

Germline Susceptibility, Somatic Exposure Signatures, and Chromatin Accessibility as Drivers of Cancer-Type Divergence Among Carriers of the Same Predisposition Variant

Rithvik Jangam

rithjan51@gmail.com

researcher

Research Article

Keywords: hereditary cancer, BRCA2, homologous recombination, mutational signatures, SBS3, chromatin accessibility, TCGA

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11322040/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Background.

People who inherit a harmful mutation in a cancer gene like *BRCA2* or *ATM* have a higher risk of cancer, but carriers of the same gene often end up with completely different cancers. I wanted to know why. I tested two ideas: that the tissue's history of DNA damage (its mutational signatures) matters, and that how open and readable a gene's pathway is in each tissue (chromatin accessibility) matters.

Methods.

I used public data from The Cancer Genome Atlas (TCGA). In Phase 1, I started from 853 harmful inherited variants in 10,389 patients and built a catalog of genes whose carriers get more than one cancer type. I compared mutational signatures across each gene's cancer types (9,493 tumors) and compared pathway openness between carrier and non-carrier tissues using ENCODE normal tissue and TCGA tumor ATAC-seq data. Because the inherited-variant table has no patient IDs, in Phase 2 I used tumors where the same genes are broken by a truncating tumor mutation as a public stand-in for carriers. I called mutations from the TCGA MC3 file (9,818 tumors), fit COSMIC signatures to 7,991 tumors, and compared mutated and non-mutated tumors within each cancer type while adjusting for mutation burden, age, and sex. I then tested whether the effect of a broken homologous recombination (HR) gene depended on the tissue.

Results.

The final catalog had 44 gene–cancer pairs across 11 genes. In Phase 1, every gene's cancer types had strongly different exposure fingerprints (average AUC 0.81), and the *BRCA2* HR pathway was more open in *BRCA2* carrier tissues in both normal tissue (+ 1.67) and tumors (+ 3.26). In Phase 2, breast tumors with a broken *BRCA2* had much more of the HR-failure signature SBS3 (median 56% of mutations vs. 12%; adjusted $p = 3.6 \times 10^{-6}$), and so did breast tumors with a broken *BRCA1* (44% vs. 12%; $p = 4.9 \times 10^{-4}$). These were the only two of 108 tests that passed correction for multiple testing. The effect depended strongly on tissue. When *BRCA1*, *BRCA2*, *PALB2*, or *BRIP1* was broken, tumors from the classic carrier tissues (breast, ovary, prostate, pancreas) had about 18 percentage points more SBS3, while tumors from all other tissues had essentially none (+ 0.9 points; interaction $p = 2.8 \times 10^{-9}$ after removing hypermutated tumors). For *BRCA2* alone, the gap was + 26 points vs. 0 (interaction $p = 8.6 \times 10^{-10}$). Predicted links for *ATM* and smoking and for *PMS2* and mismatch repair pointed the right way but were not significant.

Conclusions.

Breaking the same repair gene has very different effects depending on the tissue. Loss of *BRCA2* leaves a strong HR-failure scar mainly in the tissues where *BRCA2* carriers actually get cancer, which are also tissues where Phase 1 found the HR pathway to be more open. This supports the idea that a tissue's own biology helps decide where an inherited mutation turns into cancer. The next step is to repeat this in true inherited carriers using controlled-access data.

1. Introduction

Some cancers are partly caused by harmful mutations that people inherit from their parents. These are called germline variants, and they are present in every cell of the body from birth. A large study of 10,389 patients in The Cancer Genome Atlas (TCGA) found harmful inherited variants in about 8% of cases [1]. Most of these variants are in tumor suppressor genes, which normally help prevent cancer. According to Knudson's two-hit model, an inherited mutation acts as a first hit, and cancer develops after the second copy of the gene is damaged later in life [2].

One puzzling thing about inherited cancer syndromes is how different the outcomes can be. People with Li-Fraumeni syndrome, caused by mutations in *TP53*, can develop breast cancer, sarcomas, brain tumors, adrenal tumors, or leukemia, often at young ages [3]. Lynch syndrome, caused by mutations in DNA mismatch repair genes such as *MLH1*, *MSH2*, *MSH6*, *PMS2*, and *EPCAM*, mostly raises the risk of colorectal and endometrial cancer, but can also lead to other cancers [4]. People with *BRCA1* or *BRCA2* mutations have high but different risks of breast and ovarian cancer [5]. This means that two people with the same mutation can end up with completely different cancers, so the mutation alone cannot fully explain which cancer develops.

