHdn normalizes the number of Anchored IRRs for a given locus to the global read

512   depth. To account for chromosomal amplifications and other forms of genetic variation that

513   could alter local read depth, we performed the following normalization. For each rRE locus and

514   sample in its corresponding cancer, samtools v1.13 was used with the parameter depth -r to find

515   the read depth at each base pair within the locus and a 500 bp region surrounding the start and

516   stop positions of the TR. We calculated the average read depth at each base pair and defined this

517   as the local read depth. Finally, we calculated the local read depth-normalized Anchored IRR

518   value specific to a sample and rRE combination by dividing the unnormalized Anchored IRR

519   value from EHdn by the local read depth at the locus.

520

521   **Generation of CABOSEN cells**

522   CABOSEN cells were generated from a cabozantinib-sensitive (CABOSEN) human papillary

523   RCC xenograft tumor grown in RAG2$^{-/-}$ gammaC$^{-/-}$ mice, as described previously[50]. Tumor tissue

524   was minced with a sterile blade and the cell suspension cultured in DMEM/F-12 medium

525   (Corning) supplemented with 10%(v/v) Cosmic Calf Serum (ThermoFisher). Cells were

526   expanded and cryopreserved in growth medium supplemented with 10%(v/v) DMSO and cells

527   from passage 8 were used for analysis.

528

Analysis of rREs by gel electrophoresis

530   We performed PCR with CloneAmp HiFi PCR Mix (Takara Biosciences, Mountain

View, CA) and added DMSO to a final concentration of 5-10% (v/v) as needed. All cell lines

were tested negative for mycoplasma contamination with the MycoAlert Mycoplasma Detection

Kit (Lonza). Cell line identities were authenticated by STR profiling by the Genetic Resources

Core Facility at Johns Hopkins University, with the exception of SNU-349, which did not match

the reported STR profile of SNU-349 or any other catalogued cell line, but has a mutated *VHL*

gene and expresses high levels of *PAX8* and *CA9*, consistent with ccRCC origin. A list of

primers used to analyze the loci is available in **Table S6**.

538

Visualization of repeat expansions with ExpansionHunter and REViewer

540   To inspect the reads supporting a repeat expansion, we annotated the repeat as described

on the GitHub page for ExpansionHunter. We then profiled the region with ExpansionHunter

(v4.0.2) using the default settings[14]. The resulting reads were visualized with REViewer (v0.1.1)

using the default settings. REViewer is available at https://github.com/Illumina/REViewer. A

repeat expansion was called when the repeat tract length for one allele of the tumor sample was

greater than 100 bp and exceeded the repeat tract length of either normal allele. A locus was

considered validated if at least 10 cancer genomes had a repeat expansion.

547

Validation of rREs in independent cohorts of samples

549   Twelve pairs of matching normal and tumor samples from patients with clear cell renal

550   cell carcinoma were obtained with the patients' informed consent *ex vivo* upon surgical tumor

551   resection (Stanford IRB-approved protocols #26213 and #12597) and analyzed. Eighteen and 15

552   pairs of matching normal and tumor samples for prostate and breast cancer, respectively, were

553   obtained from the Tissue Procurement Shared Resource facility at the Stanford Cancer Institute

554   and analyzed. Nucleic acid was isolated with either the Quick Microprep Plus kit (Catalog

555   D7005) or the Zymo Quick Miniprep Plus kit (Catalog D7003) (Zymo Research, Irvine, CA).

556   Gel electrophoresis was performed as described above. A locus was considered detected if a

557   somatic repeat expansion was identified in at least one patient tumor sample compared to a

558   matching normal sample.

559

560   **Downsampling Analysis**

561   For the downsampling analysis, tumor genomes from renal cell carcinoma samples were

562   downsampled from their mean (52x) sequencing depth to 40, 30, 20, and 10x with the samtools

563   view command. EHdn was run, as described above for each of the sequencing depths, and the

564   Bonferroni-corrected p-value was plotted for the recurrent repeat expansion in *UGT2B7* (GAAA,

565   chr4:69929297-69930148).

