medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Leveraging sequences missing from the human genome to diagnose cancer

Ilias Georgakopoulos-Soares[1,2*], Ofer Yizhar Barnea[1,2*], Ioannis Mouratidis[3], Rachael Bradley[1,2], Ryder Easterlin[1,2], Candace Chan[1,2], Emmalyn Chen[4], John S. Witte[4,5,6], Martin Hemberg[7,8^], Nadav Ahituv[1,2^]

[1]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, USA.
[2]Institute for Human Genetics, University of California San Francisco, San Francisco, California, USA.
[3]Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium.
[4]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, USA
[5]Department of Urology, University of California San Francisco, San Francisco, California, USA
[6]Departments of Epidemiology and Population Health and Biomedical Data Science, Stanford University School of Medicine, Stanford, California, USA
[7]Wellcome Sanger Institute, Hinxton, UK.
[8]Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, USA

*These authors contributed equally to the work.
^Corresponding authors: mhemberg@bwh.harvard.edu, nadav.ahituv@ucsf.edu

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

ABSTRACT

Cancer diagnosis using cell-free DNA (cfDNA) can significantly improve treatment and survival but has several technical limitations. Here, we show that tumor-associated mutations create neomers, DNA sequences 11-18bp in length that are absent in the human genome, that can accurately detect cancer subtypes and features. We show that we can detect twenty-one different tumor-types with higher accuracy than state-of-the-art methods using a neomer-based classifier. Refinement of this classifier via supervised learning identified additional cancer features with even greater precision. We also demonstrate that neomers can precisely diagnose cancer from cfDNA in liquid biopsy samples. Finally, we show that neomers can be used to detect cancer-associated non-coding mutations affecting gene regulatory activity. Combined, our results identify a novel, sensitive, specific and straightforward cancer diagnostic tool.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

INTRODUCTION

Cancer is the second leading cause of death worldwide (*1*, *2*), and for most cancer types survivability is significantly higher if the tumor is detected at an early stage (*3*, *4*). Currently mass population screening is applicable only for breast and cervical cancers and utilizes physical tests like mammography and cytology screens. Detection for other cancer types, done both *en masse* and in a low and affordable resource setting, still poses a major challenge for the scientific and clinical communities (*5*). In particular, a major hurdle is to identify reliable biomarkers for the detection of cancer development at a presymptomatic stage. Detection at such an early stage would allow for patient stratification and improvement of patients' outcome by providing personalized treatments.

Circulating cell-free DNA (cfDNA) is an emerging and promising resource for cancer diagnostics and prognostics (*6*, *7*). It has a short life span (16 minutes to 2.5 hours), which makes it a highly temporal indicator of various processes occurring in the subject's body. Due to advances in sequencing technologies, cfDNA can be rapidly analyzed at a relatively low cost. Analysis of circulating tumor DNA (ctDNA) has become a prospective minimally invasive tool to screen the population and to monitor patients already diagnosed with cancer. Current applications, including identification of tissue of origin and cancer type, minimal residual disease and other technologies rely on sequencing to resolve somatic mutations (*8*) and epigenetic marks, such as DNA methylation or histone modifications that can determine the cancerous tissue (*9*, *10*). However, ctDNA still has many hurdles and caveats that need to be overcome (*11*). Some of the major hurdles include: 1) cfDNA is fragmented (180-360 base pairs) making its collection and extraction more challenging and the tumour-derived DNA makes up only a small portion (estimated to be around 0.4%; (*11*)) warranting the need for extremely sensitive biomarkers that can easily detect the presence of cancerous cells; 2) prior knowledge of specific mutations or methylation marks is required for targeted screening, and consequently the main focus has been on coding mutations which only constitute a small fraction of mutations; 3) cfDNA mutation and epigenetic diagnosis could be confounded by somatic alterations in white blood cells (*12*); 4) the diagnostic techniques used to detect methylation or histone marks are technologically complex and can have low sensitivity and specificity (*6*, *13*–*15*); 5) and to provide the most optimal cancer treatment, it needs to be diagnosed at preliminary stages when the tumor is small (~5mm in diameter). At these stages, the tumor produces minute levels of ctDNA that are difficult to detect using current methods (*6*).

Nullomers are short DNA sequences (11-18 base pairs) that are absent from the human genome (*16*, *17*). While the absence of nullomers could be due to chance, we and others have shown that a significant proportion of them is under negative selection pressures (*17*, *18*), suggesting that they could have a deleterious effect on the genome. We have also shown that these sequences could be used as DNA 'fingerprints' to identify specific human populations (*18*). As nullomers do not exist in a human genome, their appearance due to mutagenesis followed by clonal expansion could be exploited as a diagnostic method for diseases associated with a mutational burden, such as cancer.

Here, we set out to test whether nullomers could be used as a diagnostic tool to detect cancer in general and also specific subtypes. Throughout this manuscript we refer to nullomers found in the tumor genome as *neomers* to distinguish them from the more general category. We first analyzed The Cancer Genome Atlas (TCGA;(*19*)) database finding recurrent neomers created by somatic mutations that could be used to detect not only cancer subtypes with higher accuracy than leading methods (*20*) but also additional cancer features. Further analyses of cfDNA whole-genome sequencing datasets found

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

that these neomers can also be used to detect cancer subtypes in these data without the need for matched healthy control samples. Finally, we show that cancer-associated neomers can be used to detect cancer-associated mutations in gene regulatory regions and functional assays of prostate cancer associated neomers show that they have a functional effect on their regulatory activity. Combined, our results show that neomers can be used as a rapid, sensitive, specific and straightforward cancer diagnosis and also aid in the identification of gene regulatory mutations associated with cancer.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

RESULTS

Annotation of mutations that lead to nullomers

As cancer causes DNA mutations, we investigated if they can result in the resurfacing of nullomers (**Fig. 1A**). Using our previously characterized human nullomers (*18*), we analyzed whole-genome sequencing results from 2,577 patients across 21 different cancer types from The Cancer Genome Atlas (TCGA; (*21*)),(**Fig. S1A)** for resurfacing nullomers. We focused on 16bp nullomers as it is the shortest length where we detect a sufficient number of nullomers per patient, with the human reference genome having only 37.24% of all possible 16mers. The majority of the 44,599,472 single nucleotide substitutions gives rise to multiple nullomers and we identified 213,164,038 resurfacing nullomers across all cancer types. Furthermore, we identified 2,470,091 nullomers resulting from short insertions and deletions (1-100 bp). The median number of nullomers created by each substitution was two and for indels four (**Fig. S1B-C**). On average, 58.29% of substitutions in a patient resulted in one or more nullomers, with only 2.1% of the nullomers derived from coding regions. The median number of nullomers found across cancer patients was 9,107 (**Fig. 1B-C**) and the number of nullomers was directly proportional to the number of mutations (**Fig. S1D**) and nullomer length (**Fig. S1E**). As mutations were identified by comparing to healthy tissues, mutations did not need to be filtered for common variants that could otherwise result in nullomers (*18*).

As we were interested in nullomers that could be used as cancer biomarkers, we focused on the subset of nullomers that are recurrent, i.e. those found in more than one patient for a specific cancer type, termed hereafter as *neomers*. The number of neomers was proportional to the total number of mutations (**Fig. 1C**) and as both the number of patients per cancer type and the mutational load varied, the median number of neomers for each tissue type ranged from 0-98. Analysis of the most frequent neomers revealed several previously known cancer-associated mutations (**Table 1**). For example, some of the most recurrent coding neomers were the result of either the Gly12Asp, Gly12Val or Gly12Cys missense mutation in the KRAS proto-oncogene GTPase (KRAS), which are known to make up 80% of cancer-associated KRAS mutations and lead to KRAS being constitutively active (*22*, *23*). Although KRAS has been associated with several cancers, 190/215 (88%) of these mutations were found in pancreatic cancers. Several frequently occurring coding neomers were also found in other known cancer-associated genes such as tumor protein p53 (*TP53*), B-Raf proto-oncogene, serine/threonine kinase (*BRAF*) and phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (*PIK3CA*) .The mutation that most often resulted in a neomer was located in a noncoding region, within the telomerase reverse transcriptase (*TERT*) promoter, which is known to be associated with numerous cancer types (*24*). This mutation, called -124C>T or C228T, is extremely common in numerous cancer types (*25*) and thought to disrupt a G-quadruplex (*26*) leading to the binding of GAPB (*27*), an ETS transcription factor, resulting in increased *TERT* expression. We found this mutation in 97 patients with the highest incidence in glioblastoma (51%), fitting with its high prevalence rate and diagnostic use for this cancer type (*28*).

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Table 1. Common cancer-associated neomers. Six of the most common neomers created by a single mutation.