I focused on two possible explanations. The first is exposure. Over a person's lifetime, things like tobacco smoke, ultraviolet (UV) light, and natural processes in the body leave specific patterns of mutations in DNA, called mutational signatures [6]. For example, SBS4 is linked to smoking, SBS7a to UV light, SBS2 and SBS13 to a group of enzymes called APOBEC, SBS1 to aging, and SBS6, SBS15, SBS20, SBS21, SBS26, and SBS44 to broken mismatch repair. If an inherited mutation makes a person more vulnerable, the tissue that builds up the most exposure damage might be the one that becomes cancerous.

The second explanation is chromatin accessibility. DNA is tightly wrapped around proteins, and only the unwrapped, or open, regions can be read by the cell. Which regions are open is different in each tissue, and this affects where mutations build up [7]. A lab method called ATAC-seq measures which parts of the DNA are open [8]. If the genes in a predisposition gene's pathway are more open in a certain tissue, that tissue might rely on the pathway more and be hit harder when the gene is broken. Public ATAC-seq data from normal tissues (ENCODE) [9] and from tumors (TCGA) [10] make it possible to test this idea.

Past studies have mostly looked at these two explanations separately. As far as I know, no study has tested both at the same time in carriers of the same genes to see which one better explains the cancer type they develop.

There is also one very specific clue about how a broken repair gene shows up in a tumor. When *BRCA1* or *BRCA2* stops working, cells lose homologous recombination (HR), their most accurate way of fixing breaks in both strands of DNA. They fall back on sloppier repair, and this leaves a flat, spread-out mutation pattern called SBS3 [6, 23, 24]. SBS3 is basically a scar that says “HR was broken here.” That makes it a useful tool for my question: if I can see where the scar shows up and where it doesn't, I can see which tissues are actually affected when an HR gene is lost. A clinical study of tumors from *BRCA1/2* carriers also found that these genes seem to matter more in some tumor types than others [25], which made me want to test this directly with public data and connect it to chromatin.

My hypothesis was that the cancer type a carrier develops depends on more than the inherited mutation, and that exposure signatures and tissue accessibility both play a role. I did the project in two phases. **Phase 1** had four aims: (1) build a reliable list of genes whose carriers develop more than one cancer type; (2) test whether mutational signatures differ across those cancer types; (3) test whether DNA openness is higher in the tissues where carriers get cancer; and (4) compare the two explanations. **Phase 2** asked the question more directly: within the same cancer type, do tumors with one of these genes broken carry different signatures than tumors without it, and does that depend on the tissue?

2. Methods

This project had two phases. Phase 1 (Sections 2.1–2.5) looks at carriers' cancer types as a group. Phase 2 (Sections 2.6–2.7) looks at individual tumors, because I found out partway through that the public inherited-variant table can't be linked to specific patients.

2.1 Data sources

All datasets I used are listed in Table 1. Phase 1 chromatin analyses used the GRCh38 human reference genome. Phase 2 used GRCh37 (hg19), because that is the build the MC3 mutation file was made with.

Table 1
Data sources.

Dataset	What it contains	Used for
Huang et al. 2018, Table S2A [1]	853 harmful inherited variants across 33 TCGA cancer types (no patient IDs)	Aim 1
ClinVar, downloaded September 2026 [11]	Variant classifications from labs around the world (9,048,962 records)	Aim 1 check
TCGA mutational signature activities [6]	Amount of each of 65 signatures in 9,493 tumors	Aims 2 and 4
ENCODE ATAC-seq [9]	Open DNA regions in 26 normal tissues (3,023,634 peaks)	Aim 3 (main)
Corces et al. 2018 [10]	Open DNA regions in 23 TCGA cancer types (902,507 peaks)	Aim 3 (extra)
MSigDB and Reactome [12, 13]	Lists of genes in biological pathways	Aim 3
TxDb.Hsapiens.UCSC.hg38.knownGene	Locations of genes in the genome	Aim 3
TCGA MC3 public mutation file v0.2.8 [19]	3,600,963 tumor (somatic) mutations with patient barcodes	Phase 2
TCGA Clinical Data Resource [20]	Cancer type, age, and sex for each patient	Phase 2
COSMIC signatures v3.2, GRCh37 [21, 22]	Reference patterns for each mutational signature	Phase 2

2.2 Aim 1: Building the carrier list

I started with Table S2A from Huang et al. [1], which lists 853 inherited variants labeled as pathogenic or likely pathogenic using standard guidelines [14]. I kept only variants marked Pathogenic or Likely Pathogenic, which kept all 853. The public table does not include patient IDs, so I counted each variant record as one carrier. I counted carriers for each gene and cancer type, kept gene–cancer pairs with at least three carriers, and then kept only genes with at least two qualifying cancer types. I needed at least two cancer types per gene because my question is about why carriers of the same gene get different cancers. To test how sensitive this choice was, I also repeated the process with a stricter cutoff of five carriers.

