

# Expanding cancer predisposition genes with ultra-rare cancer-exclusive human variations

Roni Rasnic<sup>1\*</sup>, Nathan Linial<sup>1</sup> and Michal Linial<sup>2</sup>

<sup>1</sup>The Rachel and Selim Benin School of Computer Science and Engineering, <sup>2</sup>Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

## Corresponding Author:

Roni Rasnic

**Keywords:** UK Biobank, TCGA, rare variants, Loss of function mutation, Germline mutations, gnomAD

Tables: 1-2; Figures: 1-4; Supplemental data Table S1-S3.

## Abstract

It is estimated that up to 10% of cancer incidents are attributed to inherited genetic alterations. Despite extensive research, there are still gaps in our understanding of genetic predisposition to cancer. It was theorized that ultra-rare variants partially account for the missing heritable component. We harness the UK BioBank dataset of ~500,000 individuals, 14% of which were diagnosed with cancer, to detect ultra-rare, possibly high-penetrance cancer predisposition variants. We report on 115 cancer-exclusive ultra-rare variations (CUVs) and nominate 26 variants with additional independent evidence as cancer predisposition variants. We conclude that population cohorts are valuable source for expanding the collection of novel cancer predisposition genes.

## Introduction

Discovery of cancer predisposition genes (CPGs) has the potential to impact personalized diagnosis and advance genetic consulting. Genetic analysis of family members with high occurrences of cancer has led to the identification of variants that increase the risk of developing cancer [1]. In addition to family-based studies, efforts to identify CPGs focus on pediatric patients where the contribution of environmental factors is expected to be small. Forty percent of pediatric cancer patients belong to families with a history of cancer [2].

Tumorigenesis results from mis-regulation of one or more of the major cancer hallmarks [3]. Therefore, it is anticipated that CPGs overlap with genes that are often mutated in cancerous tissues. Indeed, CPGs most prevalent in children (TP53, APC, BRCA2, NF1, PMS2, RB1, and RUNX1) are known cancer driver genes that function as tumor suppressors, oncogenes or have a role in maintaining DNA stability [4]. Many of the CPGs mutated in somatic tissues are associated with DNA-repair pathways and homologous recombination [5]. The inherited defects in cells' ability to repair and cope with DNA damage are considered as major factors in predisposition to breast and colorectal cancers [6].

Complementary approaches for seeking CPGs are large-scale genome / exome wide association studies (GWAS) which are conducted solely based on statistical considerations without prior knowledge on cancer promoting genes [7]. Identifying CPGs from GWAS is a challenge for the following reasons: (i) Limited contribution of genetic heritability in certain cancer types; (ii) Low effect size / risk associated with each individual variant; (iii) Low-penetrance in view of individual's background [8], and (iv) Low statistical power. Large cohorts of breast cancer show that ~2% of cancer cases are associated with mutations in BRCA1 and BRCA2 which are also high-risk ovarian cancer susceptibility genes. Additionally, TP53 and PTEN are associated with early-onset and high-risk familial breast cancer. Mutations in ATM and HRAS1 mildly increase the risk for breast cancer but strongly increase the risk for other cancer types and a collection of DNA mismatch repair genes (MLH1, MSH2, MSH6, PMS2) are associated with high risk of developing cancer [9]. A large cohort of Caucasian patients with pancreatic cancer reveal 6 high risk CPGs that overlap with other cancer types (CDKN2A, TP53, MLH1, BRCA2, ATM and BRCA1) [10].

Estimates for the heritable component of predisposition to cancer were extracted from GWAS, family-based and twin studies [11-13]. These estimates vary greatly with maximal

genetic contribution associated with thyroid and endocrine gland cancers, and a minimal one with stomach cancer and leukemia [14]. Current estimates suggest that as many as 10% of cancer incidents can be attributed to inherited genetic alterations (e.g., single variants and structural variations) [15, 16]. The actual contribution of CPGs varies according to gender, age of onset, cancer types and ethnicity [17-20]. It is evident that high risk variants with large effect sizes are very rare [21]. Actually, based on the heritability as reflected in GWAS catalog, it was estimated that only a fraction of existing CPGs is presently known [22]. Therefore, instances of extremely rare mutations with high risk for developing cancer remain to be discovered.

A catalog of 114 CPGs was compiled from 30 years of research [1] with about half of the reported genes derived from family studies representing high-penetrance variants. An extended catalog was reported with a total of 152 CPGs that were tested against rare variants from TCGA germline data, covering 10,389 cancer patients from 33 cancer types and included known pediatric CPGs [23]. The contribution of BRCA1/2, ATM, TP53 and PALB2 to cancer predisposition was confirmed.

In this study we report on known and novel cancer predisposition candidate genes. We benefit from the UK-Biobank (UKBB), an invaluable resource of germline genotyping data for ~500,000 individuals. The UKBB reports on ~70,000 cancer patients and ~430,000 cancer free individuals, considered as control group. We challenge the possibility that CPGs can be identified from very rare events, henceforth called cancer-exclusive ultra-rare variants (CUVs). These CUVs are expected to exhibit high penetrance. Here we report on 115 exome variations, 72 of which are heterologous. The majority of the matching genes are novel CPG candidates. We provide indirect genomic support for some of the CUVs that occur within coding genes and discuss their contribution to tumorigenesis.