Neomer	Locus	Coordinate (hg38)	Mutation	Name	Patient #	Cancer #
AGGGCCCGGAAGGGGC	*TERT*	chr5:1295113	G>A	-124C>T C228T	97	7
TCTTGCCTACGCCATC	*KRAS*	chr12:25245350	C>T	c.35G>AGly12Asp	89	6
CTGTTGGCGTAGGCAA	*KRAS*	chr12:25245350	C>A	c.35G>TGly12Val	79	6
ACGCCACGAGCTCCAA	*KRAS*	chr12:25245351	C>G	c.34G>T Gly12Cys	47	1
GGTGCATGTTTGTGCC	*TP53*	chr17:7673802	C>T	c.818G>A Arg273Pro	35	13
GTGGGGGCAGTGCCTC	*TP53*	chr17:7675088	C>T	c.524G>A Arg175His	28	11

We also identified several neomers that are frequently created by different mutations (**Table 2**). Interestingly, some of these frequently recurrent nullomers are created by different mutations, yet are predominantly found in one cancer. For example, GTTTTTCTCCTAGACC is found 40 times in skin cancer at 31 different loci while CTGGCAGTGAGCCACG is found 21 times in liver cancer across 18 loci. The majority (98%) of these frequent neomers reside in noncoding regions, and many of them reside in intronic regions (35%). For example, CGACGTTCTGCCCACT is found in 32 loci, primarily in pancreatic and stomach cancer. Of those loci, 21/32 (65.6%) were found in distal intergenic regions nearby pancreatic cancer associated genes, such as the C-C motif chemokine ligand (*CCL4*) (*29*), and is also found in an intron of POM121 transmembrane nucleoporin like 12 (*POM121L12*) which is commonly mutated in gastrointestinal cancers (*30*) in the vicinity of the potassium voltage-gated channel modifier subfamily V member 1 (*KCNV1*) gene where promoter hypermethylation has been associated with both pancreatic (*31*) and esophageal cancer (*32*).

Table 2. Frequent cancer-associated neomers. Five of the top frequently recurring neomers created by several different mutations.

Neomer	Loci #	Patient #	Cancer #	Cancer type
CGACGTTCTGCCCACT	32	74	7	Mainly pancreatic and stomach
GTTTTTCTCCTAGACC	32	41	2	All but one in skin
CTGCAGTGGCGCAATA	30	30	11	A third of the mutations in colorectal
TTATAGGGGTCCAGTG	25	26	1	Colorectal only
CTGGCAGTGAGCCACG	23	26	2	21 in liver, 5 esophagus

Generation of a cancer subtype neomer classifier

Based on the observation that most neomers are predominantly found in one cancer type, we hypothesized that they can be used to distinguish between cancer types. We filtered neomers by keeping only those that appeared $>=r_i$ times in specific cancer type *i* (**Table S1**). Comparison of the set

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

of neomers associated with each cancer type reveals a small overlap, as indicated by the Jaccard index which is <0.04, suggesting that each cancer type has a distinct neomer signature (**Fig. 1D**). The only exceptions were esophagus and stomach cancer that are known to have similar characteristics (*33*). We also counted the number of times neomers are found in each patient, finding that patients are strongly enriched for only one set of cancer specific neomers (**Fig. 1E**).

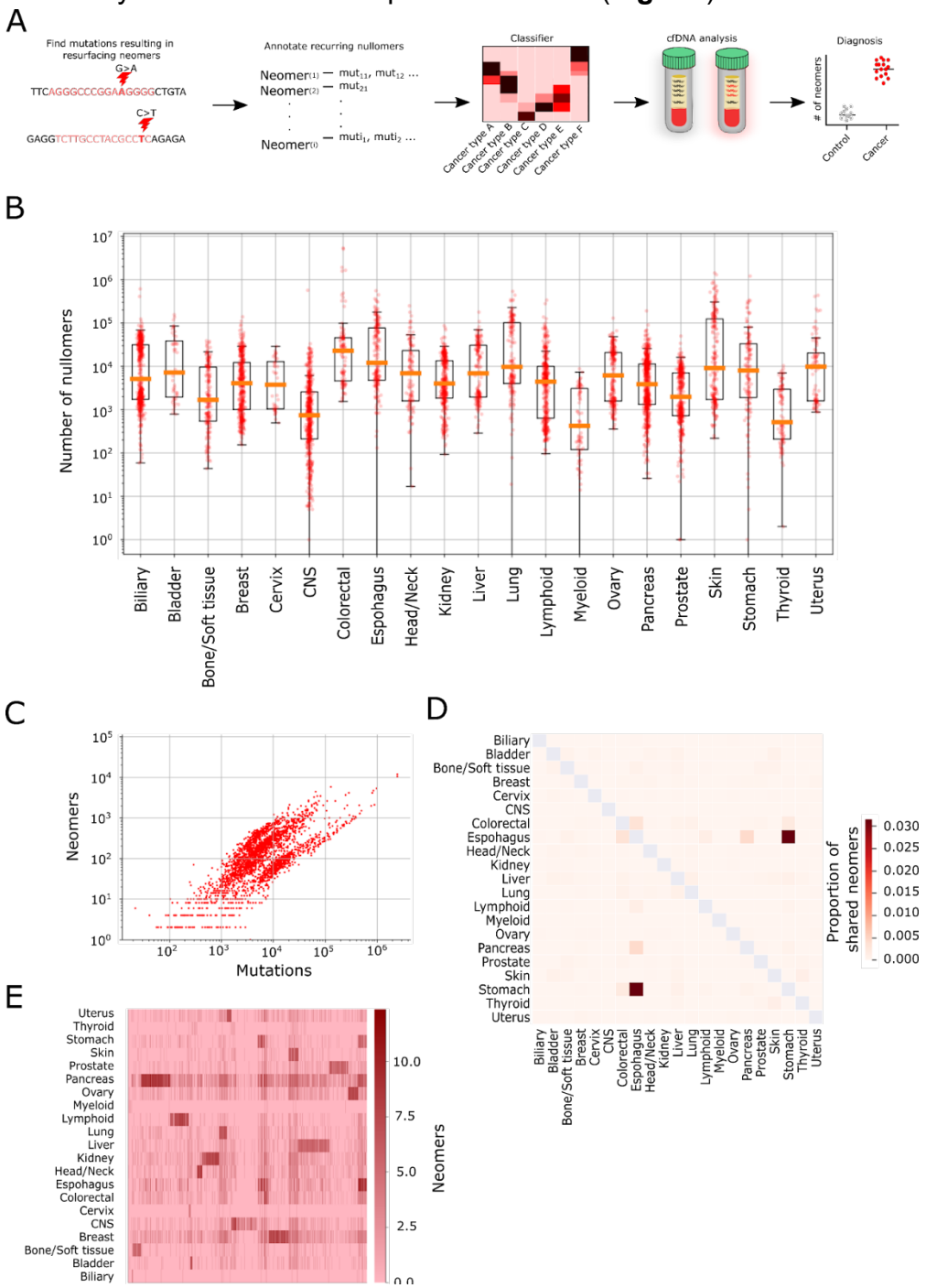

Figure 1. Neomers across 2,577 patients and 21 tissues in the TCGA dataset: (**A**) Schematic overview of our pipeline for identifying neomers and using them to distinguish and detect tumors. (**B**) Number of neomers per patient sample across tissues. Each dot represents a patient sample. (**C**) Number of neomers and the number of substitutions for 2,577 patients (Spearman's rho = 0.75). (**D**) The heatmap shows the Jaccard index for the overlap of neomer sets associated with different cancer types. (**E**) Heatmap showing the occurrence of the neomers across patients. Each row represents a patient and the intensity of the heatmap (log2-scale) shows the number of nullomers from each tissue set.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

To test if neomers can classify tumor samples, we trained a support vector machine classifier to identify tumor type. The classifier takes as input a 21-dimensional vector indicating the number of neomers found for each cancer specific set. Evaluation using 10-fold cross-validation, revealed that our classifier achieves both high sensitivity and specificity, with an F1 score of 0.92 and an accuracy of 0.99 (**Fig. 2A,B**). The performance was better than the deep learning model recently presented by Jiao et al (*20*) and also required less computational resources to train.

Neomers can distinguish additional cancer features

We next tested whether a supervised approach, i.e. using neomers that are thought to be informative based on prior biological knowledge, would improve performance. We first considered microsatellite unstable (MSI) and microsatellite stable (MSS) cancers. MSI is associated with better cancer prognosis, increased benefits from surgery and higher sensitivity to immunotherapy, but with a lack of efficacy from adjuvant treatment (*34*). Since MSI cancers are associated with a proliferation of polyA and polyT stretches, we hypothesized that neomers containing these motifs would be able to distinguish these two cancer types. We identified ten MSI samples from a cohort of 560 breast cancers (*35*) and compared to ten randomly selected MSS samples from the same cohort. We found that the polyA/T neomers were able to flawlessly separate the two categories (**Fig. 2C**).

