SCIENTIFIC REPORTS
nature research

OPEN

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

Roni Rasnic[1]✉, Nathan Linial[1] & Michal Linial[2]

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 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.

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*)[2] are known cancer driver genes that function as tumor suppressors, oncogenes or have a role in maintaining DNA stability[4]. Many of the predisposed cancer genes are associated with pathways of DNA-repair 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: (1) limited contribution of genetic heritability in certain cancer types; (2) low effect size/risk associated with each individual variant; (3) low-penetrance in view of individual's background[8], and (4) 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

[1]The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel. [2]Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel. ✉email: roni.rasnic@mail.huji.ac.il

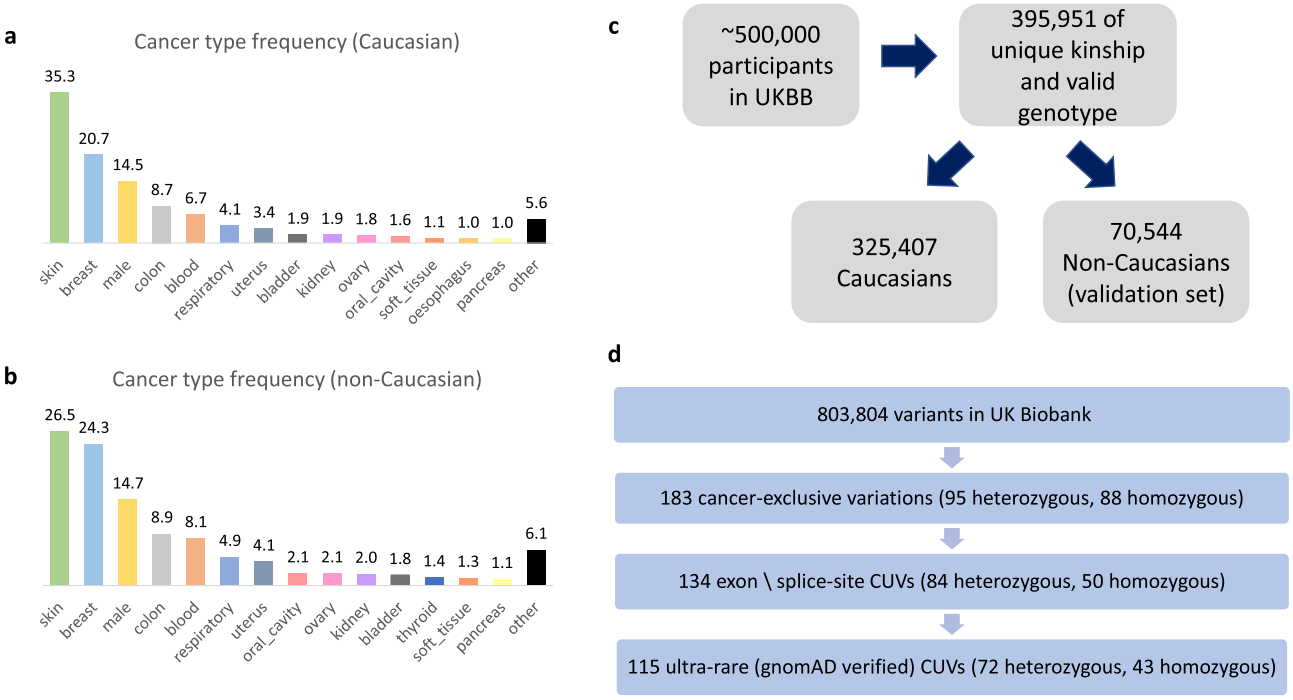

Figure 1. UK Biobank CUVs collection. The Caucasian filtered UK Biobank (UKBB) data set include 42,972 individuals who had cancer and the non-Caucasian include 6,959 such individuals. (**a**) Cancer type distribution for the Caucasian data set. (**b**) Cancer type distribution for the non-Caucasian data set. (**c**) The data of 395,951 UKBB participants was used for this study, 325,407 of which were confirmed Caucasian. (**d**) Out of 803,804 UKBB variants, we curated 72 heterozygous and 43 homozygous CUVs (total 115 CUVs).