2.3 Checking variants in ClinVar

To double-check the variants, I used ClinVar [11], a public database of variant classifications. I kept ClinVar records for my 11 genes that were labeled Pathogenic, Likely pathogenic, or Pathogenic/Likely pathogenic. Since ClinVar and Huang et al. write protein changes differently, I converted ClinVar's format to match (for example, changing “p.Arg1970Ter” to “p.R1970*”) and then matched variants by gene and protein change.

2.4 Aim 2: Comparing mutational signatures

I used a file listing how much of each mutational signature was found in 9,493 TCGA tumors. I matched the cancer type names in this file to TCGA codes (Supplementary Table S1). Some cancer types, including LGG, did not have a match and were left out. For each gene, I selected all tumors from that gene's carrier cancer types. Because I did not have patient IDs linking carriers to their tumors, this included all TCGA tumors of those types, not just carriers' tumors.

I focused on six signatures chosen ahead of time: SBS1 (aging), SBS2 and SBS13 (APOBEC), SBS4 (smoking), SBS6 (mismatch repair), and SBS7a (UV). For each gene and signature, I used a Kruskal–Wallis test [15] to check whether the signature differed across that gene's cancer types. Since I ran 66 tests, I corrected the p-values using the Benjamini–Hochberg method [16] to reduce false positives. I also built a logistic regression model for each gene that tried to tell apart the gene's most common carrier cancer type from its other cancer types using the six signatures. I measured how well it worked using AUC, where 0.5 means random guessing and 1.0 means perfect [17]. The models were tested on the same data used to build them.

2.5 Aim 3: Comparing DNA openness

I downloaded ENCODE ATAC-seq data for normal human tissues using the ENCODE website's programming interface, using one high-quality peak file per tissue (23 tissues). For cancer types without a matching tissue, I added three stand-in tissues: foreskin skin cells for melanoma (SKCM), a brain region called the posterior cingulate cortex for brain tumors (LGG and GBM), and psoas muscle for sarcoma (SARC). I matched each cancer type to its tissue of origin (Supplementary Table S1).

For each gene, I chose a list of related genes in the same biological pathway from the Reactome database [12, 13] (Supplementary Table S2). I then looked at the promoter of each pathway gene, which is the region right before a gene where the cell starts reading it. For each gene and tissue, I found the open DNA peaks that overlapped these promoters and calculated their average fold enrichment, a measure of how open the region is. I labeled each tissue as a carrier tissue if it matched one of the gene's carrier cancer types, and a non-carrier tissue otherwise, and compared the two groups using a Wilcoxon rank-sum test. I repeated the same analysis with the Corces et al. tumor data [10].

Testing different versions. I did not decide on one exact method before starting, so I ended up testing eight versions of this analysis. These included using only each gene's own promoter, using a mixed group of pathway lists, adjusting for each tissue's overall openness, using tumor data, combining normal and tumor data, and using Reactome pathways with three different promoter window sizes (2 kb, 5 kb, and 10 kb). I report all eight. Testing many versions raises the chance of finding a significant result just by luck [18], so I used a stricter cutoff of 0.05 divided by 8, or 0.00625, to decide whether a result was truly reliable. I also tried a weighted version that copied rows based on the number of carriers, but I removed it because copying rows makes the sample look bigger than it really is, which is not a valid method.

2.6 Phase 2: Tumors with the gene broken vs. tumors without it

Why I needed a stand-in. The obvious next step after Phase 1 was to compare carriers' own tumors to non-carriers' tumors with the same cancer type. When I opened the Huang et al. table to find patient IDs, I found that none of its four sheets had any, because inherited DNA can identify people and the table was de-identified on purpose. Getting the linked data requires TCGA controlled access, which has to go through a university researcher and takes weeks. So instead, I used tumors where the same gene was broken by a mutation in the tumor itself (a somatic mutation). These mutations are public and come with patient barcodes. This isn't the same thing as being an inherited carrier, but it asks a very similar biological question: what happens to a tumor when this gene stops working?