We next applied a similar strategy to distinguish patients with DNA polymerase epsilon catalytic subunit (POLE) deficiency, as these tumors are known to respond more favorably to immune checkpoint inhibitors (*36*–*38*). We identified 25 patients from the TCGA dataset labelled as POLE deficient, and searched for neomers created through a TCT>TAT or TCG>TTG mutation, which are the most common types of mutations in this context (*38*). Comparing against POLE proficient tumours, we found that the number of neomers identified for each group have very little overlap (**Fig. 2D**), and the classifier achieved an accuracy of 96%.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

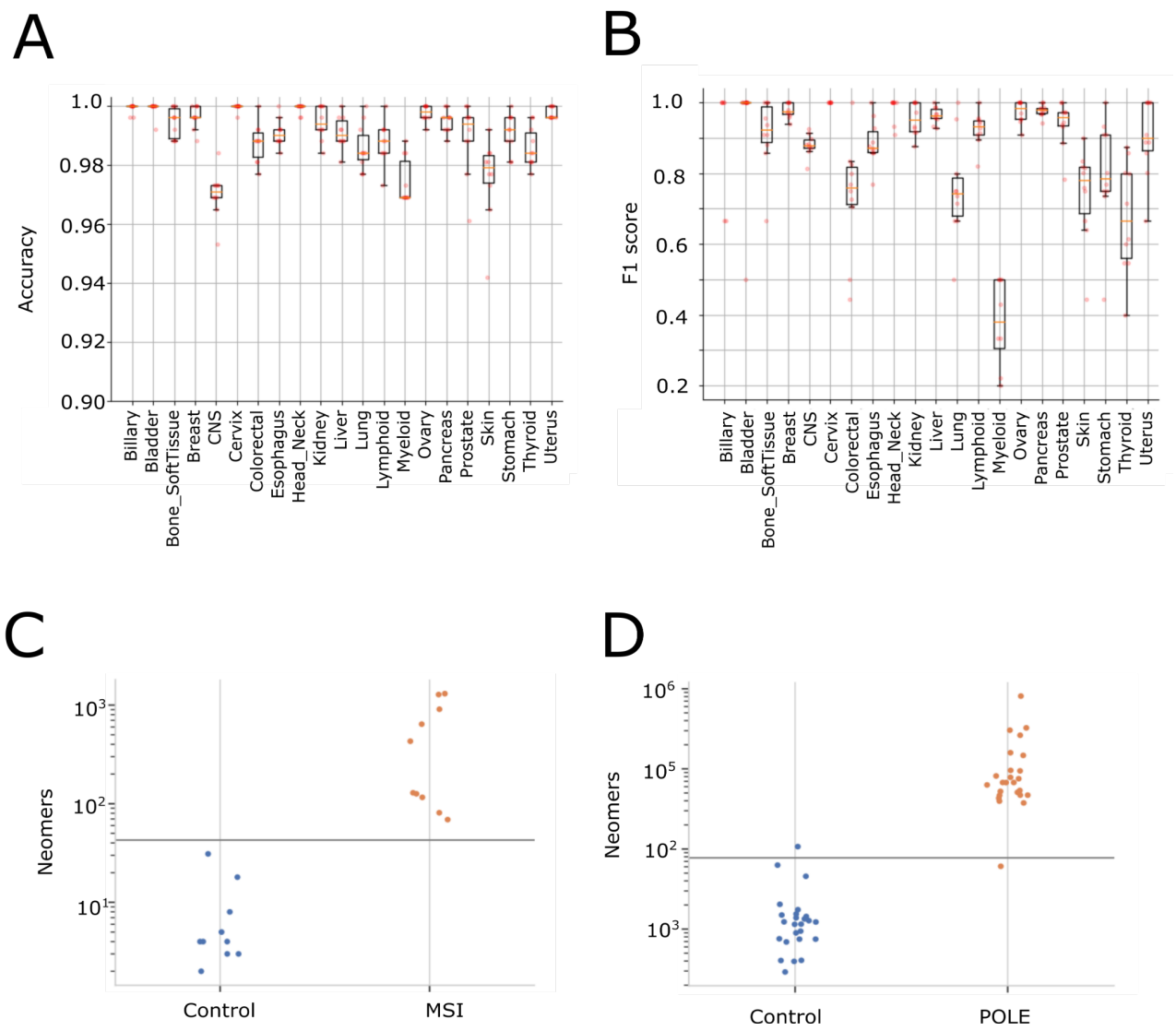

Figure 2. Neomers can distinguish cancer features. (**A**) Accuracy of classifier using an unsupervised classifier and (**B**) F1 score for the same classifier, (**C**) Separation of MSI and MSS samples using a supervised selection of nullomers. (**D**) Separation of POLE proficient and deficient samples using nullomers. In (**C**) and (**D**) vertical line displays the harmonic mean.

Neomers are enriched in cfDNA

We next tested whether neomers could be used to diagnose cancer in cfDNA. We focused on prostate cancer, due to the following reasons: 1) It is the fifth leading cause of death worldwide, the second most frequent cancer in males and is responsible for 3.8% of all cancer deaths (*39*); 2) the availability of both WGS datasets and cfDNA samples; 3) the current primary screen for this cancer, measuring levels of the prostate-specific antigen in the blood, has high false negative and false positive rates (*40*); 4) the need for more accurate screening for minimal residual disease after treatment or surgical interventions (*41*, *42*); 5) the number of neomers per this subtype (N=5,270, median per patient=29.5) is on the low end of all 21 tissues that we analyzed (**Table S1**), allowing us to asses our ability to diagnose cancer with a relatively small number of neomers; 6) localized prostate cancer has a low abundance of ctDNA

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

making it difficult to detect by ultra low pass WGS or targeted cfDNA sequencing (*43*) or via methylation (*44*) compared to metastatic (*45*), providing a challenging test for our neomer approach.

We analyzed three previously published cfDNA WGS samples from metastatic prostate cancer patients and twenty-three controls, sequenced at depths 50-80x (*46*). We first excluded all neomers variants that are not rare due to germline variants (*18*). For each prostate cancer associated neomer we characterized all possible single nucleotide substitutions in the reference genome that could give rise to this neomer. By intersecting this list of neomer creating substitutions with known germline variants identified by the gnomAD project (*47*), we calculated the probability that each neomer will be present in an individual. We excluded all neomers that are found in the population with $p>0.0005$, leaving us with 3,193 prostate neomers.

Another source of neomers in cfDNA WGS data could be sequencing errors. To exclude neomers that were observed due to these technical artifacts, we developed a Poisson model (see Methods). Since sequencing errors are assumed to be distributed uniformly, neomers arising for this reason will have a profile that differs from neomers stemming from sequences that are present in the cfDNA, even at a low allele frequency. Moreover, these neomers will also differ from ones present due to germline variants which will be found at a higher frequency. After filtering our data for neomers likely to have arisen due to germline variants or sequencing errors, we compared the enrichment of reads containing neomers associated with prostate cancer as well as the number of significantly enriched neomers in cases versus controls. We found that the median number of prostate neomers detected in the patient samples is eight, while in the healthy controls we detect two. Similarly, the mean enrichment compared to the expected number of reads was 1.35 compared to 0.93 for the healthy controls (**Fig. 3**). Taken together, these results demonstrate that our prostate neomers classifier could serve as a sensitive and specific assay for identifying cancer in cfDNA samples.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

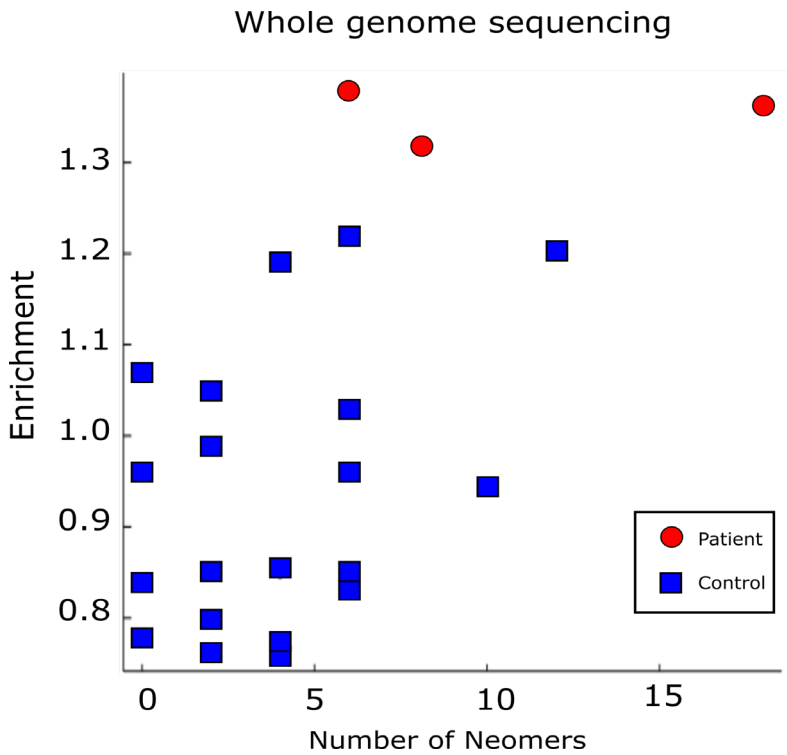

Figure 3. Identification of cancer in liquid biopsy samples using neomers. Cancer status detection in prostate patients and healthy controls from whole-genome sequencing of liquid biopsy samples.