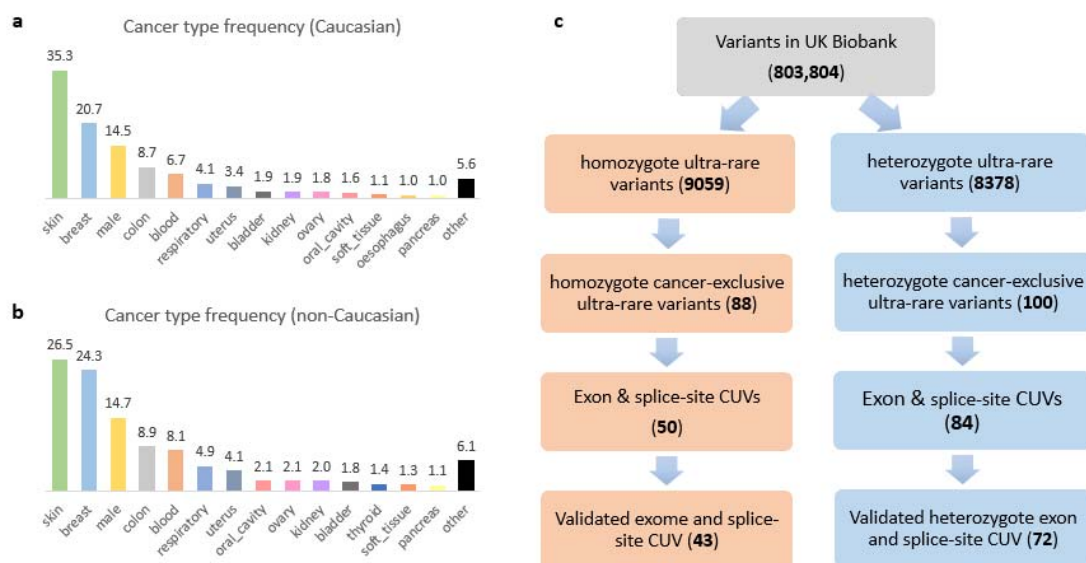
## Results

The primary UK Biobank (UKBB) data set used in the article is comprised of 325,407 Caucasian filtered UKBB participants (see Methods), 282,435 cancer-free (86.8%) and 42,972 diagnosed with at least one malignant neoplasm. Among participants with cancer, 55% were diagnosed with either skin or breast cancer. The ICD-10 codes assembly is summarized in Supplementary Table S1. A total of 13.2% of the cancer-diagnosed individuals had two or more distinct neoplasms diagnosed. The validation UKBB data set includes 70,544 non-Caucasian, filtered participants, 63,585 cancer free (90.1%). Fig. 1a and Fig. 1b provide further details on different cancer type prevalence in these sets.

### Compilation of cancer-exclusive ultra-rare variants (CUVs)

We scanned 803,804 genetic markers in our prime data set for cancer-exclusive variations. 188 variations met our initial criteria, appearing at least twice in individuals diagnosed with cancer and not appearing in cancer-free individuals. One hundred of the variations were heterozygous and 88 were homozygous. In order to target variations with additional supporting evidence, we considered only coding exome and splice-region variants. An

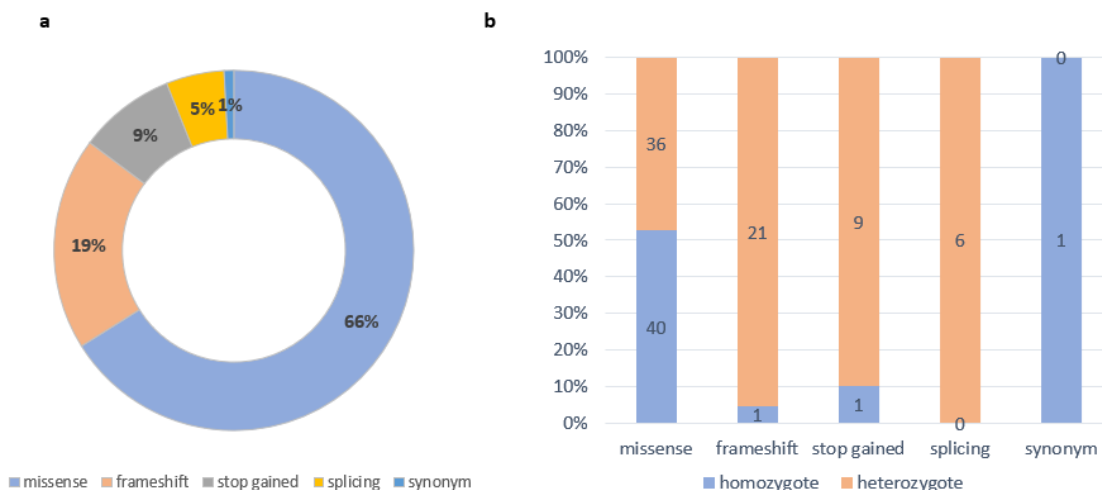
additional rarity filtration performed using the gnomAD data set reduced the list to 115 variants (associated with 108 genes), 72 heterozygous and 43 homozygous (Fig. 1b). The detailed list of all 115 CUVs can be found in Supplementary Table S2.



### Figure 1. UK Biobank CUVs collection

The Caucasian filtered UK Biobank (UKBB) data set include 42,972 individuals who had cancer and the non-Caucasian filtered include 6,959 such individuals. **(a)** Cancer type distribution for the Caucasian data set. **(b)** Cancer type distribution for the non-Caucasian data set. **(c)** Out of 803,304 UKBB variants, we curated 72 heterozygous and 43 homozygous CUVs (total 115 CUVs).

Most (66%) of the CUV are missense variants. There is a strong enrichment for loss of function (LoF) variants (i.e., frameshift, splicing disruption and stop gains), which account for 33% of the CUVs. Only a single homozygous CUV is synonymous (Fig. 2a). The distribution of variation types varies greatly between homozygous and heterozygous CUVs (Fig. 2b). Missense variants are 93% of the homozygous variants, but only 50% of the heterozygous CUVs. The heterozygous CUVs are highly enriched for LoF variants which constitute the other 50%.



**Figure 2. Exomic CUVs are mostly genomically disruptive**

The partition of variant types for the compiled list of 115 exomic CUVs. The list is dominated by transcript disruptive variations (99.1%) that include missense, frameshift, stop gain and splicing sites. **(a)** Distribution of variation types among the exomic CUVs. **(b)** Dispersion of variant type among heterozygous and homozygous CUVs.

### Cancer-exclusive ultra-rare variants (CUVs) overlap with known cancer predisposition genes

From the listed CUVs, 26 variants in 23 genes were previously defined as cancer inducing (Table 1). Specifically, 22 CUVs within 19 genes appear in the updated list of CPG catalog [23] and 24 CUVs within 21 genes are known cancer driver genes (Fig. 3a), as defined by either COSMIC [27] or the consensus gene catalog of driver genes (listing 299 genes, coined C299) [28]. More than half of the cancer associated variants result in loss of function (LoF). Many of the affected genes are tumor suppressor genes (TSGs), among which are prominent TSGs such as APC, BRCA1 and BRCA2 (Table 1), each identified by two distinct CUVs.