566

567   **Benchmarking the Local Read-Depth Normalization (LRDN) filter**

568   We benchmarked the local read depth filter *in silico* by observing its behavior with

569   simulated reads. First, we created a reference genome containing artificially expanded repeats.

570   We randomly selected 10 TRs located in chromosome 1 that were less than the sequencing read

571 length of 100 bp. We artificially expanded these TRs in chromosome 1 of GRCh37 with the

572 BioPython python package (version 1.79). Next, we used wgsim (version 0.3.1-r13) to simulate

573 reads from the reference file with the command "wgsim -N 291269925 -1 100 -2 100

574 reference_file.fasta output.read1.fastq output.read2.fastq". The number of reads (specified by the

575 -N option) was calculated to achieve 30x coverage of chromosome 1. The resulting pair of files,

576 hereinafter referred to as the base fastq files, contained a copy number of 2 for all of the

577 expansions.

578 To simulate copy number amplification, the read simulation process was repeated using

579 reference files that contained only the artificially expanded repeats and their surrounding 1,000

580 bp flanks. We created 10 pairs of fastq files, each with an increasing copy number. We specified

581 the copy number by multiplying the number of reads to generate (wgsim -N option) by the

582 required number. To generate the final set of fastq files, we concatenated each pair of copy

583 number-amplified fastq files with the base fastq files. The end result is 8 pairs of fastq files that

584 contain reads of chromosome 1 and a copy number amplification varying from 2 to 10 of the

585 expanded repeats.

586 The base fastq file with a copy number of 2, in addition to the eight copy number-

587 amplified fastq files, were aligned to chromosome 1 of GRCh37 with bwa-mem (v 0.6) with the

588 default options. The resulting SAM files were converted to BAM format with samtools (v 1.15)

589 with the default options. Finally, we ran the EHdn profile command (v 0.9.0) with the minimum

590 anchor mapping quality set to 50 and maximum IRR mapping quality set to 40. Finally, the

591 Anchored IRR values were extracted by overlapping the STR coordinates with the *de novo*

592 repeat expansion calls.

593

594   **Short-read and long-read DNA sequencing**

595   We sequenced Caki-1 and 786-O with both short-read sequencing (60x sequencing

596   coverage, 150 bp paired-end sequencing on a NovaSeq 6000 instrument) and long-read DNA

597   sequencing (50x sequencing coverage, PacBio HiFi sequencing on a Sequel IIe instrument). We

598   aligned the long reads to GRCh37 with pbmm2 v1.7.0, using the parameters --sort --min-

599   concordance-perc 70.0 --min-length 50. We aligned the short reads to GRCh37 with Sentieon

600   (v202112.01) with parameters -K 10000000 -M, an implementation of BWA-MEM, and

601   analyzed the samples with EHdn, as described above. We included loci containing at least one

602   sample with an Anchored IRR value >0 for further analysis. Anchored IRR values >0 arise when

603   the repeat length exceeds the sequencing read length. To benchmark EHdn against long-read

604   sequencing data, we manually determined the TR length of a given locus in the long-read

605   sequencing data. If the TR length in the long-read sequencing data exceeded the short-read

606   sequencing read length of 150 bp, we considered that locus confirmed.

607   The PacBio HiFi data were aligned to GRCh37 with pbmm2 v1.7.0 and visualized at the

608   *UGT2B7* locus with Tandem Repeat Genotyper v0.2.0

609   (https://github.com/PacificBiosciences/trgt).