Neomers alter promoter activity

Only a small number of mutations in gene regulatory elements that affect gene expression have been found to be associated with cancer (*48*, *49*). As the vast majority of our cancer-associated nullomers reside in noncoding sequences, we tested whether nullomers could identify cancer-associated gene regulatory mutations that have a functional effect. Of note, our top neomer mutation was in the *TERT* promoter (**Table 1**), which is associated with numerous cancers (*24*). Focusing on prostate cancer, we selected five neomers for luciferase reporter assays using the following criteria: i) neomers that reside in a promoter based on ENCODE annotations (*50*); ii) the gene regulated by the promoter is associated with prostate cancer. Our list included neomers in: 1) a promoter between two divergent genes, *RPS2* and the lncRNA gene *SNHG9* (**Fig. 4A**), both of which are overexpressed in prostate cancer (*51*); 2) a promoter between two divergent genes, *TMEM127* and *CIAO1* (**Fig. 4B**), with the former being downregulated in prostate cancer (*52*); 3) a promoter between two divergent genes, *TTC23* and *LRRC28*, with the former showing aberrant splicing that relates to therapy resistance in prostate cancer cells (*53*); 4) The promoter of *GNAI2*, a protein that interacts with *CXCR5*, which positively correlates with prostate cancer progression (*54*); 5) A promoter between two divergent genes, *PRICKLE4* and *FRS3*, with the latter thought to affect malignant but not benign prostate cells (*55*). We cloned the promoter sequence with and without the neomer into a luciferase promoter assay vector and compared their activity in androgen-sensitive human prostate adenocarcinoma cells (LNCaP). For two out of the five assayed promoters, we observed a significant effect on reporter activity (**Fig. 4C**). For the *RPS2-SNHG9* promoter, the neomer led to significantly increased activity in line with this gene being

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

overexpressed in cancer (*51*). For the *TMEM127-CIAO1* promoter the neomer completely abolished activity, fitting with its observed downregulation in prostate cancer (*52*). Combined, our experimental results show that neomers could have a significant effect on promoter activity and could potentially be used to identify cancer associated *cis*-regulatory mutations.

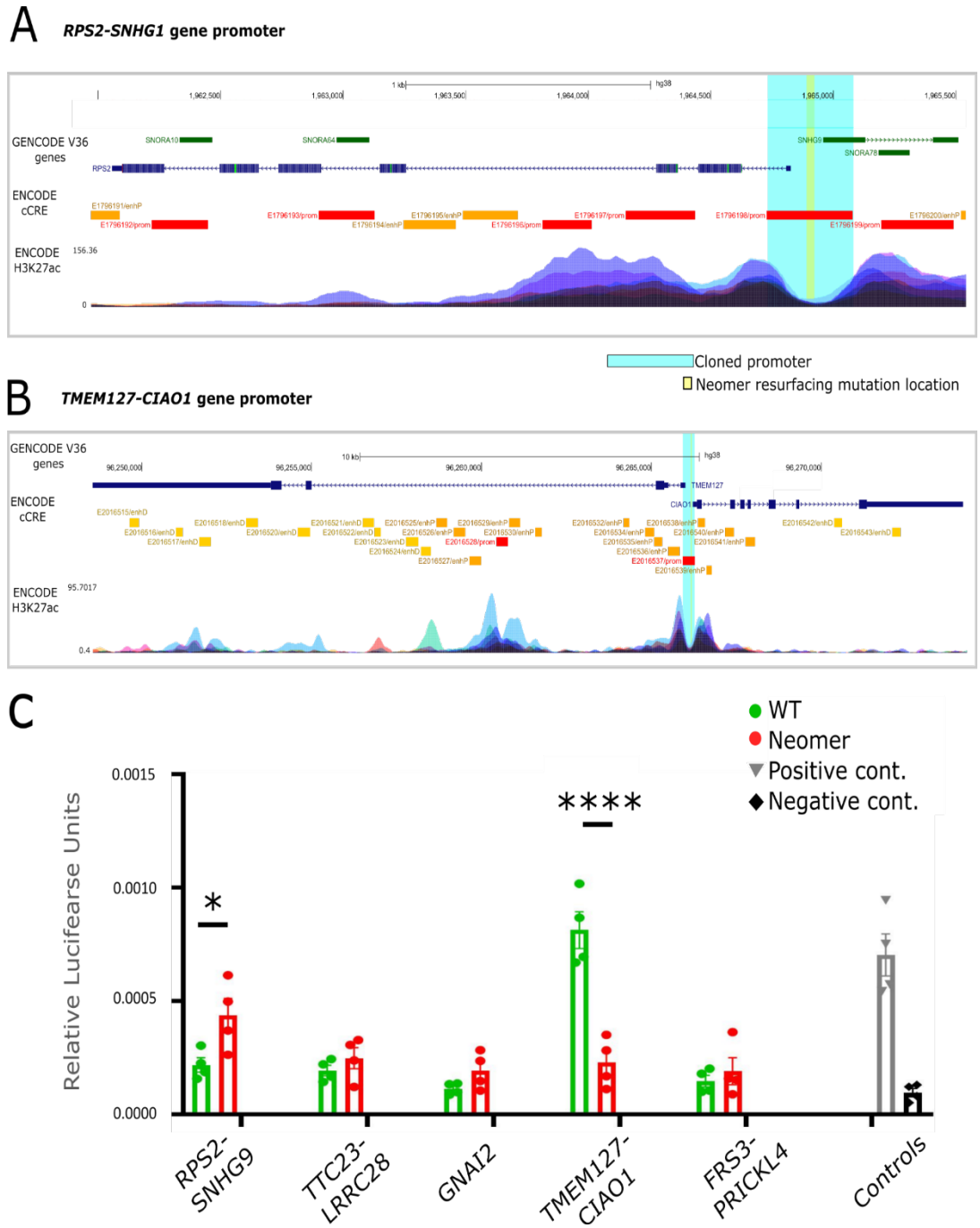

Figure 4. Neomer resurfacing mutations in gene promoters affect gene expression. (**A**-**B**) UCSC browser track snapshot of the location of neomer resurfacing mutation in prostate cancer RPS2-SNHG9 (**A**) and TMEM127-CIAO1 (**B**). Presented are genebody location (GENCODE V28), regulatory element according to ENCODE cCRE (yellow = enhancer, red = promoter), and the ENCODE layered epigenetic enhancer mark H3K27ac. (**C**) Relative luciferase (LUC) units from dual-luciferase reporter assay of 5 constructs containing either WT or nullomer resurfacing SNP in the promoter area driving LUC expression. All RLU are compared to Renila expression. Significance is calculated using two-way ANOVA with multiple testing wih Šidák correction.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

DISCUSSION

Cancer is a DNA mutation causing disease. Here, we show that by analyzing cancer WGS datasets we can find mutations that lead to the generation of neomers, short sequences normally not present in the genome but present in multiple cancer genomes. Further analyses of these neomers shows that they can be used to classify not only cancer tissue origin but also additional cancer features, such as MSI or POLE deficiency with high accuracy. Analysis of cfDNA WGS datasets finds that neomers could be used to distinguish patients from healthy individuals. Finally, using reporter assays, we demonstrate that neomers have a functional effect on regulatory sequences.

Our analyses used 2,577 patients with 21 different cancerous tissues to develop a cancer tissue of origin classifier. Overall, we observed different recurring neomers in each cancer type, other than esophagus and stomach, likely because these cancers share very similar As it is based on neomer detection in sequencing datasets, read mapping is not needed, requiring only a single pass over the data, thus making it extremely effective from a computational standpoint. Additional genomes from tumor tissues, controls and cfDNA could improve this classifier. In general, the classifier has better performance for cancer types with more patients and high mutation burden (**Fig 1,2**). Thus, obtaining WGS datasets from tumor, matched control and cfDNA from myeloid, thyroid and prostate cancers would be extremely helpful in improving the ability of our neomer classifier to detect these cancer types. In addition, our analyses showed that other than tissue origin, neomers can also be used to detect other cancer features. It would be interesting to test whether neomers could diagnose additional tumor features and also detect other cancer characteristics such as chance of recurrence, drug response, mortality and others.