**Table 1.** CUVs overlap with known cancer predisposition or driver genes

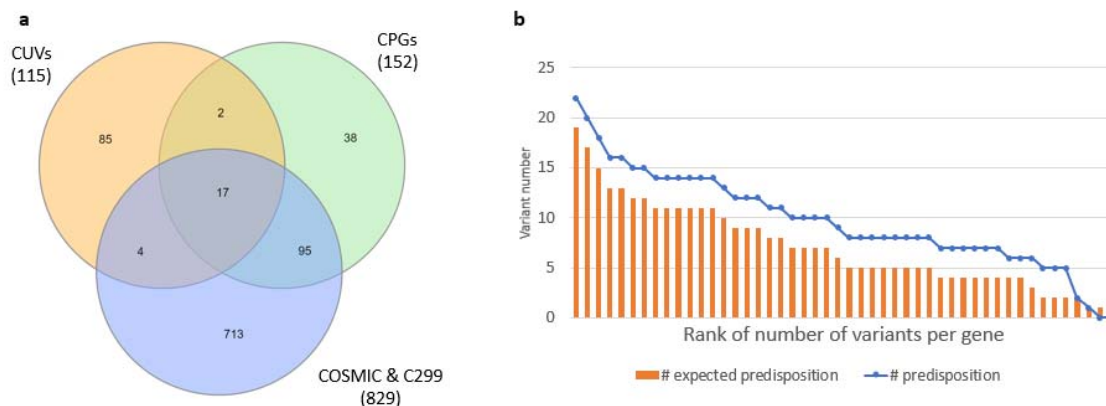
| hg19        | Effect        | Gene          | COSMIC | C299 | CPG | Function <sup>a</sup> |
|-------------|---------------|---------------|--------|------|-----|-----------------------|
| 1:155205517 | missense      | <b>GBA</b>    |        |      | Y   | Enzyme                |
| 2:48027130  | missense      | <b>MSH6</b>   | Y      | Y    | Y   | DNA repair            |
| 3:10183771  | missense      | <b>VHL</b>    | Y      | Y    | Y   | Ubq-complex           |
| 3:30730003  | splice region | <b>TGFBR2</b> | Y      | Y    |     | Kinase                |
| 3:37048480  | splice region | <b>MLH1</b>   | Y      | Y    | Y   | TSG                   |
| 5:112173671 | frameshift    | <b>APC</b>    | Y      | Y    | Y   | TSG                   |
| 5:112175255 | frameshift    | <b>APC</b>    | Y      | Y    | Y   | TSG                   |
| 9:101891277 | stop gain     | <b>TGFBR1</b> |        |      | Y   | Kinase                |
| 9:131341997 | missense      | <b>SPTAN1</b> |        | Y    |     | Cytoskeletal          |

|              |               |               |   |   |   |                  |
|--------------|---------------|---------------|---|---|---|------------------|
| 10:43609079  | frameshift    | <b>RET</b>    | Y | Y | Y | Kinase           |
| 10:88659605  | missense      | <b>BMPR1A</b> | Y |   | Y | Kinase           |
| 10:89717630  | stop gain     | <b>PTEN</b>   | Y | Y | Y | TSG, Phosphatase |
| 11:44193237  | missense      | <b>EXT2</b>   | Y |   | Y | TSG, Enzyme      |
| 11:71720337  | missense      | <b>NUMA1</b>  | Y |   |   | MT Spindle pole  |
| 11:108192066 | missense      | <b>ATM</b>    | Y | Y | Y | DDR, Kinase      |
| 13:32890621  | frameshift    | <b>BRCA2</b>  | Y | Y | Y | TSG, DNA repair  |
| 13:32914296  | missense      | <b>BRCA2</b>  | Y | Y | Y | TSG, DNA repair  |
| 13:48878061  | frameshift    | <b>RB1</b>    | Y | Y | Y | TSG              |
| 13:103524611 | frameshift    | <b>ERCC5</b>  | Y |   | Y | DNA repair       |
| 16:2121553   | missense      | <b>TSC2</b>   | Y | Y | Y | TSG              |
| 17:29654601  | missense      | <b>NF1</b>    | Y | Y | Y | RAS regulator    |
| 17:41244383  | frameshift    | <b>BRCA1</b>  | Y | Y | Y | TSG, DNA repair  |
| 17:41246296  | stop gain     | <b>BRCA1</b>  | Y | Y | Y | TSG, DNA repair  |
| 18:3451996   | frameshift    | <b>TGIF1</b>  |   | Y |   | TGF ligand       |
| 21:36421256  | splice region | <b>RUNX1</b>  | Y | Y | Y | TF               |
| 22:30067894  | missense      | <b>NF2</b>    | Y | Y | Y | Cytoskeletal     |

<sup>9</sup>Function abbreviation: DDR, DNA damage response; TSG, tumor suppressor gene; TF, transcription factor; MT, microtubule; Ubq, ubiquitin.

The heterozygous CUVs are enriched for known cancer predisposition genes. Twenty-five of the cancer associated CUVs are heterologous and one is homologous. Notably, there is an inherent imbalance in the initial variant sampling performed by the UKBB. As the UKBB use DNA arrays for obtaining genomic data, the identifiability of ultra-rare exome variants is restricted by the selection of SNP markers and the design of the array. There are 6450 heterozygous ultra-rare exome variants from 2938 genes which pass our biobank-ethnic and the gnomAD allele frequency filtration. A total of 1604 of the filtered ultra-rare variants are from 105 known CPGs, with some genes that are over-represented in the ultra-rare variants (Supplemental Table S3). For example, the exomic region of BRCA2 is covered by 226 such SNP marker variants, while most genes have none.

In order to account for the disproportional number of the ultra-rare variant of some CPGs, we calculated the expected number of cancer predisposed genes when gradually removing highly-represented genes from the ultra-rare variant collection for heterozygotes. As shown in Fig. 3b, there is an enrichment towards CPGs and even more so as we remove variants of over-represented genes (e.g., BRCA2). We calculated p-values for each data-point using a two-side binomial test. The results are available in Supplemental Table S3.



### Figure 3. CUVs list is enriched with cancer predisposition genes

Out of the 108 genes in the CUVs list, 23 are known cancer genes. **(a)** Venn diagram of the genes associated with CUVs, known cancer driver genes (as reported in COMICS) and the consensus CPGs. **(b)** Expected number of known CPG CUV (orange) versus the actual number of known CPG in heterozygote CUVs (blue). An unbalanced representation of genes in ultra-rare variants of UKBB results in over-representation of some genes. We therefore ranked the genes based on number of ultra-rare variants (Supplementary Table S3). For each rank, we present the expected number of CUVs from CPGs and the actual number observed for CUVs from CPGs.

### Independent genetic validation

Due to the extremely rare nature of the CUVs, we require additional support for the collection of the CPG candidates. We seek independent genetic validation of the non-cancer related CUVs. We apply three sources for validation: (i) the filtered Caucasian UKBB cohort; (ii) the matched filtered, non-Caucasian UKBB cohort; (iii) the collection of germline variants from TCGA, as reported in gnomAD. The complete list of genetically validated novel CPG candidates is listed in Table 2.

Within the Caucasian cohort, we consider the following as additional genomic evidence: (i) a gene with 2 CUVs, or (ii) any CUV seen in more than two individuals diagnosed with cancer. We found 7 genes that have 2 distinct CUVs, 3 of which are already known CPGs: BRCA1, BRCA2 and APC. The other 4 genes are likely novel CPG candidates: DSP, KCNH2, MYBPC3 and SCN5A. There are 9 CUVs which we detected in three individuals with cancer. Three of them are known predisposition or driver genes: NF1, ATM and TGFBR2. The other 6 genes are CPG candidates that were not previously assigned as such. This set includes PCDHB16, DNAH3, ENDOU, AGR2, HIST1H2BO and NAV3. Interestingly, a certain homozygous CUV in the gene ICAM1 appeared in 4 individuals with cancer in our filtered Caucasian cohort.