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. Notably, the presented CUVs were extracted from UKBB DNA array and therefore only cover the array pre-selected 803,804 SNPs. 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 UKBB data set used in the article is comprised of 325,407 Caucasian UKBB participants (see Methods, Fig. 1c), 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 clinical 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 participants, among them 63,585 are cancer-free (90.1%). Figure 1a,b provide further details on different cancer type prevalence in these sets.

Non-melanoma skin cancer is mostly attributed to environmental factors rather than genetic association[24]. However, based on evidence for hereditary links for non-melanoma skin cancer predisposition[25,26], we included these individuals in our analysis. In addition, focusing on extremely rare variations enables the identification of existing, yet overlooked genetic associations.

Compilation of cancer-exclusive ultra-rare variants (CUVs). We scanned 803,804 genetic markers in our prime data set for cancer-exclusive variations. 183 variations met our initial criteria, appearing at least twice in individuals diagnosed with cancer and not appearing in cancer-free individuals. Among them, 95 were heterozygous and 88 were homozygous variations. In order to target variations with additional supporting evi-

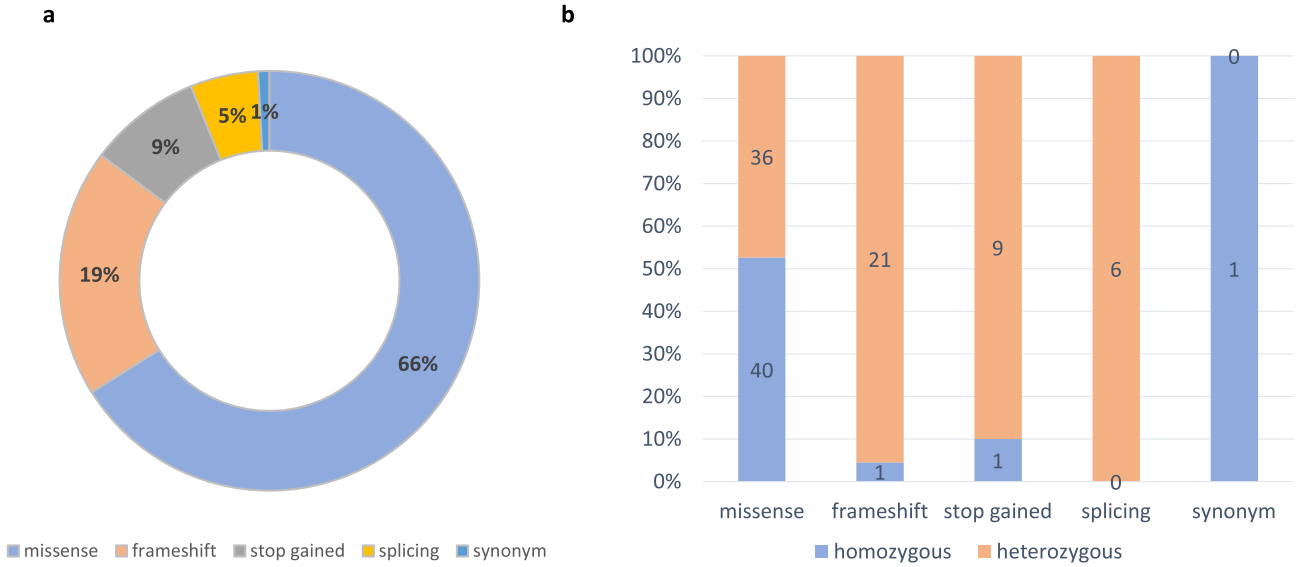

Figure 2. Exomic CUVs are mostly gene disruptive. The partition of variant types for the compiled list of 115 exomic CUVs. The list is dominated by transcript disruptive variations (99.1%) that include missense, frameshift, stop gain and splicing sites. (**a**) Distribution of variation types among the exomic CUVs. (**b**) Dispersion of variant types among heterozygous and homozygous CUVs.

dence, we considered only coding exome and splice-region variants. To assure the CUVs rarity in the general population, we applied an additional filter based on the gnomAD data set (see Methods). The resulting final list is comprised of 115 variants (associated with 108 genes), 72 heterozygous and 43 homozygous (Fig. 1d). The detailed list of all 115 CUVs can be found in Supplementary Table S2.