Neomers could also be combined with other cancer biomarkers and risk factors to improve the diagnostic positive predictive value. For example, it was recently shown that combining a blood test that detects both protein biomarkers and DNA mutations along with positron emission tomography-computed tomography (PET-CT) could detect multiple cancers (*56*). Adding neomers to known cancer-associated coding mutations in the screening of cfDNA could increase sensitivity and specificity. In summary, adding neomer-based diagnostics to existing cancer biomarkers and risk factors could improve the power to detect various cancer subtypes.

As nullomers/neomers do not exist in the human genome they could also be exceptional candidates for neoantigens, to be targeted via immunotherapy. Previous work has shown that minimal absent words, short sequences that are absent from a genome or proteome, could be used to identify phosphorylation sites of high confidence, some of which could be associated with cancer (*57*). Nullomers were also shown to be effective in identifying unique peptides that are exceedingly distant from human peptides that potentially could be used as antibodies against Trypanosoma cruzi (*58*) or SARS-CoV-2 (*59*). Analysis of the Immune Epitope Database of validated antigens (*60*) found that 13 of the recurrent coding neomers can create neoantigens with predicted strong binding levels that were subsequently validated (**Table S2**). From the 1,700 neoantigens with strong binding levels, only 1.72

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

(p-value<1e-8, hypergeometric test) is expected to correspond to a neomer, suggesting that missense mutations also resulting in neomers are 7-fold more likely to also generate strongly binding neoantigens.

We used a sequence enrichment based assay to detect nullomers in cfDNA from blood taken from prostate cancer cases and controls. Alternate assays could potentially be used for future rapid diagnosis of cancer via nullomers. These could include the use of CRISPR-based detection tools that utilize Cas12 or Cas13 (*61*). For example, recent use of Cas13 in a microwell array system allowed the rapid screening of over 4,500 targets for 169 human-associated viruses with high sensitivity and specificity (*62*). In addition, with nullomer-based diagnostics potentially not needing large amounts of starting material, cfDNA could be collected from urine or saliva, which were shown to be a viable but reduced source of cfDNA (*63*, *64*).

Neomers could be used as a novel tool to identify cancer-associated gene regulatory mutations. Amongst the 210 prostate cancer promoter neomers, we selected five promoters and found that two of them significantly affected promoter activity due to the neomer. Their difference in activity was in line with the gene's expression change in prostate cancer, with *RPS2-SNHG9* having increased activity fitting with its overexpression in prostate cancer (*51*) and *TMEM127-CIAO1* abolishing activity, in line with its observed downregulation in prostate cancer (*52*). Although these promoters were selected based on their prostate cancer association, future high-throughput assays, such as massively parallel reporter assays (MPRAs; (*65*)) that can test thousands of sequences and variants for their regulatory activity, could be used to test the effect of neomers on gene regulation in an unbiased manner.

In summary, we show that neomers can provide a powerful tool for cancer diagnosis. As they can easily be detected via sequence or CRISPR-based tools, it should be straightforward to integrate them in current routine cancer diagnostic tests and their use could increase the sensitivity and specificity of these tests. Combining neomer-based screening with clinical characteristics and additional diagnostic tools/features could increase the positive predictive value. In addition, as cfDNA could also be isolated from urine and saliva and detection of these sequences only requires a relatively small amount of DNA, neomer-based diagnosis could be carried out in a non-invasive manner. Our work also suggests that neomers could be used to highlight cancer-associated gene regulatory mutations which have been difficult to identify. Further high-throughput characterization of these mutations could allow the detection of bona fide cancer-associated functional regulatory mutations that could be used for diagnosis and treatment.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

METHODS

Computational characterization of nullomers

The GRCh38 reference assembly of the human genome was used throughout the study. Nullomer extraction was performed for kmer lengths up to 17 base pairs using the algorithm described in (*18*). By definition, the reverse complement of a nullomer will also be a nullomer. Throughout this manuscript when counting nullomers, the reverse complement of nullomer *i* was also considered separately, unless *i* is a palindrome.

Substitutions and indels identified from whole genome sequencing (WGS) of tumor samples from 2,577 individuals across 21 tissues were obtained from https://dcc.icgc.org/releases/PCAWG/ (*21*). Nullomer extraction was performed for kmer lengths up to 17 base pairs using the same algorithm described in (*18*). Recurrent nullomers (neomers) (r_i) were annotated as those that resulted from substitutions or indels across two or more patients within a cancer type. When possible, r_i was chosen to get ~10,000 neomers from each tissue, otherwise it was set to 2 (**Table S1**).

Classification of tumor tissue of origin using neomers

We trained a classifier to distinguish tissue of origin for a cancer sample based on observed neomers using the libSVM package to train a support vector machine classifier with a linear kernel. We used 10-fold cross validation whereby the classifier was evaluated on a held out fraction of the data. The set of neomers for each cancer type was recalculated for each round to only include the patients in the training set.

Supervised selection of neomers

The MMR status of each biopsy sample was derived from (Zou et al. 2021). The model was trained on neomers identified in MSI samples and the performance of the algorithm evaluated. For the MSI versus the MSS samples, we counted the number of neomers that contained either AAAAAAAA or TTTTTTTT repeats. The threshold for determining MSI or MSS was set as the harmonic mean of the maximum number of counts in the MSS set and the minimum number of counts in the MSI set. The POLE deficiency status of each biopsy sample was derived from (Zou et al. 2021) and we used a similar strategy to that of MMR status, but we instead counted neomers created through either a TCT>TAT or TCG>TTG mutation. Since the number of patients in each category was limited, we only used 5-fold cross validation.

Comparison to validated neoantigens

We downloaded a list of 1,967 validated neoantigens from http://biopharm.zju.edu.cn/download.neoantigen/iedb_validated.zip. Requiring both predicted strong binding and a positive validation, left us with 1,700 neoantigens. To evaluate the enrichment of neoantigens corresponding to neomers, we assumed a hypergeometric distribution with 1,700 draws

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

from an urn with 188,659 white balls (total number of neomers) and 186,067,892 black balls (number of nullomers found with lower recurrency than what was required for neomers).

Neomer identification in ctDNA samples

ctDNA samples were derived from (*46*). To filter out common population variants, we obtained variant information from the gnomAD v2 (*47*) and annotated all neomers that were generated due to these variants. Variants that were not single base pair substitutions were not considered. The FASTQ files were scanned for neomers by searching for exact matches to the 16-mers of interest. In addition, the quality scores of bases matching a neomer were evaluated to make sure that none of them fell below a threshold.

We developed a Poisson model where the expected number of neomers of type i is given by $CaeN_i/3$, where C is the coverage, a is the allele frequency, e is the error rate, N_i is the number of loci where a substitution could result in the creation of neomer i, and the division by 3 is to account for the fact that only one of three substitutions will create the neomer. For the analysis of the WGS data we used e=0.015 and a=0.006 while C was calculated from the reads and N_i is derived from the reference genome. Moreover, we excluded neomers that were present at a level that would be expected if a>0.25 to further remove neomers detected due to germline variants or experimental artefacts.

When processing the data, reads were not mapped to the genome, which means that the coverage is based on the total length of the sequenced reads rather than on the mappable reads. Each read is searched for neomers and neomers containing a base with quality <10 are excluded. The permissive threshold was chosen since otherwise all of the reads from the control samples would have been excluded.

Promoter luciferase assays

Promoter sequences with and without the neomer (**Table S3**) were synthetically generated and cloned into the modified Promega promoter assay luciferase vector pGL4.11b (a gift from Dr. Rick Myers, HudsonAlpha) by BioMatik Inc and Sanger sequence verified. LNCaP cells were plated at an initial density of 2*10^5 cells/well in 24-well tissue culture plates and maintained in RPMI medium, 10% FBS supplemented with L-Glutamine and Penicillin/Streptomycin. Plasmids together with a renilla expressing plasmid, pGL4.74 (Promega), at a ratio of 10:1 luciferase:renilla were transfected using the X-tremeGENE™ HP DNA Transfection Reagent (Roche) using 1:4 ratio of DNA (ug) to reagent (ul). 72 hours post transfection luciferase and renilla levels were measured using the Dual-Luciferase Reporter Assay System (Promega) following the manufacturer's protocol using a GloMax Explorer Multimode Microplate Reader (Promega). Luciferase activity was normalized to renilla levels and presented as Relative Luciferase Units (RLU). Statistical analysis was performed using Prism version 9.0.2 (GraphPad). All values were reported as means (AVG) and standard errors (SE). p values < 0.05 were considered statistically significant.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Conflicts of interest

I.G.S, O.Y.B, I.M, M.H and N.A are co-founders and equity holders of Neomer Diagnostics. Certain authors of this manuscript have filed a patent application covering embodiments and concepts disclosed in the manuscript.