The non-Caucasian UKBB cohort provides additional independent genomic evidence. There are 5 CUVs that appear at least once in an individual with cancer from the non-Caucasian cohort. CUVs from the genes MYO1E, SARDH and ISLR appeared in two distinct individuals with cancer from this non-Caucasian cohort, while CUVs from PCDHB16 and known CPG BMP1A appeared in a single individual with cancer.

TCGA germline variants were obtained using exome sequencing and thus offer an additional separate source for CUV validation. Clearly, the appearance of CUVs in TCGA germline data

is not anticipated, as we discuss variants that are ultra-rare in both UKBB and gnomAD. The TCGA collection within gnomAD includes only 7,269 samples. We identified 10 CUVs that were also observed in TCGA gnomAD germline data, one of a known cancer driver gene TGIF1, and 9 novel CPG candidates: PCDHB16, EGFLAM, AKR1C2, MAP3K15, MRPL39, DNAH3, WDFY4, HSPB2 and ZFC3H1.

Based on the above support, we compiled a list of validated CPGs which includes 21 genes that are novel CPGs. Among these genes 12 CUVs are heterozygous, 8 are homozygous and MYBPC3 has both heterozygous and homozygous CUVs. Two of these genes have multiple validation evidence. DNAH3 with a homozygous CUV which appears in 3 individuals with cancer in the Caucasian cohort and within TCGA germline variant collection. PCDHB16 with a homozygous CUV which appeared in 3 individuals in the Caucasian cohort, one individual in the non-Caucasian cohort and in the TCGA gnomAD resource. In addition, non-CPG cancer-driver genes with validated CUVs include TGFBR2 and TGIF1 that are also very likely CPG candidates.

Some of the prominent genes in our list were signified by additional independent studies. For example, a novel oncolytic agent targeting ICAM1 against bladder cancer is now in phase 1 of a clinical trial [38]. Additionally, DNAH3 was identified as novel predisposition gene using exome sequencing in a Tunisian family with multiple non-BRCA breast cancer instances [35].

**Table 2.** Novel validated CPG candidates

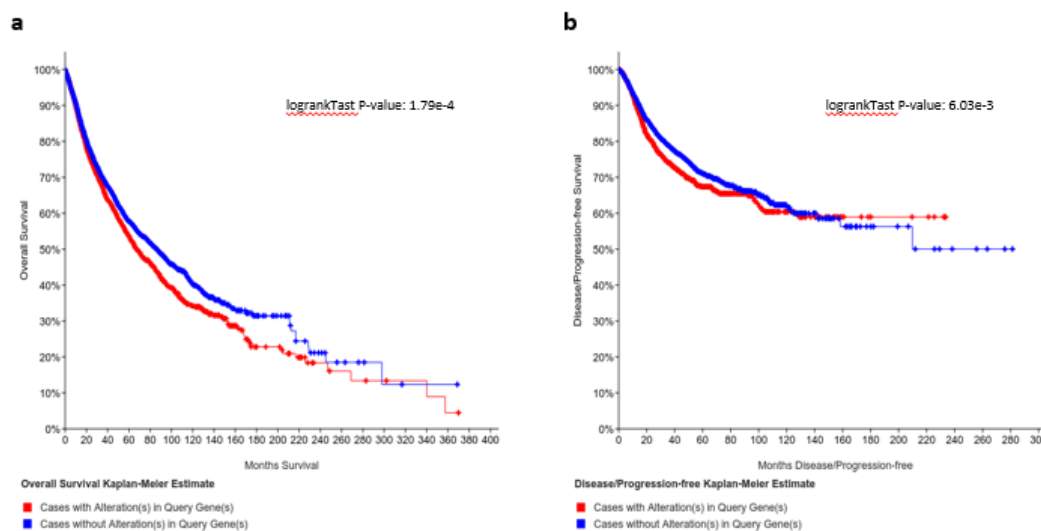
| Gene Symbol | zygote form | # people per CUV | Distinct CUVs | non-Caucasian cohort | TCGA germline | function in tumorigenesis  | Ref. |
|-------------|-------------|------------------|---------------|----------------------|---------------|--|------|
| AGR2        | hetero      | 3                |               |                      |               | Affects cell migration, transformation and metastasis. Wnt signaling, tumor antigen. | [34] |
| AKR1C2      | hetero      | 2                |               |                      | <b>Y</b>      | Exerts an inhibitory effect on oncogenesis   | [35] |
| DNAH3       | homo        | 3                |               |                      | <b>Y</b>      | Cancer predisposed genes in Tunisian family  | [36] |
| DSP         | hetero      | 2                | <b>Y</b>      |                      |               | Affects cell adhesion. Suppressed by TGF- $\beta$                                    |      |
| EGFLAM      | hetero      | 2                |               |                      | <b>Y</b>      | Promotes matrix assembly   |      |
| ENDOU       | homo        | 3                |               |                      |               | Cancer biomarker   | [37] |
| HIST1H2BO   | hetero      | 3                |               |                      |               | Affects major signaling pathways   |      |
| HSPB2       | hetero      | 2                |               |                      | <b>Y</b>      | Epigenetically regulated   | [38] |
| ICAM1       | homo        | 4                |               |                      |               | Biomarker, under a clinical trial  | [39] |
| ISLR        | homo        | 2                |               | <b>Y</b>             |               | Marker for mesenchymal stem cells. Deregulated gene in cancer                        | [40] |
| KCNH2       | hetero      | 2                | <b>Y</b>      |                      |               | Affects proliferation and migration  |      |
| MAP3K15     | hetero      | 2                |               |                      | <b>Y</b>      | Contributes to cell migration  |      |
| MRPL39      | hetero      | 2                |               |                      | <b>Y</b>      | Tumor suppressor by targeting miR-130  | [41] |
| MYBPC3      | both        | 2                | <b>Y</b>      |                      |               | Cytoskeletal modifier  |      |



|         |        |   |   |   |   |  |      |
|---------|--------|---|---|---|---|--|------|
| MYO1E   | homo   | 2 |   | Y |   | Stimulates upregulation of motility and invasion       | [42] |
| NAV3    | hetero | 3 |   |   |   | Acts as a suppressor of breast cancer                  | [43] |
| PCDHB16 | homo   | 3 |   | Y | Y |  |      |
| SARDH   | homo   | 2 |   | Y |   | Acts as tumor suppressor                               | [44] |
| SCN5A   | hetero | 2 | Y |   |   | Promotes breast cancer, possess anti-pancreatic cancer | [45] |
| WDFY4   | hetero | 2 |   |   | Y | Presentats viral, tumor antigen on dendritic cells     | [46] |
| ZFC3H1  | homo   | 2 |   |   | Y | Indirect activating DNA repair                         |      |