Most (66%) of the CUVs 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 variant set, but only 50% of the heterozygous CUVs. The heterozygous CUVs are highly enriched for LoF variants which constitute the other 50%.

Cancer-exclusive ultra-rare variants overlap with known cancer predisposition genes. From the listed CUVs, 26 variants were previously defined as cancer inducing genes (in 23 genes, 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 determined 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 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. Notably, 10 of the variants had at least one appearance in non-melanoma skin cancer.

The heterozygous CUVs are enriched for known cancer predisposition genes. Twenty-five of the cancer associated CUVs are heterozygous and one is homozygous. However, 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 DNA array. There are 6,450 heterozygous ultra-rare exome variants from 2,938 genes which pass our biobank-ethnic and the gnomAD allele frequency filtration. A total of 1,604 of the filtered ultra-rare variants overlap with 105 known CPGs, as some genes are over-represented among 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 collection of heterozygous ultra-rare variants. 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*). The statistical significance estimates (p-values) for each data-point are available in Supplemental Table S3 (see Methods).

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: (1) the filtered Caucasian UKBB cohort; (2) the matched filtered, non-Caucasian UKBB cohort; (3) 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. Ten out of the 23 novel CPGs were identified based on appearances in individuals with non-melanoma skin cancer.

Within the Caucasian cohort, we consider the following as additional genomic evidence: (1) a gene with 2 CUVs, or (2) 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

hg19	Effect	Ref	Alt	Gene	COSMIC	C299	CPG	Function[a]
1:155205517	Missense	T	C	*GBA*			Y	Enzyme
2:48027130	Missense	G	A	*MSH6*	Y	Y	Y	DNA repair
3:10183771	Missense	T	G	*VHL*	Y	Y	Y	Ubq-complex
3:30730003	Splice region	G	A	*TGFBR2**	Y	Y		Kinase
3:37048480	Splice region	A	G	*MLH1*	Y	Y	Y	TSG
5:112173671	Frameshift	AG	A	*APC*	Y	Y	Y	TSG
5:112175255	Frameshift	G	GA	*APC*	Y	Y	Y	TSG
9:101891277	Stop gain	C	T	*TGFBR1**			Y	Kinase
9:131341997	Missense	T	G	*SPTAN1*		Y		Cytoskeletal
10:43609079	Frameshift	TCCCTGAG	T	*RET*	Y	Y	Y	Kinase
10:88659605	Missense	T	C	*BMPR1A**	Y		Y	Kinase
10:89717630	Stop gain	C	T	*PTEN**	Y	Y	Y	TSG, Phosphatase
11:44193237	Missense	G	C	*EXT2**	Y		Y	TSG, Enzyme
11:71720337	Missense	C	A	*NUMA1**	Y			MT Spindle pole
11:108192066	Missense	A	C	*ATM**	Y	Y	Y	DDR, Kinase
13:32890621	Frameshift	GC	G	*BRCA2*	Y	Y	Y	TSG, DNA repair
13:32914296	Missense	A	G	*BRCA2*	Y	Y	Y	TSG, DNA repair
13:48878061	Frameshift	AC	A	*RB1*	Y	Y	Y	TSG
13:103524611	Frameshift	GA	A	*ERCC5**	Y		Y	DNA repair
16:2121553	Missense	C	G	*TSC2*	Y	Y	Y	TSG
17:29654601	Missense	G	T	*NF1**	Y	Y	Y	RAS regulator
17:41244383	Frameshift	GC	G	*BRCA1**	Y	Y	Y	TSG, DNA repair
17:41246296	Stop gain	C	A	*BRCA1*	Y	Y	Y	TSG, DNA repair
18:3451996	Frameshift	CG	C	*TGIF1*		Y		TGF ligand
21:36421256	Splice region	C	T	*RUNX1*	Y	Y	Y	TF
22:30067894	Missense	T	C	*NF2*	Y	Y	Y	Cytoskeletal

Table 1. CUVs overlap with known cancer predisposition or driver genes. [a]Function abbreviation: *DDR* DNA damage response, *TSG* tumor suppressor gene, *TF* transcription factor, *MT* microtubule, *Ubq* ubiquitin. *Variants with at least one appearance in non-melanoma skin cancer.