610

611   **Analysis of rRE loci**

612   To determine if rREs were associated with any human diseases, rREs were mapped to

613   genes with GREAT (v4.0.4, default settings)[51]. The resulting genes were analyzed with Enrichr

614   using Jensen Diseases[52]. To determine whether repeat expansions were associated with

615   microsatellite instability-high (MSI-High) cancers, we obtained data from Hause et al[5]. The

616   percentage of MSI-high cancers was obtained from colon adenocarcinoma (COAD), stomach

617   adenocarcinoma (STAD), kidney renal cell carcinoma (KIRC), ovarian serous

618   cystadenocarcinoma (OV), prostate adenocarcinoma (PRAD), head and neck squamous cell

619   carcinoma (HNSC), liver hepatocellular carcinoma (LIHC), bladder urothelial carcinoma

620   (BLCA), glioblastoma multiforme (GBM), skin cutaneous melanoma (SKCM), thyroid

621   carcinoma (THCA), and breast invasive carcinoma (BRCA) and compared to the number of

622   repeat expansions and the percentage of patients with at least one repeat expansion in the

623   corresponding cancer type from the PCAWG dataset. We also overlapped cancer genomes

624   containing rREs with the microsatellite mutation rate, which we term the STR mutation rate, and

625   MSI calls from Fujimoto et al[29]. The association of rREs with STR mutation rate was assessed

626   with the two-tailed Wilcoxon rank sum test. The association of rREs with MSI calls was assessed

627   with Chi-square test with Yates' correction.

628   To determine whether rREs are associated with known mutational signatures, we

629   downloaded mutational signatures from the ICGC DCC

630   (https://dcc.icgc.org/releases/PCAWG/mutational_signatures/Signatures_in_Samples). We

631   performed a multiple linear regression for each single-base-substitution (SBS) and doublet-base-

632   substitution (DBS) signatures to identify predictors of the number of rREs present in a sample.

633   To choose the predictors, we performed best subset selection on DBS and SBS signatures and

634   included age as a possible confounding factor. We used the statsmodels v0.12.2 in Python and,

635   specifically, the ordinary least squares model found in the statsmodels.api.OLS module to

636   estimate the coefficients of the selected predictors in their corresponding multiple linear

637   regression model[53].

638   To determine whether repeat expansions were associated with a difference in cytotoxic

639   activity, we calculated cytotoxic activity as previously described for four cancers that had

640   matching RNA-seq and WGS[41]. For each locus, we compared cytolytic activity for patients with

641   a repeat expansion to patients without a detected repeat expansion using a Welch's t-test with

642   correction for multiple hypothesis testing (Benjamini-Hochberg FDR q-value < 0.05). rREs were

643   annotated with genic elements using annotatr (v1.18.1)[33].

644   To determine if rREs were associated with regulatory elements, we downloaded

645   candidate cis-regulatory elements (cCREs)[34] and mapped them to GRCh37 with LiftOver

646   (UCSC)[54]. We determined the distance between rREs and cCREs with the bedtools closest

647   command (v2.27.1)[55], and compared this distance to the simple repeats catalog[56]. To compare the

648   distance to ENCODE cCREs, a Welch's t-test was performed.

649   To determine if prostate cancer rREs were associated with prostate cancer susceptibility

650   loci[36], we calculated the distance to three sets of loci using the "bedtools closest" command. We

651   calculated the distance between (1) rREs present in prostate cancer samples and prostate cancer

652   susceptibility loci, (2) rREs not present in cancer samples and cancer susceptibility loci, and (3)

653   simple repeats and cancer susceptibility loci. To compare the distances between these three

654   associations, we performed a Welch's t-test with FDR correction (Benjamini-Hochberg).

655   To determine whether rREs were associated with replication timing, we downloaded

656   Repli-seq replication timing data from seven cell lines from the ENCODE website (NCI-H460,

657   T470, A549, Caki2, G401, LNCaP, and SKNMC)[57]. We selected regions for which all cell lines

658   had concordant signals for analysis (early or late replication designations agreed for each cell

659   line at a given locus). We determined whether there was a difference in the distribution of rREs

660   across early- and late-replicating regions compared to the simple repeats catalog with

661   bootstrapping ($n=10,000$). We sampled 54 loci (the number of rREs that are present in a

662   concordant replication region) from rREs and simple repeats. A Welch's t-test was performed on

663   the bootstrapped samples to estimate a p-value. We applied FDR correction (Benjamini-

664   Hochberg) to the estimated p-values. To determine whether rRE status in *UGT2B7* was

665   associated with survival outcome in clear cell RCC patients (TCGA abbreviation: KIRC), we

666   used Welch's t-test quartile.