Acknowledgments

We want to first and foremost thank the individuals who participated in this study. We also want to thank Dr. Felix Fang at UCSF for his advice and support. This work was supported in part by the Benioff Initiative for Prostate Cancer Research. MH was supported by core funding from the Wellcome Trust and core funding from the Evergrande Center.

Author contributions

I.G.S., O.Y.B., I.M., M.H. and N.A. conceived the study. I.G.S., I.M., R.E., M.H. wrote the code, I.G.S., O.Y.B., I.M., M.H. and N.A. performed the analyses and generated the visualizations, O.Y.B, R.B., C.C and E.C. carried out experimental assays, E.C. and J.S.W. provided cfDNA samples, M.H. and N.A. provided resources and supervised the research, I.G.S., O.Y.B., I.M., M.H. and N.A. wrote the manuscript.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

References

1. Cancer, (available at https://www.who.int/news-room/fact-sheets/detail/cancer).

2. R. L. Siegel, K. D. Miller, A. Jemal, Cancer statistics, 2020. *CA Cancer J. Clin.* **70**, 7–30 (2020).

3. N. Hawkes, Cancer survival data emphasise importance of early diagnosis. *BMJ*. **364** (2019), doi:10.1136/bmj.l408.

4. R. Etzioni, N. Urban, S. Ramsey, M. McIntosh, S. Schwartz, B. Reid, J. Radich, G. Anderson, L. Hartwell, The case for early detection. *Nat. Rev. Cancer*. **3**, 243–252 (2003).

5. Cancer, (available at https://www.who.int/cancer/detection/en/).

6. A. J. Bronkhorst, V. Ungerer, S. Holdenrieder, The emerging role of cell-free DNA as a molecular marker for cancer management. *Biomol Detect Quantif*. **17**, 100087 (2019).

7. E. Heitzer, L. Auinger, M. R. Speicher, Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living. *Trends Mol. Med.* **26**, 519–528 (2020).

8. O. A. Zill, K. C. Banks, S. R. Fairclough, S. A. Mortimer, The landscape of actionable genomic alterations in cell-free circulating tumor DNA from 21,807 advanced cancer patients. *Clin. Cancer Res.* (2018) (available at https://clincancerres.aacrjournals.org/content/24/15/3528.abstract).

9. S. Saghafinia, M. Mina, N. Riggi, D. Hanahan, G. Ciriello, Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors. *Cell Rep.* **25**, 1066–1080.e8 (2018).

10. R. Sadeh, I. Sharkia, G. Fialkoff, A. Rahat, J. Gutin, A. Chappleboim, M. Nitzan, I. Fox-Fisher, D. Neiman, G. Meler, Z. Kamari, D. Yaish, T. Peretz, A. Hubert, J. E. Cohen, A. Salah, M. Temper, A. Grinshpun, M. Maoz, S. Abu-Gazala, A. Ben Ya'acov, E. Shteyer, R. Safadi, T. Kaplan, R. Shemer, D. Planer, E. Galun, B. Glaser, A. Zick, Y. Dor, N. Friedman, ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin. *Nat. Biotechnol.* (2021), doi:10.1038/s41587-020-00775-6.

11. G. Barbany, C. Arthur, A. Lieden, M. Nordenskjöld, R. Rosenquist, B. Tesi, K. Wallander, E. Tham, Cell-free tumour DNA testing for early detection of cancer--a potential future tool. *J. Intern. Med.* **286**, 118–136 (2019).

12. P. Razavi, B. T. Li, D. N. Brown, B. Jung, E. Hubbell, R. Shen, W. Abida, K. Juluru, I. De Bruijn, C. Hou, O. Venn, R. Lim, A. Anand, T. Maddala, S. Gnerre, R. Vijaya Satya, Q. Liu, L. Shen, N. Eattock, J. Yue, A. W. Blocker, M. Lee, A. Sehnert, H. Xu, M. P. Hall, A. Santiago-Zayas, W. F. Novotny, J. M. Isbell, V. W. Rusch, G. Plitas, A. S. Heerdt, M. Ladanyi, D. M. Hyman, D. R. Jones, M. Morrow, G. J. Riely, H. I. Scher, C. M. Rudin, M. E. Robson, L. A. Diaz Jr, D. B. Solit, A. M. Aravanis, J. S. Reis-Filho, High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nat. Med.* **25**, 1928–1937 (2019).

13. L. Ji, T. Sasaki, X. Sun, P. Ma, Z. A. Lewis, R. J. Schmitz, Methylated DNA is over-represented in whole-genome bisulfite sequencing data. *Front. Genet.* **5**, 341 (2014).

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

14. M.-B. Worm Ørntoft, Review of Blood-Based Colorectal Cancer Screening: How Far Are Circulating Cell-Free DNA Methylation Markers From Clinical Implementation? *Clin. Colorectal Cancer*. **17**, e415–e433 (2018).

15. K. Warton, G. Samimi, Methylation of cell-free circulating DNA in the diagnosis of cancer. *Front Mol Biosci*. **2**, 13 (2015).

16. G. Hampikian, T. Andersen, in *Biocomputing 2007* (WORLD SCIENTIFIC, 2006), pp. 355–366.

17. D. Vergni, D. Santoni, Nullomers and High Order Nullomers in Genomic Sequences. *PLoS One*. **11**, e0164540 (2016).

18. I. Georgakopoulos-Soares, O. Y. Barnea, I. Mouratidis, M. Hemberg, N. Ahituv, Absent from DNA and protein: genomic characterization of nullomers and nullpeptides across functional categories and evolution. *Cold Spring Harbor Laboratory* (2020), p. 2020.03.02.972422.

19. The Cancer Genome Atlas Program (2018), (available at https://www.cancer.gov/tcga).

20. W. Jiao, G. Atwal, P. Polak, R. Karlic, E. Cuppen, PCAWG Tumor Subtypes and Clinical Translation Working Group, A. Danyi, J. de Ridder, C. van Herpen, M. P. Lolkema, N. Steeghs, G. Getz, Q. Morris, L. D. Stein, PCAWG Consortium, A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* **11**, 728 (2020).

21. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes. *Nature*. **578**, 82–93 (2020).

22. I. A. Prior, P. D. Lewis, C. Mattos, A comprehensive survey of Ras mutations in cancer. *Cancer Res.* **72**, 2457–2467 (2012).

23. C. Muñoz-Maldonado, Y. Zimmer, M. Medová, A Comparative Analysis of Individual RAS Mutations in Cancer Biology. *Front. Oncol.* **9**, 1088 (2019).

24. J. Vinagre, A. Almeida, H. Populo, R. Batista, J. Lyra, V. Pinto, R. Coelho, R. Celestino, H. Prazeres, L. Lima, M. Melo, A. G. da Rocha, A. Preto, P. Castro, L. Castro, F. Pardal, J. M. Lopes, L. L. Santos, R. M. Reis, J. Cameselle-Teijeiro, M. Sobrinho-Simoes, J. Lima, V. Maximo, P. Soares, Frequency of TERT promoter mutations in human cancers. *Nat. Commun.* **4**, 2185 (2013).

25. B. Heidenreich, P. S. Rachakonda, K. Hemminki, R. Kumar, TERT promoter mutations in cancer development. *Curr. Opin. Genet. Dev.* **24**, 30–37 (2014).

26. J. H. Song, H.-J. Kang, L. A. Luevano, V. Gokhale, K. Wu, R. Pandey, H.-H. Sherry Chow, L. H. Hurley, A. S. Kraft, Small-Molecule-Targeting Hairpin Loop of hTERT Promoter G-Quadruplex Induces Cancer Cell Death. *Cell Chem Biol*. **26**, 1110–1121.e4 (2019).

27. R. J. A. Bell, H. T. Rube, A. Kreig, A. Mancini, S. D. Fouse, R. P. Nagarajan, S. Choi, C. Hong, D. He, M. Pekmezci, J. K. Wiencke, M. R. Wrensch, S. M. Chang, K. M. Walsh, S. Myong, J. S.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Song, J. F. Costello, Cancer. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. *Science*. **348**, 1036–1039 (2015).

28. B. Powter, S. A. Jeffreys, H. Sareen, A. Cooper, D. Brungs, J. Po, T. Roberts, E.-S. Koh, K. F. Scott, M. Sajinovic, J. Y. Vessey, P. de Souza, T. M. Becker, Human TERT promoter mutations as a prognostic biomarker in glioma. *J. Cancer Res. Clin. Oncol.* **147**, 1007–1017 (2021).

29. J. M. Romero, B. Grünwald, G.-H. Jang, P. P. Bavi, A. Jhaveri, M. Masoomian, S. E. Fischer, A. Zhang, R. E. Denroche, I. M. Lungu, A. De Luca, J. M. S. Bartlett, J. Xu, N. Li, S. Dhaliwal, S.-B. Liang, D. Chadwick, F. Vyas, P. Bronsert, R. Khokha, T. L. McGaha, F. Notta, P. S. Ohashi, S. J. Done, G. M. O'Kane, J. M. Wilson, J. J. Knox, A. Connor, Y. Wang, G. Zogopoulos, S. Gallinger, A Four-Chemokine Signature Is Associated with a T-cell-Inflamed Phenotype in Primary and Metastatic Pancreatic Cancer. *Clin. Cancer Res.* **26**, 1997–2010 (2020).