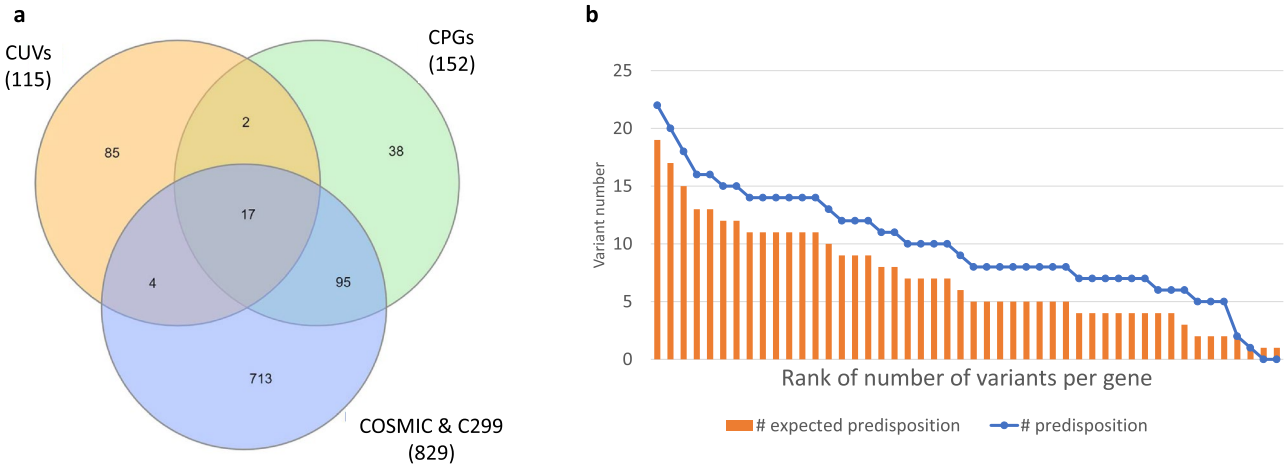

Figure 3. CUVs list is enriched with cancer predisposition genes. Out of the 108 genes in the CUVs list, 23 are known cancer genes. (**a**) Venn diagram of the genes associated with CUVs, known cancer driver genes (as reported in COSMIC) and the consensus CPGs. (**b**) Expected number of known CPG CUV (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.

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	39
AKR1C2	Hetero	2			Y	Exerts an inhibitory effect on oncogenesis	40
DNAH3	Homo	3			Y	Cancer predisposed genes in Tunisian family	27
*DSP**	Hetero	2	Y			Affects cell adhesion. Suppressed by TGF-β	
*EGFLAM**	Hetero	2			Y	Promotes matrix assembly	
*ENDOU**	Homo	3				Cancer biomarker	41
HIST1H2BO	Hetero	3				Affects major signaling pathways	
HSPB2	Hetero	2			Y	Epigenetically regulated	42
*ICAM1**	Homo	4				Biomarker, under a clinical trial	26
*ISLR**	Homo	2		Y		Marker for mesenchymal stem cells. Deregulated gene in cancer	43
KCNH2	Hetero	2	Y			Affects proliferation and migration	
MAP3K15	Hetero	2			Y	Contributes to cell migration	
MRPL39	Hetero	2			Y	Tumor suppressor by targeting miR-130	44
MYBPC3	Both	2	Y			Cytoskeletal modifier	
*MYO1E**	Homo	2		Y		Stimulates upregulation of motility and invasion	45
*NAV3**	Hetero	3				Acts as a suppressor of breast cancer	46
*PCDHB16**	Homo	3		Y	Y		
SARDH	Homo	2		Y		Acts as tumor suppressor	47
*SCN5A**	Hetero	2	Y			Promotes breast cancer, possess anti-pancreatic cancer	48
*WDFY4**	Hetero	2			Y	Presentats viral, tumor antigen on dendritic cells	49
ZFC3H1	Homo	2			Y	Indirect activating DNA repair	

Table 2. Novel validated CPG candidates. *Variants with at least one appearance in non-melanoma skin cancer.