### Somatic mutations in novel CPGs significantly decrease survival rate

There is substantial overlap between CPGs and known cancer driver genes (Fig. 3a). This overlap suggests that somatic mutations in validated CPG candidates may have an impact on patients' survival rate. We tested this hypothesis for the 21 novel CPG candidates (Table 2) using a curated set of 32 non-redundant TCGA studies (compiled in cBioPortal [1]) that cover 10,953 patients. By testing the impact of alteration in the 21 novel CPGs in somatic data we expect to provide a functional link between the germline CPG findings and the matched mutated genes in somatic cancer samples. Altogether, 3,846 (35%) of the patients had somatic mutations in one or more of the genes. The median survival of patients with somatic mutations in these genes is 67.4 months, while the median for patients without somatic mutations in any of these genes is much longer (86.3 months). Applying the Kaplan-Meier survival estimate yields a p-value of  $1.78e-4$  in the Logrank test (Fig. 4a). The Kaplan-Meier disease/progression-free estimate was also worse for patients with somatic mutations in the 21 novel CPGs with a p-value of  $6.03e-3$  (Fig. 4b).



**Figure 4. Somatic mutations in CPG candidate effect cancer patient survival and disease progression**

The effect of somatic mutations in the 21 novel CPG candidate (Table 2) on the somatic mutations from TCGA cancer patients was tested (via cBioPortal). **(a)** Meier-Kaplan survival rate estimate. **(b)** Meier-Kaplan Disease / Progression-free estimate.

We conclude that the CUV-based CPG candidate genes from UKBB carry a strong signature that is manifested in patients' survival, supporting the notion that these genes belong to an extended set of previously overlooked CPGs.

### **Homozygous variations are mainly recessive**

In order to ascertain whether the homozygous variations found are indicative of the heterozygous form of the variant as well, we viewed the heterozygous prevalence within the UKBB Caucasian population. In only a single variant in the gene MYO1E was the prevalence in healthy individuals significantly lower than in individuals with cancer ( $p$ -value: 0.04). As most of the variations have a strong cancer predisposition effect as homozygous variations, it seems that their influence is explained by a recessive inheritance mode. This phenomenon might explain the significant depletion of known CPGs within the homozygous variations in our list.

Inspecting the heritability model of previously reported CPGs [1] is in accord with our findings, showing that while about two-thirds of the genes comply with a dominant inheritance, the rest are likely to be recessive. Notably, in the most updated CPG catalog, 15% of the genes were assigned with both inheritance patterns. In our ultra-rare list, only MYBPC3 is associated with both heterozygous and homozygous variations.

### **Discussion**

We present a list of 115 CUVs from 108 genes. Among them 24 variants (from 21 genes) are associated with known cancer genes. Most of these variants (21) overlap with known cancer predisposition genes. Expanding the number of currently identified CPGs is crucial for better understanding of tumorigenesis and identifying various processes causing high cancer penetrance. Genetic consulting, family planning and appropriate treatment is a direct outcome of an accurate and exhaustive list of CPGs.

Known cancer predisposition variants only partially explain the cases of inherited cancer incidents. CPGs identification has already impacted cancer diagnostics, therapy and prognosis [1]. Genomic tests and gene panel for certain cancer predisposition markers are commonly used for early detection and in preventative medicine [29, 30]. It is likely that CPGs based on ultra-rare variants are not saturated. For example, additional CPGs including CDKN2A, NF1, and NBN were associated with an increased risk for breast cancer [31]. Specifically, CDKN2A has been also detected as a CPG in families of patients with pancreatic cancer [32].

With the impending availability of UKBB exome sequencing (150,000 exomes), we will be able to revisit the identified variants and further refine the list of candidate CPGs (i.e., remove false-positives and add evidence to support true CPGs). Inspecting the function of genes associated with the 108 identified genes further supports the importance of protein modification (e.g. kinases and phosphatase function), chromatin epigenetic signatures [33], membrane signaling, DNA repair systems and more.

The inheritably rare nature of CUVs raise doubts on the reliability of their initial identification [26]. Therefore, we only considered as candidate CPGs those genes that are supported by additional independent genomic evidence from either the UKBB or the TCGA cohort. We nominate 23 genes as CPG candidates, two of which are known cancer drivers. As we have shown (Fig. 4), somatic mutations in the non-driver validated CPG candidates resulted in a significant negative effect on the patients' survival rate.

## **Materials and Methods**

### **Study population**

The UK Biobank (UKBB) has recruited ~500,000 people from the general population of the UK, using National Health Service patient registers, with no exclusion criteria [24]. Participants were between 40 and 69 years of age at the time of recruitment, between 2006 and 2010. To avoid biases due to familial relationships, we removed 75,853 samples keeping only one representative of each kinship group of related individuals. 312 additional samples had mismatching sex (between the self-reported and the genetics-derived) and 726 samples had only partial genotyping. We restricted our initial exploratory analysis to individuals that were genetically verified as Caucasians who also declared themselves as 'white'. The rest of the participants not matching this criterion were used for the complementing analysis. The filtered Caucasian cohort includes 325,407 individuals (42,972 of whom had cancer) and the filtered non-Caucasian cohort includes 70,544 individuals (6,959 had cancer). Additional data from gnomAD was used for variant rarity filtering and TCGA-germline validation [25].

### **Rare variants reliability**

Our CUV collection includes variants that appeared at least twice in the filtered Caucasian cohort, thereby evading many SNP-genotyping inaccuracies [26]. We further ascertain the validity of prominent variants with additional genomic evidence.

### **Cancer type definition**

The UKBB provides an ICD-10 code for each diagnosed condition. We considered an individual diagnosed with malignant neoplasm (ICD-10 codes C00-C97) as individuals with cancer, and otherwise as cancer-free individuals. The codes were aggregated to improve data readability using the assembly described in Supplementary Table S1.

**Ethical approval:** UK Biobank approval was obtained as part of the project 26664. Ethical approval for this study was obtained from the committee for ethics in research involving human subjects, for the faculty of medicine, The Hebrew University, Jerusalem, Israel (Approval number - 13082019).

### **Data availability**

Most of the data that support the findings of this study are available from the UKBB. However, restrictions apply to the availability of these data, which were used under license

for the current study, and so are not publicly available. Data are available from the authors upon a justified request and with permission of the UKBB. Data extracted from gnomAD is available from the authors upon request.