Mutation data. I downloaded the public MC3 mutation file [19], which has 3,600,963 mutations called from TCGA tumor exomes. I kept mutations that passed quality filters (FILTER = PASS, plus the “wga” flags, because most ovarian samples were whole-genome amplified and would otherwise be lost), and I kept one primary tumor per patient. That left 2,875,946 mutations in 9,818 tumors. I got each patient's cancer type, age at diagnosis, and sex from the TCGA Clinical Data Resource [20]. I defined each tumor's mutation burden as its total number of mutations.

Measuring signatures in each tumor. For the 2,714,937 single-letter mutations, I looked up the DNA letter on each side using the hg19 reference genome, which puts every mutation into one of 96 standard categories (for example, a C turning into a T with an A before it and a G after it). Because C > T on one strand is the same as G > A on the other, I flipped every mutation so the changed letter was always C or T, which is the normal convention. I then used the MutationalPatterns R package [22] to figure out how much each of 17 COSMIC signatures [21] contributed to each tumor's mutations. I picked these 17 because they are the ones commonly found in TCGA exomes, and using fewer signatures makes the fit less likely to make things up. I only did this for tumors with at least 30 single-letter mutations (7,991 tumors), since fewer than that isn't enough to fit reliably. For each tumor, I turned the results into proportions (what fraction of its mutations came from each signature). I combined some related signatures: APOBEC = SBS2 + SBS13, UV = SBS7a + SBS7b, and mismatch repair (MMR) = SBS6 + SBS15 + SBS26 + SBS44. SBS3 (HR failure), SBS4 (smoking), and SBS1 (aging) were used on their own.

Which tumors count as “mutated.” For each of the 11 catalog genes, I called a tumor mutated if it had a truncating mutation in that gene: nonsense, frameshift, splice site, start-loss, or stop-loss. These usually break the protein, which is closest to what the inherited variants do. As a sensitivity check, I also re-ran everything counting missense mutations too, even though many missense changes don't actually break the gene.

Tests. For each gene–cancer pair in the Aim 1 catalog, I compared mutated and non-mutated tumors of that cancer type on six signatures (SBS1, APOBEC, SBS3, SBS4, UV, and MMR). I only tested pairs with at least three mutated and three non-mutated tumors. I used two tests: a Wilcoxon rank-sum test, and a

linear regression of the signature proportion on mutation status while adjusting for log mutation burden, age, and sex (sex was dropped for single-sex cancers). The adjustment matters a lot. Tumors with huge numbers of mutations pick up mutations in almost every gene just by chance, so without it, “has a *BRCA2* mutation” could just mean “has a lot of mutations.” I treated the adjusted model as my main result and corrected all p-values with the Benjamini–Hochberg method [16].

Predictions made ahead of time. Before running Phase 2, I wrote down 11 specific predictions based on the biology: tumors with a broken *BRCA1*, *BRCA2*, *PALB2*, or *BRIP1* should have more SBS3 in breast or ovarian cancer; tumors with a broken *ATM* or *CHEK2* should have more smoking damage (SBS4) in lung cancer; and tumors with a broken *PMS2* should have more mismatch-repair damage in endometrial or prostate cancer. I corrected these 11 separately from the full exploratory set.

2.7 Phase 2 follow-up: Is the effect tissue-specific?

Phase 2 showed a strong SBS3 effect in breast cancer, so I ran a follow-up to test the part of my hypothesis that is about tissues. I made one group of HR genes (a tumor counted as HR-mutated if it had a truncating mutation in *BRCA1*, *BRCA2*, *PALB2*, or *BRIP1*), and also looked at *BRCA1* and *BRCA2* separately. I defined the carrier tissues as breast (BRCA), ovary (OV), prostate (PRAD), and pancreas (PAAD), the four cancers most strongly linked to inherited *BRCA1/2* variants in clinical genetics [5, 25]. I chose these four before running the follow-up, but after seeing the Phase 2 results, so this part is a follow-up test and not a fully pre-planned one.

I ran three analyses. (A) For each cancer type with at least three mutated tumors, I estimated how much extra SBS3 the mutated tumors had, using the same adjusted regression as before. (B) Using all tumors together, I fit one model: $\text{SBS3} \sim \text{mutated} \times \text{carrier tissue} + \text{log mutation burden} + \text{age} + \text{cancer type}$. The mutated $\times$ carrier tissue term (the interaction) tests whether breaking the gene has a bigger effect in carrier tissues than in other tissues. Including cancer type as a variable means each tumor is only compared to others of its own type. (C) I pooled breast and ovarian tumors and estimated the overall effect, adjusting for cancer type.

Hypermutated tumors. Some tumors, especially stomach, colon, and endometrial cancers with broken mismatch repair, have thousands of mutations. In these tumors, a truncating mutation in *BRCA2* is often a random “passenger” mutation that doesn't actually knock out the gene. To make sure these weren't driving the results, I repeated all three analyses after removing tumors with 500 or more mutations. I report both versions.