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 *BMPR1A* 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 23 validated CPGs which includes 21 genes that are novel CPGs. Among these genes 12 CUVs are heterozygous, 8 are homozygous and *MYBPC3* is supported by 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[29]. Additionally, *DNAH3* was identified as novel predisposition gene using exome sequencing in a Tunisian family with multiple non-BRCA breast cancer instances[30].

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[31,32]) 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

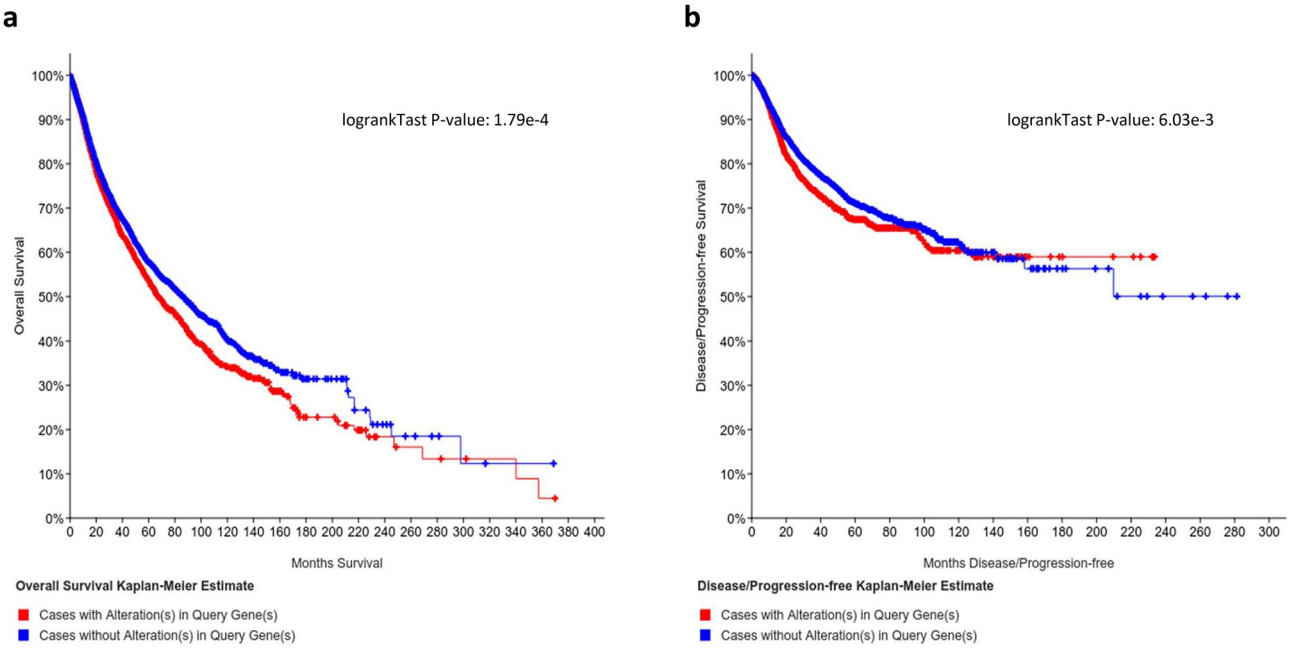

Figure 4. Somatic mutations in CPG candidate effect cancer patient survival and disease progression. The effect of somatic mutations in the 21 novel CPG candidate (Table 2) on the survival rate of TCGA cancer patients was tested via cBioPortal. (**a**) Meier–Kaplan survival rate estimate. (**b**) Meier–Kaplan disease/progression-free estimate.

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). Cancer types in this analysis are represented by varied number of patients and percentage of individuals with somatic mutations in any of the novel CPGs (Supplemental Table S4). The trend in most cancer types match the presented pan-cancer analysis. Survival and disease/progression estimate for each cancer type are available in Supplementary Figures S1–S24. Hazard Ratios and confidence intervals were calculated (see Materials and Methods and Supplemental Table S4).

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 26 variants (from 23 genes) are associated with known cancer genes. Most of these variants (22) 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[33,34]. It is likely that CPGs based on ultra-rare variants are not saturated. For example, additional CPGs including *CDKN2A* and *NF1* were associated with an increased risk for breast cancer[35]. Specifically, *CDKN2A* has been also detected as a CPG in families of patients with pancreatic cancer[36]. 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[37], membrane signaling, DNA repair systems and more.