667   To identify motifs enriched and depleted in the rRE catalog, we followed the same

668   method used in the Motif-Scan python module (v1.3.0)[58]. We compared our rRE catalog to

669   Simple Repeats (Tandem Repeat Finder, TRF) as a control. For each unique motif present, we

670   built a contingency table specifying the count of rREs and Simple Repeats with and without the

671   motif. Two one-tailed Fisher's exact tests were applied to the table to test for significance in both

672   directions, enrichment and depletion. The "stats" module in the Scipy python package (v1.7.0)

673   was used to conduct the significance test. Since multiple hypothesis tests were performed, we

674   applied FDR correction (Benjamini-Hochberg) for multiple hypothesis testing to the p-values,

675   with a cutoff (FDR) of 0.01.

676   For the comparison of SNVs in COSMIC genes to rREs, we first divided cancer genomes

677   into two categories: rRE cohort and non-rRE cohort. The rRE cohort contains all of the genomes

678   that have at least one rRE detected (n = 615) and the non-rRE cohort contains all of the genomes

679   that have no rREs detected (n = 1897). We then looked at the number of donors in the rRE cohort

680   that have at least one mutation on a given gene (COSMIC Tier 1 genes) i and the number of

681   donors in the non-rRE cohort that have at least one mutation on a given gene i with a

682   contingency table. We then calculated the p-value (Fisher's exact test) for the significance of

683   associating genes to either rRE or non-rRE cohort. This p-value calculation is repeated for all

684   COSMIC genes and then an FDR at 0.05 significance level (Benjamini-Hochberg) was

685   employed to correct for multiple hypothesis testing.

686

Estimation of expansions in the general population

688   To estimate the frequency of the rREs in the general population, ExpansionHunter

689   Denovo (version 0.9.0) was run on 1000 Genomes Project samples[59] (n = 2,504) (GRCh38) and

690   Medical Genome Reference Bank[60] samples (n = 4,010) (GRCh37 lifted over to GRCh38).

691   The genomic coordinates of the 160 rREs (GRCh37) were padded with 1,000 bp and

692   translated to GRCh38 coordinates with the UCSC LiftOver. Then, these rRE coordinates

693   (GRCh38) were overlapped with loci from the population samples containing the Anchored IRR

694   calls. rREs that overlapped with matching motifs in the population samples were selected for

695   further analysis. We next sought to identify expanded rREs in the population samples to quantify

696   their prevalence. To do so, we converted their global-normalized Anchored IRR values to be

697   comparable to ICGC values. This step was necessary because sequencing read lengths from the

698   PCAWG dataset are generally 100 bp while the read lengths from 1000Genomes and Medical

699   Genome Reference Bank are 150 bp. The conversion follows the formula (Anchored IRR, 100

700   bp) = 0.5 + 1.5 * (Anchored IRR, 150 bp)[15]. A sample in the population samples was counted as

701   expanded if its Anchored IRR value was greater than the 99th percentile of Anchored IRR values

702   in the normal samples from the PCAWG dataset, a threshold that is comparable to the threshold

703   used to call expansions in tumor samples (**Extended Data Figure 4**). In future rRE catalogs, for

704   the rare instance where the estimated frequency of repeat expansions in the population samples is

705   higher than expected, these data could be used to further filter rREs to improve the detection of

706   cancer-specific repeat expansions.