## References

1. Rahman, N. (2014) Realizing the promise of cancer predisposition genes, *Nature*. **505**, 302-8.
2. Zhang, J., Walsh, M. F., Wu, G., Edmonson, M. N., Gruber, T. A., Easton, J., Hedges, D., Ma, X., Zhou, X. & Yergeau, D. A. (2015) Germline mutations in predisposition genes in pediatric cancer, *New England Journal of Medicine*. **373**, 2336-2346.
3. Hanahan, D. & Weinberg, R. A. (2011) Hallmarks of cancer: the next generation, *Cell*. **144**, 646-74.
4. Vogelstein, B. & Kinzler, K. W. (2004) Cancer genes and the pathways they control, *Nature medicine*. **10**, 789.
5. Bertelsen, B., Tuxen, I. V., Yde, C. W., Gabrielaite, M., Torp, M. H., Kinalis, S., Oestrup, O., Rohrberg, K., Spangaard, I., Santoni-Rugiu, E., Wadt, K., Mau-Sorensen, M., Lassen, U. & Nielsen, F. C. (2019) High frequency of pathogenic germline variants within homologous recombination repair in patients with advanced cancer, *NPJ Genom Med*. **4**, 13.
6. Easton, D. F. (1999) How many more breast cancer predisposition genes are there?, *Breast Cancer Research*. **1**, 14.
7. Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. & Manolio, T. A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, *Proc Natl Acad Sci U S A*. **106**, 9362-7.
8. Galvan, A., Ioannidis, J. P. & Dragani, T. A. (2010) Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer, *Trends Genet*. **26**, 132-41.
9. Baria, K., Warren, C., Roberts, S. A., West, C. M. & Scott, D. (2001) Chromosomal radiosensitivity as a marker of predisposition to common cancers?, *Br J Cancer*. **84**, 892-6.
10. Hu, C., Hart, S. N., Polley, E. C., Gnanaolivu, R., Shimelis, H., Lee, K. Y., Lilyquist, J., Na, J., Moore, R., Antwi, S. O., Bamlet, W. R., Chaffee, K. G., DiCarlo, J., Wu, Z., Samara, R., Kasi, P. M., McWilliams, R. R., Petersen, G. M. & Couch, F. J. (2018) Association Between Inherited Germline Mutations in Cancer Predisposition Genes and Risk of Pancreatic Cancer, *JAMA*. **319**, 2401-2409.
11. Verkasalo, P. K., Kaprio, J., Koskenvuo, M. & Pukkala, E. (1999) Genetic predisposition, environment and cancer incidence: a nationwide twin study in Finland, 1976–1995, *International Journal of Cancer*. **83**, 743-749.
12. Frank, S. A. (2004) Genetic predisposition to cancer - insights from population genetics, *Nat Rev Genet*. **5**, 764-72.
13. Law, P. J., Timofeeva, M., Fernandez-Rozadilla, C., Broderick, P., Studd, J., Fernandez-Tajes, J., Farrington, S., Svinti, V., Palles, C. & Orlando, G. (2019) Association analyses identify 31 new risk loci for colorectal cancer susceptibility, *Nature communications*. **10**, 2154.
14. Czene, K., Lichtenstein, P. & Hemminki, K. (2002) Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database, *Int J Cancer*. **99**, 260-6.

15. Economopoulou, P., Dimitriadis, G. & Psyrri, A. (2015) Beyond BRCA: new hereditary breast cancer susceptibility genes, *Cancer treatment reviews*. **41**, 1-8.
16. Grant, R. C., Selander, I., Connor, A. A., Selvarajah, S., Borgida, A., Briollais, L., Petersen, G. M., Lerner-Ellis, J., Holter, S. & Gallinger, S. (2015) Prevalence of germline mutations in cancer predisposition genes in patients with pancreatic cancer, *Gastroenterology*. **148**, 556-564.
17. Petersen, G. M., Amundadottir, L., Fuchs, C. S., Kraft, P., Stolzenberg-Solomon, R. Z., Jacobs, K. B., Arslan, A. A., Bueno-de-Mesquita, H. B., Gallinger, S., Gross, M., Helzlsouer, K., Holly, E. A., Jacobs, E. J., Klein, A. P., LaCroix, A., Li, D., Mandelson, M. T., Olson, S. H., Risch, H. A., Zheng, W., Albanes, D., Bamlet, W. R., Berg, C. D., Boutron-Ruault, M. C., Buring, J. E., Bracci, P. M., Canzian, F., Clipp, S., Cotterchio, M., de Andrade, M., Duell, E. J., Gaziano, J. M., Giovannucci, E. L., Goggins, M., Hallmans, G., Hankinson, S. E., Hassan, M., Howard, B., Hunter, D. J., Hutchinson, A., Jenab, M., Kaaks, R., Kooperberg, C., Krogh, V., Kurtz, R. C., Lynch, S. M., McWilliams, R. R., Mendelsohn, J. B., Michaud, D. S., Parikh, H., Patel, A. V., Peeters, P. H., Rajkovic, A., Riboli, E., Rodriguez, L., Seminara, D., Shu, X. O., Thomas, G., Tjonneland, A., Tobias, G. S., Trichopoulos, D., Van Den Eeden, S. K., Virtamo, J., Wactawski-Wende, J., Wang, Z., Wolpin, B. M., Yu, H., Yu, K., Zeleniuch-Jacquotte, A., Fraumeni, J. F., Jr., Hoover, R. N., Hartge, P. & Chanock, S. J. (2010) A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33, *Nat Genet*. **42**, 224-8.
18. Wolpin, B. M., Rizzato, C., Kraft, P., Kooperberg, C., Petersen, G. M., Wang, Z., Arslan, A., Beane-Freeman, L., Bracci, P. M., Buring, J., Canzian, F., Duell, E. J., Gallinger, S., Giles, G., Goodman, G. E., Goodman, P. J., Jacobs, E. J., Kamineni, A., Klein, A. P., Kolonel, L. N., Kulke, M. H., Li, D., Malats, N., Olson, S. H., Risch, H. A., Sesso, H. D., Visvanathan, K., White, E., Zheng, W., Abnet, C. C., Albanes, D., Andreotti, G., Austin, M. A., Barfield, R., Basso, D., Berndt, S. I., Boutron-Ruault, M. C., Brotzman, M., Buchler, M. W., Bueno-de-Mesquita, H. B., Bugert, P., Burdette, L., Campa, D., Caporaso, N. E., Capurso, G., Chung, C., Cotterchio, M., Costello, E., Elena, J., Funel, N., Gaziano, J. M., Giese, N. A., Giovannucci, E. L., Goggins, M., Gorman, M. J., Gross, M., Haiman, C. A., Hassan, M., Helzlsouer, K. J., Henderson, B. E., Holly, E. A., Hu, N., Hunter, D. J., Innocenti, F., Jenab, M., Kaaks, R., Key, T. J., Khaw, K. T., Klein, E. A., Kogevinas, M., Krogh, V., Kupcinskis, J., Kurtz, R. C., LaCroix, A., Landi, M. T., Landi, S., Le Marchand, L., Mambrini, A., Mannisto, S., Milne, R. L., Nakamura, Y., Oberg, A. L., Owzar, K., Patel, A. V., Peeters, P. H., Peters, U., Pezzilli, R., Piepoli, A., Porta, M., Real, F. X., Riboli, E., Rothman, N., Scarpa, A., Shu, X. O., Silverman, D. T., Soucek, P., Sund, M., Talar-Wojnarowska, R., Taylor, P. R., Theodoropoulos, G. E., et al. (2014) Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer, *Nat Genet*. **46**, 994-1000.
19. Long, J., Cai, Q., Sung, H., Shi, J., Zhang, B., Choi, J. Y., Wen, W., Delahanty, R. J., Lu, W., Gao, Y. T., Shen, H., Park, S. K., Chen, K., Shen, C. Y., Ren, Z., Haiman, C. A., Matsuo, K., Kim, M. K., Khoo, U. S., Iwasaki, M., Zheng, Y., Xiang, Y. B., Gu, K., Rothman, N., Wang, W., Hu, Z., Liu, Y., Yoo, K. Y., Noh, D. Y., Han, B. G., Lee, M. H., Zheng, H., Zhang, L., Wu, P. E., Shieh, Y. L., Chan, S. Y., Wang, S., Xie, X., Kim, S. W., Henderson, B. E., Le Marchand, L., Ito, H., Kasuga, Y., Ahn, S. H., Kang, H. S., Chan, K. Y., Iwata, H., Tsugane, S., Li, C., Shu, X. O., Kang, D. H. & Zheng, W. (2012) Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer, *PLoS Genet*. **8**, e1002532.
20. Thomas, G., Jacobs, K. B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., Crenshaw, A., Cancel-Tassin, G., Staats, B. J., Wang, Z., Gonzalez-Bosquet, J., Fang, J., Deng, X., Berndt, S. I., Calle, E. E., Feigelson, H. S., Thun, M.