Numerous CUVs are present in individuals with non-melanoma skin cancer. For the most part non-melanoma skin cancers are attributed to environmental factors. Nevertheless, studies show that there are in fact genetic components associated with the majority of non-melanoma skin cancers[25,26]. Accordingly, CUVs can unveil such rare genetic associations.

We chose to focus on cancer-exclusive variants to shed light on mostly overlooked ultra-rare cancer predisposition variants. Naturally, additional ultra-rare variants in the data-set are presumably cancer inducing. Detecting these variants requires developing a broader model expanding the scope to somewhat less rare, possibly lower-penetrance variants. The impending availability of UKBB exome sequencing (150,000 exomes), will enable us to revisit the identified variants, to further refine the list of candidate CPGs (i.e., removing false-positives and adding evidence to support true CPGs) and to develop a less strict detection model.

The inheritably rare nature of CUVs raise concerns on the reliability of their initial identification[38]. We overcome this hurdle by only considering 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 UKBB has recruited ~ 500,000 people from the general population of the UK, using National Health Service patient registers, with no exclusion criteria[39]. 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. We derived the kinship group from the familial information provided by the UKBB .fam files. Additionally, 312 samples had mismatching sex (between the self-reported and the genetics-derived) and 726 samples had only partial genotyping.

We divided the remaining 395,951 participants into two groups: (1) 'Caucasians'—individuals that were both genetically verified as Caucasians and declared themselves as 'white'. (2) 'non-Caucasians'—individuals not matching the previous criterion. The Caucasian cohort includes 325,407 individuals (42,972 of whom had cancer) and the non-Caucasian cohort includes 70,544 individuals (6,959 had cancer). We used the Caucasian cohort for our primary analysis and the non-Caucasian cohort for additional validation purposes.

Variant filtration pipeline. We considered a heterozygous variation as cancer-exclusive when there were at least 2 cancer patients exhibiting the variation and no healthy individuals with the variation in the Caucasian cohort. We considered a homozygous variation as cancer-exclusive when there were at least 2 cancer patients exhibiting the variation (i.e., homozygous to the alternative SNP) and no healthy individuals with the homozygous variation in the Caucasian cohort. The ensemble Variant effect predictor[40] was used to annotate the variants.

We applied two additional filtration steps for the exome/splicing-region variants. The first filter was applied using the 'non-Caucasian' data set, we filtered heterozygous variations with MAF > 0.01% and homozygous variations with homozygous frequency > 0.01% in this set. This filtration step is meant to diminish variations which are mostly ethnic artifacts. The second filter was applied to assure the variations rarity. We applied the same filter (heterozygous variations with MAF > 0.01% and homozygous variations with homozygous frequency > 0.01%), using gnomAD v2.1.1[41]. The used gnomAD threshold was based on the summation of gnomAD v2.1.1 exomes and genomes. We also used gnomAD for the TCGA-germline validation, by extracting TCGA appearances from the database.

Statistical analysis. The UKBB ultra-rare variants are enriched with CPGs variants. We accounted for this imbalance by calculating the expected number of cancer predisposed genes when gradually removing highly-represented genes from the ultra-rare variant collection for heterozygotes. We calculated p-values for each data-point using a two-side binomial test.

We downloaded survival data from cBioPortal. The data only included survival months. We used Cox regression without covariates to calculate Hazard Ratio and confidence intervals. The results are listed in Supplementary Table S4.

Rare variants reliability. Our CUV collection includes variants that appeared at least twice in the filtered Caucasian cohort, thereby evading many SNP-genotyping inaccuracies[38]. 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. All methods were performed in accordance with the relevant guidelines and regulations. UKBB 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).

UKBB received ethical approval from the NHS National Research Ethics Service North West (11/NW/0382). UKBB participants provided informed consent forms upon recruitment.

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.