30. C. E. Antal, A. M. Hudson, E. Kang, C. Zanca, C. Wirth, N. L. Stephenson, E. W. Trotter, L. L. Gallegos, C. J. Miller, F. B. Furnari, T. Hunter, J. Brognard, A. C. Newton, Cancer-Associated Protein Kinase C Mutations Reveal Kinase's Role as Tumor Suppressor. *Cell*. **160** (2015), pp. 489–502.

31. A. Vincent, N. Omura, S. M. Hong, A. Jaffe, J. Eshleman, Genome-wide analysis of promoter methylation associated with gene expression profile in pancreatic adenocarcinoma. *Clin. Cancer Res.* (2011) (available at https://clincancerres.aacrjournals.org/content/17/13/4341.short).

32. E. Xu, J. Gu, E. T. Hawk, K. K. Wang, M. Lai, M. Huang, J. Ajani, X. Wu, Genome-wide methylation analysis shows similar patterns in Barrett's esophagus and esophageal adenocarcinoma. *Carcinogenesis*. **34**, 2750–2756 (2013).

33. J. M. Leers, S. R. DeMeester, N. Chan, S. Ayazi, A. Oezcelik, E. Abate, F. Banki, J. C. Lipham, J. A. Hagen, T. R. DeMeester, *J. Thorac. Cardiovasc. Surg.*, in press.

34. F. Battaglin, M. Naseem, H.-J. Lenz, M. E. Salem, Microsatellite instability in colorectal cancer: overview of its clinical significance and novel perspectives. *Clin. Adv. Hematol. Oncol.* **16**, 735–745 (2018).

35. S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, D. C. Wedge, P. Van Loo, Y. S. Ju, M. Smid, A. B. Brinkman, S. Morganella, M. R. Aure, O. C. Lingjærde, A. Langerød, M. Ringnér, S.-M. Ahn, S. Boyault, J. E. Brock, A. Broeks, A. Butler, C. Desmedt, L. Dirix, S. Dronov, A. Fatima, J. A. Foekens, M. Gerstung, G. K. J. Hooijer, S. J. Jang, D. R. Jones, H.-Y. Kim, T. A. King, S. Krishnamurthy, H. J. Lee, J.-Y. Lee, Y. Li, S. McLaren, A. Menzies, V. Mustonen, S. O'Meara, I. Pauporté, X. Pivot, C. A. Purdie, K. Raine, K. Ramakrishnan, F. G. Rodríguez-González, G. Romieu, A. M. Sieuwerts, P. T. Simpson, R. Shepherd, L. Stebbings, O. A. Stefansson, J. Teague, S. Tommasi, I. Treilleux, G. G. Van den Eynden, P. Vermeulen, A. Vincent-Salomon, L. Yates, C. Caldas, L. van't Veer, A. Tutt, S. Knappskog, B. K. T. Tan, J. Jonkers, Å. Borg, N. T. Ueno, C. Sotiriou, A. Viari, P. A. Futreal, P. J. Campbell, P. N. Span, S. Van Laere, S. R. Lakhani, J. E. Eyfjord, A. M. Thompson, E. Birney, H. G. Stunnenberg, M. J. van de Vijver, J. W. M. Martens, A.-L. Børresen-Dale, A. L. Richardson, G. Kong, G. Thomas, M. R. Stratton, Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*. **534**, 47–54 (2016).

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

36. B. Garmezy, J. S. Gheeya, K. Z. Thein, P. G. Pilie, W. Wang, J. Rodon Ahnert, K. R. Shaw, F. Meric-Bernstam, T. A. Yap, Correlation of pathogenic POLE mutations with clinical benefit to immune checkpoint inhibitor therapy. *J. Clin. Orthod.* **38**, 3008–3008 (2020).

37. F. Wang, Q. Zhao, Y.-N. Wang, Y. Jin, M.-M. He, Z.-X. Liu, R.-H. Xu, Evaluation of POLE and POLD1 Mutations as Biomarkers for Immunotherapy Outcomes Across Multiple Cancer Types. *JAMA Oncol*. **5**, 1504–1506 (2019).

38. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, M. R. Stratton, Signatures of mutational processes in human cancer. *Nature*. **500**, 415–421 (2013).

39. P. Rawla, Epidemiology of Prostate Cancer. *World J. Oncol.* **10**, 63–89 (2019).

40. M. J. Barry, Prostate-Specific–Antigen Testing for Early Diagnosis of Prostate Cancer. *New England Journal of Medicine*. **344** (2001), pp. 1373–1377.

41. F. C. Cackowski, R. S. Taichman, Minimal Residual Disease in Prostate Cancer. *Advances in Experimental Medicine and Biology* (2018), pp. 47–53.

42. N. P. Murray, Minimal residual disease in prostate cancer patients after primary treatment: theoretical considerations, evidence and possible use in clinical management. *Biol. Res.* **51**, 32 (2018).

43. S. T. Hennigan, S. Y. Trostel, N. T. Terrigino, O. S. Voznesensky, R. J. Schaefer, N. C. Whitlock, S. Wilkinson, N. V. Carrabba, R. Atway, S. Shema, R. Lake, A. R. Sweet, D. J. Einstein, F. Karzai, J. L. Gulley, P. Chang, G. J. Bubley, S. P. Balk, H. Ye, A. G. Sowalsky, Low Abundance of Circulating Tumor DNA in Localized Prostate Cancer. *JCO Precis Oncol*. **3** (2019), doi:10.1200/PO.19.00176.

44. M. T. Bjerre, M. Nørgaard, O. H. Larsen, S. Ø. Jensen, S. H. Strand, P. Østergren, M. Fode, J. Fredsøe, B. P. Ulhøi, M. M. Mortensen, J. B. Jensen, M. Borre, K. D. Sørensen, Epigenetic Analysis of Circulating Tumor DNA in Localized and Metastatic Prostate Cancer: Evaluation of Clinical Biomarker Potential. *Cells*. **9** (2020), doi:10.3390/cells9061362.

45. M. C. Maia, M. Salgia, S. K. Pal, Harnessing cell-free DNA: plasma circulating tumour DNA for liquid biopsy in genitourinary cancers. *Nat. Rev. Urol.* **17**, 271–291 (2020).

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

46. P. Ulz, S. Perakis, Q. Zhou, T. Moser, J. Belic, I. Lazzeri, A. Wölfler, A. Zebisch, A. Gerger, G. Pristauz, E. Petru, B. White, C. E. S. Roberts, J. S. John, M. G. Schimek, J. B. Geigl, T. Bauernhofer, H. Sill, C. Bock, E. Heitzer, M. R. Speicher, Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nat. Commun.* **10**, 4666 (2019).

47. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Genome Aggregation Database Consortium, B. M. Neale, M. J. Daly, D. G. MacArthur, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. **581**, 434–443 (2020).

48. R. C. Poulos, M. A. Sloane, L. B. Hesson, J. W. H. Wong, The search for cis-regulatory driver mutations in cancer genomes. *Oncotarget*. **6**, 32509–32525 (2015).

49. K. Elliott, E. Larsson, Non-coding driver mutations in human cancer. *Nat. Rev. Cancer* (2021), doi:10.1038/s41568-021-00371-z.