707   To compare the length of TRs in normal samples with and without a matching rRE in a

708   tumor sample, donors in the Prost-AdenoCA and Kidney-RCC cohorts whose data are available

709   for download through the Cancer Collaboratory were included (n=253). We used

710   ExpansionHunter (v5.0.0) with the default options to genotype prostate and kidney cancer rREs

711   in the normal samples of the selected donors. When there were two alleles of an rRE in a sample,

712   both alleles were included and treated as distinct data points. For each rRE, we tested whether

713   the distribution of genotypes from donors who have an expansion in their tumor samples differed

714   from donors who did not have an expansion. Student's t-test was used to compute p-values, and

715   FDR-correction (Benjamini-Hochberg) to adjust for multiple hypothesis testing.

716

717   **Association of rREs with gene expression**

718   Matching RNA-seq and WGS data were available for Kidney-RCC, Ovary-AdenoCA,

719   Panc-AdenoCA, and Panc-Endocrine. RNA-seq data from these samples were obtained from

720   DCC (https://dcc.icgc.org/) and values were converted to transcripts per million (TPM).

721   Normalized gene expression (TPM) values were compared for samples with and without an rRE

722   (Welch's *t*-test, with FDR correction). For isoform analysis, normalized gene expression counts

723   were compared for samples with and without a repeat expansion using the DESeq2 (v1.32.0)

724   package in R v4.0.5. We used the DESeq function to calculate the $\log_2$ fold change values for 3

725   isoforms of the *UGT2B7* gene (ENST00000305231.7, ENST00000508661.1,

726   ENST00000502942.1) and performed a Wald test with FDR correction using the Benjamini-

727   Hochberg procedure (threshold q-value < 0.01).

728

729   **Design, synthesis, and characterization of Syn-TEFs and PAs**

730   Synthetic transcription elongation factors (Syn-TEFs) and polyamides (PAs) were

731   designed to target a GAAA repeat (Syn-TEF3 and PA3) or a control G**G**AA repeat (Syn-TEF4

732   and PA4). Syn-TEF3, Syn-TEF4, PA3, and PA4 were synthesized and purified to a minimum of

733   95% compound purity by WuXi Apptec and used without further characterization. HPLC

734   conditions for chemical characterization: 1.0 mL/min, Solvent A: 0.1% (v/v) trifluoroacetic acid

735   (TFA) in H2O, Solvent B: 0.075% (v/v) TFA in acetonitrile, Gemini, Column: C18 5 μm 110A

736   150*4.6mm. Full results of characterization can be found in **Fig. S2**.

737

738   **Treatment of RCC cell lines with synthetic transcription elongation factors (Syn-TEFs)**

739   Caki-1, and 786-O, and Caki-2 cells were obtained from ATCC and grown in RPMI 1640 media

740   with L-glutamine (Gibco Catalog 11875093), supplemented with 10% (v/v) FBS. A498 and

741   ACHN cells were obtained from ATCC and grown in DMEM media with glucose, L-glutamine,

742   and sodium pyruvate (Corning Catalog 10-013-CV), supplemented with 10% (v/v) FBS. RCC-4

743   cells were obtained from Amato Giacca (Stanford University) and grown in DMEM media with

744   glucose, L-glutamine, and sodium pyruvate (Corning Catalog 10-013-CV), supplemented with

745   10% (v/v) FBS. Cell lines were confirmed by STR profiling (Genetic Resource Core Facility,

746   Johns Hopkins University) and tested negative for mycoplasma. Cells were seeded in 96-well

747   plates on day 0. On day 1, cells were treated with the indicated molecules. Molecules were

748 dissolved in DMSO (vehicle) and added to cells (0.1% (v/v) DMSO final concentration). On day

749 4 (72 h later), relative metabolic activity was measured as a proxy for relative cell density, with

750 the Cell Counting Kit (CCK-8; Dojindo Molecular Technologies) per the manufacturer's

751 instructions. Absorbance (450 nm) of cells treated with molecules was normalized to DMSO

752 (0.1%(v/v)) or no treatment. Absorbance was measured with an Infinite M1000 microplate

753 reader (Tecan, Mannedorf, Switzerland).