Received: 6 February 2020; Accepted: 28 July 2020
Published online: 10 August 2020

References

1. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* https://doi.org/10.1038/nature12981 (2014).
2. Zhang, J. *et al.* Germline mutations in predisposition genes in pediatric cancer. *N. Engl. J. Med.* https://doi.org/10.1056/NEJMoa1508054 (2015).
3. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* https://doi.org/10.1016/j.cell.2011.02.013 (2011).
4. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat. Med.* https://doi.org/10.1038/nm1087 (2004).
5. Bertelsen, B. *et al.* High frequency of pathogenic germline variants within homologous recombination repair in patients with advanced cancer. *npj Genom. Med.* https://doi.org/10.1038/s41525-019-0087-6 (2019).
6. Easton, D. F. How many more breast cancer predisposition genes are there?. *Breast Cancer Res.* https://doi.org/10.1186/bcr6 (1999).
7. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* https://doi.org/10.1073/pnas.0903103106 (2009).
8. Galvan, A., Ioannidis, J. P. A. & Dragani, T. A. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends Genet.* https://doi.org/10.1016/j.tig.2009.12.008 (2010).
9. Baria, K., Warren, C., Roberts, S. A., West, C. M. & Scott, D. Chromosomal radiosensitivity as a marker of predisposition to common cancers?. *Br. J. Cancer* https://doi.org/10.1054/bjoc.2000.1701 (2001).
10. Hu, C. *et al.* Association between inherited germline mutations in cancer predisposition genes and risk of pancreatic cancer. *J. Am. Med. Assoc.* https://doi.org/10.1001/jama.2018.6228 (2018).
11. Verkasalo, P. K., Kaprio, J., Koskenvuo, M. & Pukkala, E. Genetic predisposition, environment and cancer incidence: a nationwide twin study in Finland, 1976–1995. *Int. J. Cancer* https://doi.org/10.1002/(SICI)1097-0215(19991210)83:6<743::AID-IJC8>3.0.CO;2-Q (1999).
12. Frank, S. A. Genetic predisposition to cancer—insights from population genetics. *Nat. Rev. Genet.* https://doi.org/10.1038/nrg1450 (2004).
13. Law, P. J. *et al.* Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat. Commun.* https://doi.org/10.1038/s41467-019-09775-w (2019).
14. Czene, K., Lichtenstein, P. & Hemminki, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int. J. Cancer* https://doi.org/10.1002/ijc.10332 (2002).
15. Economopoulou, P., Dimitriadis, G. & Psyrri, A. Beyond BRCA: new hereditary breast cancer susceptibility genes. *Cancer Treat. Rev.* https://doi.org/10.1016/j.ctrv.2014.10.008 (2015).
16. Grant, R. C. *et al.* Prevalence of germline mutations in cancer predisposition genes in patients with pancreatic cancer. *Gastroenterology* https://doi.org/10.1053/j.gastro.2014.11.042 (2015).
17. Petersen, G. M. *et al.* A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat. Genet.* https://doi.org/10.1038/ng.522 (2010).
18. Wolpin, B. M. *et al.* Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nat. Genet.* https://doi.org/10.1038/ng.3052 (2014).
19. Long, J. *et al.* Genome-wide association study in East Asians identifies novel susceptibility loci for breast cancer. *PLoS Genet.* https://doi.org/10.1371/journal.pgen.1002532 (2012).
20. Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.* https://doi.org/10.1038/ng.91 (2008).
21. Mancuso, N. *et al.* The contribution of rare variation to prostate cancer heritability. *Nat. Genet.* https://doi.org/10.1038/ng.3446 (2015).
22. Jiao, S. *et al.* Estimating the heritability of colorectal cancer. *Hum. Mol. Genet.* https://doi.org/10.1093/hmg/ddu087 (2014).
23. Huang, K.-L. *et al.* Pathogenic germline variants in 10,389 adult cancers. *Cell* https://doi.org/10.1016/j.cell.2018.03.039 (2018).
24. Griffin, L. L., Ali, F. R. & Lear, J. T. Non-melanoma skin cancer. *Clin. Med. J. R. Coll. Physicians Lond.* https://doi.org/10.7861/clinmedicine.16-1-62 (2016).
25. Nikolaou, V., Stratigos, A. J. & Tsao, H. Hereditary nonmelanoma skin cancer. *Semin. Cutan. Med. Surg.* https://doi.org/10.1016/j.sder.2012.08.005 (2012).
26. Roberts, M. R., Asgari, M. M. & Toland, A. E. Genome-wide association studies and polygenic risk scores for skin cancer: clinically useful yet?. *Br. J. Dermatol.* https://doi.org/10.1111/bjd.17917 (2019).
27. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gku1075 (2015).
28. Bailey, M. H. *et al.* Comprehensive characterization of cancer driver genes and mutations. *Cell* https://doi.org/10.1016/j.cell.2018.02.060 (2018).
29. Annels, N. E. *et al.* 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.* https://doi.org/10.1158/1078-0432.CCR-18-4022 (2019).
30. Hamdi, Y. *et al.* Family specific genetic predisposition to breast cancer: results from Tunisian whole exome sequenced breast cancer cases. *J. Transl. Med.* https://doi.org/10.1186/s12967-018-1504-9 (2018).
31. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* https://doi.org/10.1158/2159-8290.CD-12-0095 (2012).
32. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* https://doi.org/10.1126/scisignal.2004088 (2013).
33. Easton, D. F. *et al.* Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med.* https://doi.org/10.1056/NEJMsr1501341 (2015).
34. Couch, F. J. *et al.* Associations between cancer predisposition testing panel genes and breast cancer. *JAMA Oncol.* https://doi.org/10.1001/jamaoncol.2017.0424 (2017).