50. Consortium, Encode Project, I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kelllis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassman, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, M. Snyder, M. J. Pazin, R. F. Lowdon, L. A. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E. Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elinitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guigo, R. C. Hardison, T. J. Hubbard, M. Kellis, W. J. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Starnatoyannopoulos, S. A. Tennebaum, Z. Weng, K. P. White, B. Wold, J. Khatun, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, M. C. Giddings, B. E. Bernstein, C. B. Epstein, N. Shoresh, J. Ernst, P. Kheradpour, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, L. D. Ward, R. C. Altshuler, M. L. Eaton, M. Kellis, S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Gao, D. Gonzalez, A. Gordon, H. P. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, B. A. Risk, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandu, L. Schaeffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, Y. Hayashizaki, J. Harrow, M. Gerstein, T. J. Hubbard, A. Reymond, S. E. Antonarakis, G. J. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, T. R. Gingeras, K. R. Rosenbloom, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, W. J. Kent, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, T. S. Furey, L. Song, L. L. Grasfeder, P. G. Giresi, B. K. Lee, A. Battenhouse, N. C. Sheffield, J. M. Simon, K. A. Showers, A. Safi, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, E. Birney, V. R. Iyer, J. D. Lieb, G. E. Crawford, G. Li, K. S. Sandhu, M. Zheng, P. Wang, O. J. Luo, A. Shahab, M. J. Fullwood, X. Ruan, Y. Ruan, R. M. Myers, F. Pauli, B. A. Williams, J. Gertz, G. K. Marinov, T. E. Reddy, J. Vielmetter, E. C. Partridge, D. Trout, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, B. King, M. A. Muratet, I. Antoshechkin, K. M. Newberry, K. McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. L. Parker, S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, C. Gunter, J. S. Newberry, S. E. Levy, D. M. Absher, A. Mortazavi, W. H. Wong, B. Wold, M. J. Blow, A. Visel, L. A. Pennachio, L. Elnitski, E. H. Margulies, S. C. Parker, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, J. Chrast, C. Davidson, T. Derrien, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix, C. Howald, T. Hunt, I. Jungreis, M. Kay, E. Khurana, F. Kokocinski, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, A. Tanzer, E. Tapanan, M. L. Tress, M. J. van Baren, N. Walters, S. Washieti, L. Wilming, A. Zadissa, Z. Zhengdong, M. Brent, D. Haussler, M. Kellis, A. Valencia, M. Gerstein, A. Raymond, R. Guigo, J. Harrow, T. J. Hubbard, S. G. Landt, S. Frietze, A. Abyzov, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyenger, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Larnarre-Vincent, J. Leng, J. Lian, M. Lindahl-Allen, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen, Z. Ouyang, D. Patacsil, B. Pei, D. Raha, L. Ramirez, B. Reed, J. Rozowsky, A. Sboner, M. Shi, C. Sisu, T. Slifer, H. Witt, L. Wu, X. Xu, K. K. Yan, X. Yang, K. Y. Yip, Z. Zhang, K. Struhl, S. M. Weissman, M. Gerstein, P. J. Farnham, M. Snyder, S. A. Tenebaum, L. O. Penalva, F. Doyle, S. Karmakar, S. G. Landt, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, D. Patacsil, T. Slifer, A. Victorsen, X. Yang, M. Snyder, K. P. White, T. Auer, L. Centarin, M. Eichenlaub, F. Gruhl, S. Heerman, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, Z. Weng, T. W. Whitfield, J. Wang, P. J. Collins, S. F. Aldred, N. D. Trinklein, E. C. Partridge, R. M. Myers, J. Dekker, G. Jain, B. R. Lajoie, A. Sanyal, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. S. Hansen, L. Boatman, E. Haugen, R. Humbert, G. Jain, A. K. Johnson, E. M. Johnson, T. M. Kutyavin, B. R. Lajoie, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, E. Rynes, P. Sabo, M. E. Sanchez, R. S.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Sandstrom, A. Sanyal, A. O. Shafer, A. B. Stergachis, S. Thomas, R. E. Thurman, B. Vernot, J. Vierstra, S. Vong, H. Wang, M. A. Weaver, Y. Yan, M. Zhang, J. A. Akey, M. Bender, M. O. Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, R. Kaul, J. Dekker, J. A. Stamatoyannopoulos, I. Dunham, K. Beal, A. Brazma, P. Flicek, J. Herrero, N. Johnson, D. Keefe, M. Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. P. Wilder, S. Batzoglou, A. Sidow, N. Hussami, S. Kyriazopoulou-Panagiotopoulou, M. W. Libbrecht, M. A. Schaub, A. Kundaje, R. C. Hardison, W. Miller, B. Giardine, R. S. Harris, W. Wu, P. J. Bickel, B. Banfai, N. P. Boley, J. B. Brown, H. Huang, Q. Li, J. J. Li, W. S. Noble, J. A. Bilmes, O. J. Buske, M. M. Hoffman, A. O. Sahu, P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, Z. Weng, S. Iyer, X. Dong, M. Greven, X. Lin, J. Wang, H. S. Xi, J. Zhuang, M. Gerstein, R. P. Alexander, S. Balasubramanian, C. Cheng, A. Harmanci, L. Lochovsky, R. Min, X. J. Mu, J. Rozowsky, K. K. Yan, K. Y. Yip, E. Birney, An integrated encyclopedia of DNA elements in the human genome. *Nature*. **489**, 57–74 (2012).

51. A. Ohkia, Y. Hu, M. Wang, F. U. Garcia, M. E. Stearns, Evidence for prostate cancer-associated diagnostic marker-1: immunohistochemistry and in situ hybridization studies. *Clin. Cancer Res.* **10**, 2452–2458 (2004).

52. Y. Qin, Y. Deng, C. J. Ricketts, S. Srikantan, E. Wang, E. R. Maher, P. L. M. Dahia, The tumor susceptibility gene TMEM127 is mutated in renal cell carcinomas and modulates endolysosomal function. *Hum. Mol. Genet.* **23**, 2428–2439 (2014).

53. E. Bowler, S. Porazinski, S. Uzor, P. Thibault, M. Durand, E. Lapointe, K. M. A. Rouschop, J. Hancock, I. Wilson, M. Ladomery, Hypoxia leads to significant changes in alternative splicing and elevated expression of CLK splice factor kinases in PC3 prostate cancer cells. *BMC Cancer*. **18**, 355 (2018).

54. C. P. El-Haibi, P. Sharma, R. Singh, P. Gupta, D. D. Taub, S. Singh, J. W. Lillard Jr, Differential G protein subunit expression by prostate cancer cells and their interaction with CXCR5. *Mol. Cancer*. **12**, 64 (2013).

55. T. Valencia, A. Joseph, N. Kachroo, S. Darby, S. Meakin, V. J. Gnanapragasam, Role and expression of FRS2 and FRS3 in prostate cancer. *BMC Cancer*. **11**, 484 (2011).

56. A. M. Lennon, A. H. Buchanan, I. Kinde, A. Warren, A. Honushefsky, A. T. Cohain, D. H. Ledbetter, F. Sanfilippo, K. Sheridan, D. Rosica, C. S. Adonizio, H. J. Hwang, K. Lahouel, J. D. Cohen, C. Douville, A. A. Patel, L. N. Hagmann, D. D. Rolston, N. Malani, S. Zhou, C. Bettegowda, D. L. Diehl, B. Urban, C. D. Still, L. Kann, J. I. Woods, Z. M. Salvati, J. Vadakara, R. Leeming, P. Bhattacharya, C. Walter, A. Parker, C. Lengauer, A. Klein, C. Tomasetti, E. K. Fishman, R. H. Hruban, K. W. Kinzler, B. Vogelstein, N. Papadopoulos, Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. *Science*. **369** (2020), doi:10.1126/science.abb9601.

57. G. Koulouras, M. C. Frith, Significant non-existence of sequences in genomes and proteomes. *Nucleic Acids Res.* (2021), doi:10.1093/nar/gkab139.

58. D. Vergni, R. Gaudio, D. Santoni, The farther the better: Investigating how distance from human self affects the propensity of a peptide to be presented on cell surface by MHC class I molecules,

medRxiv preprint doi: https://doi.org/10.1101/2021.08.15.21261805; this version posted August 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

the case of Trypanosoma cruzi. *PLoS One*. **15**, e0243285 (2020).

59. D. Santoni, D. Vergni, In the search of potential epitopes for Wuhan seafood market pneumonia virus using high order nullomers. *J. Immunol. Methods*. **481-482**, 112787 (2020).

60. R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, B. Peters, The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).

61. M. J. Kellner, J. G. Koob, J. S. Gootenberg, O. O. Abudayyeh, F. Zhang, SHERLOCK: nucleic acid detection with CRISPR nucleases. *Nat. Protoc.* **14**, 2986–3012 (2019).

62. C. M. Ackerman, C. Myhrvold, S. G. Thakku, C. A. Freije, H. C. Metsky, D. K. Yang, S. H. Ye, C. K. Boehm, T.-S. F. Kosoko-Thoroddsen, J. Kehe, T. G. Nguyen, A. Carter, A. Kulesa, J. R. Barnes, V. G. Dugan, D. T. Hung, P. C. Blainey, P. C. Sabeti, Massively multiplexed nucleic acid detection with Cas13. *Nature*. **582**, 277–282 (2020).

63. E. Augustus, K. Van Casteren, L. Sorber, P. van Dam, G. Roeyen, M. Peeters, A. Vorsters, A. Wouters, J. Raskin, C. Rolfo, K. Zwaenepoel, P. Pauwels, The art of obtaining a high yield of cell-free DNA from urine. *PLoS One*. **15**, e0231058 (2020).

64. S. Ding, X. Song, X. Geng, L. Liu, H. Ma, X. Wang, L. Wei, L. Xie, X. Song, Saliva-derived cfDNA is applicable for EGFR mutation detection but not for quantitation analysis in non-small cell lung cancer. *Thorac Cancer*. **10**, 1973–1983 (2019).

65. F. Inoue, N. Ahituv, Decoding enhancers using massively parallel reporter assays. *Genomics*. **106**, 159–164 (2015).