- J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F. R., Giovannucci, E., Willett, W. C., Cussenot, O., Valeri, A., Andriole, G. L., Crawford, E. D., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Jr., Hoover, R., Hayes, R. B., Hunter, D. J. & Chanock, S. J. (2008) Multiple loci identified in a genome-wide association study of prostate cancer, *Nat Genet.* **40**, 310-5.
21. Mancuso, N., Rohland, N., Rand, K. A., Tandon, A., Allen, A., Quinque, D., Mallick, S., Li, H., Stram, A., Sheng, X., Kote-Jarai, Z., Easton, D. F., Eeles, R. A., consortium, P., Le Marchand, L., Lubwama, A., Stram, D., Watya, S., Conti, D. V., Henderson, B., Haiman, C. A., Pasaniuc, B. & Reich, D. (2016) The contribution of rare variation to prostate cancer heritability, *Nat Genet.* **48**, 30-5.
22. Jiao, S., Peters, U., Berndt, S., Brenner, H., Butterbach, K., Caan, B. J., Carlson, C. S., Chan, A. T., Chang-Claude, J., Chanock, S., Curtis, K. R., Duggan, D., Gong, J., Harrison, T. A., Hayes, R. B., Henderson, B. E., Hoffmeister, M., Kolonel, L. N., Le Marchand, L., Potter, J. D., Rudolph, A., Schoen, R. E., Seminara, D., Slattery, M. L., White, E. & Hsu, L. (2014) Estimating the heritability of colorectal cancer, *Hum Mol Genet.* **23**, 3898-905.
23. Huang, K. L., Mashl, R. J., Wu, Y., Ritter, D. I., Wang, J., Oh, C., Paczkowska, M., Reynolds, S., Wyczalkowski, M. A., Oak, N., Scott, A. D., Krassowski, M., Cherniack, A. D., Houlihan, K. E., Jayasinghe, R., Wang, L. B., Zhou, D. C., Liu, D., Cao, S., Kim, Y. W., Koire, A., McMichael, J. F., Huchtagowder, V., Kim, T. B., Hahn, A., Wang, C., McLellan, M. D., Al-Mulla, F., Johnson, K. J., Cancer Genome Atlas Research, N., Lichtarge, O., Boutros, P. C., Raphael, B., Lazar, A. J., Zhang, W., Wendl, M. C., Govindan, R., Jain, S., Wheeler, D., Kulkarni, S., Dipersio, J. F., Reimand, J., Meric-Bernstam, F., Chen, K., Shmulevich, I., Plon, S. E., Chen, F. & Ding, L. (2018) Pathogenic Germline Variants in 10,389 Adult Cancers, *Cell.* **173**, 355-370 e14.
24. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T. & Collins, R. (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age, *PLoS Med.* **12**, e1001779.
25. Karczewski, K. & Francioli, L. (2017) The genome Aggregation Database (gnomAD), *MacArthur Lab.*
26. Weedon, M. N., Jackson, L., Harrison, J. W., Ruth, K. S., Tyrrell, J., Hattersley, A. T. & Wright, C. F. (2019) Assessing the analytical validity of SNP-chips for detecting very rare pathogenic variants: implications for direct-to-consumer genetic testing, *BioRxiv*, 696799.
27. Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C. & Ward, S. (2014) COSMIC: exploring the world's knowledge of somatic mutations in human cancer, *Nucleic acids research.* **43**, D805-D811.
28. Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J. & Reardon, B. (2018) Comprehensive characterization of cancer driver genes and mutations, *Cell.* **173**, 371-385. e18.
29. Easton, D. F., Pharoah, P. D., Antoniou, A. C., Tischkowitz, M., Tavtigian, S. V., Nathanson, K. L., Devilee, P., Meindl, A., Couch, F. J. & Southey, M. (2015) Gene-panel sequencing and the prediction of breast-cancer risk, *New England Journal of Medicine.* **372**, 2243-2257.
30. Couch, F. J., Shimelis, H., Hu, C., Hart, S. N., Polley, E. C., Na, J., Hallberg, E., Moore, R., Thomas, A. & Lilyquist, J. (2017) Associations between cancer predisposition testing panel genes and breast cancer, *JAMA oncology.* **3**, 1190-1196.