35. LaDuca, H. *et al.* 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.* https://doi.org/10.1038/s41436-019-0633-8 (2020).
36. Chaffee, K. G. *et al.* Prevalence of germ-line mutations in cancer genes among pancreatic cancer patients with a positive family history. *Genet. Med.* https://doi.org/10.1038/gim.2017.85 (2018).
37. Wang, Q. Cancer predisposition genes: molecular mechanisms and clinical impact on personalized cancer care: examples of lynch and HBOC syndromes. *Acta Pharmacol. Sin.* https://doi.org/10.1038/aps.2015.89 (2016).
38. Weedon, M. N. *et al.* Assessing the analytical validity of SNP-chips for detecting very rare pathogenic variants: implications for direct-to-consumer genetic testing. *bioRxiv* https://doi.org/10.1101/696799 (2019).
39. Sudlow, C. *et al.* UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* https://doi.org/10.1371/journal.pmed.1001779 (2015).
40. McLaren, W. *et al.* The ensemble variant effect predictor. *Genome Biol.* https://doi.org/10.1186/s13059-016-0974-4 (2016).
41. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* https://doi.org/10.1101/531210 (2019).
42. Yin, A. A. *et al.* Novel predictive epigenetic signature for temozolomide in non-G-CIMP glioblastomas. *Clin Epigenetics.* **11**, 76 (2019).
43. Martinez-Aguilar, J. *et al.* Proteomics of thyroid tumours provides new insights into their molecular composition and changes associated with malignancy. *Sci Rep.* **6**, 23660 (2016).
44. Yu, M. J. *et al.* Long noncoding RNA MRPL39 inhibits gastric cancer proliferation and progression by directly targeting miR-130. *Genet Test Mol Biomarkers*. **22**, 656–663 (2018).
45. Ouderkirk-Pecone, J. L. *et al.* Myosin 1e promotes breast cancer malignancy by enhancing tumor cell proliferation and stimulating tumor cell de-differentiation. *Oncotarget*. **7**, 46419–46432 (2016).
46. Cohen-Dvashi, H. *et al.* Navigator-3, a modulator of cell migration, may act as a suppressor of breast cancer progression. *EMBO Mol Med.* **7**, 299–314 (2015).
47. He, H. *et al.* Alteration of the tumor suppressor SARDH in sporadic colorectal cancer: A functional and transcriptome profiling-based study. *Mol Carcinog.* **58**, 957–966 (2019).
48. Mao, W. *et al.* The emerging role of voltage-gated sodium channels in tumor biology. *Front Oncol.* **9**, 124 (2019).
49. Theisen, D. J. *et al.* WDFY4 is required for cross-presentation in response to viral and tumor antigens. *Science.* **362**, 694–699 (2018).

Acknowledgements

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.

Author contributions

M.L., N.L. and R.R. designed and guided this research and wrote the manuscript. Data collection, processing and analysis were performed by R.R.. All coauthors contributed to the current version of the manuscript. All coauthors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-70494-0.

Correspondence and requests for materials should be addressed to R.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.