754 For microscopy, Caki-1 and 786-O cells were plated on glass-bottom 96-well plates

755 under standard culture conditions. One day after plating, media containing either no drug,

756 0.1%(v/v) DMSO, 50 μM Syn-TEF3, or 50 μM Syn-TEF4 was added, and the cells were

757 incubated for 72 hours at 37°C. As a control, wells that received no treatment were incubated

758 with 70%(v/v) ethanol for 30 seconds prior to staining. Cells were then stained with propidium

759 iodide and Hoechst 33342 from the Live-Dead Cell Viability Assay Kit (Millipore Sigma,

760 Catalog CBA415) according to manufacturers' instructions and immediately imaged at 10x

761 magnification with a 0.17 numerical aperture CFI60 objective on a Keyence BZ-X710

762 microscope. Four replicates were measured for each treatment condition, and the experiment was

763 repeated three times. Quantitation was conducted using FIJI software. For statistical analyses, a

764 one-way ANOVA with multiple comparisons was conducted with GraphPad Prism.

765

Methods References

766

767   50.   Zhao, H., Nolley, R., Chan, A. M. W., Rankin, E. B. & Peehl, D. M. Cabozantinib inhibits

768   tumor growth and metastasis of a patient-derived xenograft model of papillary renal cell

769   carcinoma with MET mutation. *Cancer Biol. Ther.* **18**, 863–871 (2017).

770   51.   McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions.

771   *Nat. Biotechnol.* **28**, 495–501 (2010).

772   52.   Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment

773   analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

774   53.   Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with

775   Python. in *Proceedings of the 9th Python in Science Conference* (eds. van der Walt, S. &

776   Millman, J.) 92–96 (2010). doi:10.25080/Majora-92bf1922-011.

777   54.   Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006

778   (2002).

779   55.   Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic

780   features. *Bioinformatics* **26**, 841–842 (2010).

781   56.   Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids*

782   *Res.* **27**, 573–580 (1999).

783   57.   Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome.

784   *Nature* **489**, 57–74 (2012).

785   58.   Sun, H. *et al.* Quantitative integration of epigenomic variation and transcription factor

786   binding using MAmotif toolkit identifies an important role of IRF2 as transcription

787   activator at gene promoters. *Cell Discov.* **4**, 38 (2018).

788   59.   Altshuler, D. M. *et al. An integrated map of genetic variation from 1,092 human genomes.*

789   *Nature* vol. 491 (2012).

790   60.   Pinese, M. *et al.* The Medical Genome Reference Bank contains whole genome and

791   phenotype data of 2570 healthy elderly. *Nat. Commun.* **11**, 435 (2020).

792

Data availability

794   Whole-genome sequencing data (both short- and long-read DNA sequencing) from 786-

795   O and Caki-1 cell lines are deposited in NCBI with accession PRJNA868795.

796

805

806 **Author Contributions**

807 G.S.E. conceived the study. G.S.E., G.G., A.C.F., J.T.L., M.A.E., M.P.S., and M.G. supervised

808 research. G.S.E., G.G., R.A., A.S, E.D., J.P., C.M.B., K.Z., R.K.C.Y., and A.A.E. analyzed data.

809 G.S.E., C.R.H., L.R., A.A., A.A., K.V.K., R.A.K., D.A.S., S.M.W., and T.J.M. conducted wet

810 lab experiments. G.S.E. and M.P.S. wrote the manuscript with input from all the authors.

811

812 **Competing Interests declaration**

813 G.S.E. and M.P.S. are inventors on a patent application describing anti-proliferative agents. E.D.

814 and M.E. are shareholders and currently or formerly employed by Illumina and Pacific

815 Biosciences.

816

817 **Additional Information**

818 **Supplementary Information** is available for this paper.

819 Correspondence and requests for materials should be addressed to gerwin@stanford.edu,

820 mpsnyder@stanford.edu, or mark@gersteinlab.org.