31. LaDuca, H., Polley, E. C., Yussuf, A., Hoang, L., Gutierrez, S., Hart, S. N., Yadav, S., Hu, C., Na, J., Goldgar, D. E., Fulk, K., Smith, L. P., Horton, C., Profato, J., Pesaran, T., Gau, C. L., Pronold, M., Davis, B. T., Chao, E. C., Couch, F. J. & Dolinsky, J. S. (2019) A clinical guide to hereditary cancer panel testing: evaluation of gene-specific cancer associations and sensitivity of genetic testing criteria in a cohort of 165,000 high-risk patients, *Genet Med*.
32. Chaffee, K. G., Oberg, A. L., McWilliams, R. R., Majithia, N., Allen, B. A., Kidd, J., Singh, N., Hartman, A. R., Wenstrup, R. J. & Petersen, G. M. (2018) Prevalence of germ-line mutations in cancer genes among pancreatic cancer patients with a positive family history, *Genet Med*. **20**, 119-127.
33. Wang, Q. (2016) Cancer predisposition genes: molecular mechanisms and clinical impact on personalized cancer care: examples of Lynch and HBOC syndromes, *Acta Pharmacol Sin*. **37**, 143-9.
34. Patel, P., Clarke, C., Barraclough, D. L., Jowitt, T. A., Rudland, P. S., Barraclough, R. & Lian, L. Y. (2013) Metastasis-promoting anterior gradient 2 protein has a dimeric thioredoxin fold structure and a role in cell adhesion, *J Mol Biol*. **425**, 929-43.
35. Jin, Y. X., Zhou, X. F., Chen, Y. Y., Jin, W. X., Wang, Y. H., Ye, D. R., Sun, Y. H., Li, Y. F., Wang, Q. X., Zhang, X. H., Wang, O. C. & Chen, E. D. (2019) Up-Regulated AKR1C2 is correlated with favorable prognosis in thyroid carcinoma, *J Cancer*. **10**, 3543-3552.
36. Hamdi, Y., Boujemaa, M., Ben Rekaya, M., Ben Hamda, C., Mighri, N., El Benna, H., Mejri, N., Labidi, S., Daoud, N., Naouali, C., Messaoud, O., Chargui, M., Ghedira, K., Boubaker, M. S., Mrad, R., Boussem, H., Abdelhak, S. & Consortium, P. E. C. (2018) Family specific genetic predisposition to breast cancer: results from Tunisian whole exome sequenced breast cancer cases, *J Transl Med*. **16**, 158.
37. Grundmann, U., Romisch, J., Siebold, B., Bohn, H. & Amann, E. (1990) Cloning and expression of a cDNA encoding human placental protein 11, a putative serine protease with diagnostic significance as a tumor marker, *DNA Cell Biol*. **9**, 243-50.
38. Yin, A. A., He, Y. L., Etcheverry, A., Liu, Y. H., Aubry, M., Barnholtz-Sloan, J., Liu, B. L., Mosser, J., Lu, Z. F. & Zhang, X. (2019) Novel predictive epigenetic signature for temozolomide in non-G-CIMP glioblastomas, *Clin Epigenetics*. **11**, 76.
39. Annels, N. E., Mansfield, D., Arif, M., Ballesteros-Merino, C., Simpson, G. R., Denyer, M., Sandhu, S. S., Melcher, A. A., Harrington, K. J., Davies, B., Au, G., Grose, M., Bagwan, I., Fox, B., Vile, R., Mostafid, H., Shafren, D. & Pandha, H. S. (2019) Phase I Trial of an ICAM-1-Targeted Immunotherapeutic-Coxsackievirus A21 (CVA21) as an Oncolytic Agent Against Non Muscle-Invasive Bladder Cancer, *Clin Cancer Res*. **25**, 5818-5831.
40. Martinez-Aguilar, J., Clifton-Bligh, R. & Molloy, M. P. (2016) Proteomics of thyroid tumours provides new insights into their molecular composition and changes associated with malignancy, *Sci Rep*. **6**, 23660.
41. Yu, M. J., Zhao, N., Shen, H. & Wang, H. (2018) Long Noncoding RNA MRPL39 Inhibits Gastric Cancer Proliferation and Progression by Directly Targeting miR-130, *Genet Test Mol Biomarkers*. **22**, 656-663.
42. Ouderkirk-Pecone, J. L., Goreczny, G. J., Chase, S. E., Tatum, A. H., Turner, C. E. & Krendel, M. (2016) Myosin 1e promotes breast cancer malignancy by enhancing tumor cell proliferation and stimulating tumor cell de-differentiation, *Oncotarget*. **7**, 46419-46432.
43. Cohen-Dvashi, H., Ben-Chetrit, N., Russell, R., Carvalho, S., Lauriola, M., Nisani, S., Mancini, M., Nataraj, N., Kedmi, M., Roth, L., Kostler, W., Zeisel, A., Yitzhaky, A., Zylberg, J., Tarcic, G., Eilam, R., Wigelman, Y., Will, R., Lavi, S., Porat, Z., Wiemann, S., Ricardo, S.,

- Schmitt, F., Caldas, C. & Yarden, Y. (2015) Navigator-3, a modulator of cell migration, may act as a suppressor of breast cancer progression, *EMBO Mol Med.* **7**, 299-314.
44. He, H., Chen, E., Lei, L., Yan, B., Zhao, X., Zhu, Z., Li, Q., Zhang, P., Zhang, W., Xing, J., Du, L., Dong, J. & Yang, J. (2019) Alteration of the tumor suppressor SARDH in sporadic colorectal cancer: A functional and transcriptome profiling-based study, *Mol Carcinog.* **58**, 957-966.
45. Mao, W., Zhang, J., Korner, H., Jiang, Y. & Ying, S. (2019) The Emerging Role of Voltage-Gated Sodium Channels in Tumor Biology, *Front Oncol.* **9**, 124.
46. Theisen, D. J., Davidson, J. T. t., Briseno, C. G., Gargaro, M., Lauron, E. J., Wang, Q., Desai, P., Durai, V., Bagadia, P., Brickner, J. R., Beatty, W. L., Virgin, H. W., Gillanders, W. E., Mosammaparast, N., Diamond, M. S., Sibley, L. D., Yokoyama, W., Schreiber, R. D., Murphy, T. L. & Murphy, K. M. (2018) WDFY4 is required for cross-presentation in response to viral and tumor antigens, *Science.* **362**, 694-699.



## Acknowledgement

We would also like to thank Nadav Brandes from the School of Computer Science and Engineering at the Hebrew University of Jerusalem for useful discussion and valuable comments. We thank Irene Unterman from the Medical School at the Hebrew University of Jerusalem for reading the manuscript. We thank the CSE system at the Hebrew University of Jerusalem team for their technical support.

## Competing Interests Statement

The authors declare that they have no competing interests.

## Supplementary files

**Table S1.** ICD10-code assembly

**Table S2.** The collection of 115 exome CUVs.

**Table S3.** Number of rare variants per gene, p-values for expected vs. observed CUVs from CPGs.