2.8 Aim 4 status

My original plan for Aim 4 was to build models that predict cancer type using signatures alone, openness alone, and both together. The Phase 1 openness data could only be measured per gene and tissue, not per tumor, so it had the same value for every tumor in a model and couldn't be compared fairly. Phase 2 answers a related question in a different way, by testing whether the signature effect of a broken gene depends on the tissue. The full head-to-head model is described in Future Directions.

2.9 Software

I did all analyses in R version 4.6.1. Phase 1 used readxl, dplyr, httr, jsonlite, msigdb, TxDb.Hsapiens.UCSC.hg38.knownGene, org.Hs.eg.db, and pROC. Phase 2 used data.table, readxl, Biostrings, BSgenome.Hsapiens.UCSC.hg19, MutationalPatterns 3.22.0 [22], and ggplot2. Figures were made with ggplot2, pheatmap, and matplotlib.

3. Results

3.1 The final list of 11 genes

From the 853 variants, I found 507 different gene–cancer pairs. Keeping pairs with at least three carriers left 57 pairs across 24 genes. Keeping only genes with at least two qualifying cancer types left a final list of 44 gene–cancer pairs across 11 genes (Fig. 1; Table 2). *BRCA1* and *BRCA2* had the most carriers (70 each), mostly in ovarian and breast cancer, while *ATM* and *BRCA2* each appeared in eight different cancer types. When I used the stricter cutoff of five carriers, only 11 pairs across four genes remained (*ATM*, *BRCA1*, *BRCA2*, *NF1*), so I kept the cutoff of three to include more genes.

TP53, the gene behind Li–Fraumeni syndrome, did not make the list. Its 13 carriers were spread across eight cancer types, and only sarcoma reached three carriers (SARC 3; GBM, LGG, and OV 2 each; BRCA, KIRC, LUAD, and UCEC 1 each). This likely happened because TCGA mostly includes adult patients, while Li–Fraumeni cancers often appear at younger ages.

Table 2
Final carrier list. Numbers show carriers per cancer type (TCGA codes).

Gene	Carrier cancer types (number of carriers)	Total
<i>ATM</i>	BRCA 9, PRAD 7, STAD 7, LUAD 6, BLCA 4, PAAD 4, LGG 3, SKCM 3	43
<i>BRCA1</i>	OV 36, BRCA 21, UCEC 4, CESC 3, PRAD 3, STAD 3	70
<i>BRCA2</i>	OV 27, BRCA 20, PAAD 5, SKCM 4, STAD 4, THCA 4, LUSC 3, SARC 3	70
<i>BRIP1</i>	BRCA 4, LGG 3, OV 3, STAD 3	13
<i>BUB1B</i>	GBM 3, LUSC 3	6
<i>CHEK2</i>	BLCA 4, BRCA 3, LUAD 3, SKCM 3	13
<i>GJB2</i>	LUAD 4, LIHC 3, STAD 3	10
<i>NF1</i>	PCPG 5, SARC 5	10
<i>PALB2</i>	STAD 5, BRCA 3, COAD 3	11
<i>PMS2</i>	PRAD 3, UCEC 3	6
<i>VHL</i>	PCPG 6, KIRC 4	10

3.2 ClinVar check

ClinVar had between 80 (*BUB1B*) and 5,743 (*BRCA2*) harmful variant records for each of my genes. Of the 753 variants with protein change information, 137 matched a harmful ClinVar record (103 Pathogenic, 31 Pathogenic/Likely pathogenic, and 3 Likely pathogenic). Since I only searched ClinVar's harmful records, this confirms those 137 variants, but it does not show whether ClinVar disagrees about the others. A variant that ClinVar labels differently simply would not show up as a match.

3.3 Every gene's cancer types have their own exposure fingerprint

This was the strongest result in the whole project. For every one of the 11 genes, the mutational signatures were very different across the cancer types its carriers develop. For example, among *ATM*'s cancer types, the smoking signature (SBS4) was high in lung cancer, the UV signature (SBS7a) was high in melanoma, the APOBEC signatures (SBS2 and SBS13) were high in bladder cancer, and the aging signature (SBS1) was high in stomach cancer (Table 3). The Kruskal–Wallis tests were significant for every gene after correction. The strongest result for each gene ranged from SBS1 in *VHL* (adjusted $p = 2.2 \times 10^{-26}$) to SBS2 and SBS7a in *ATM*, *BRCA2*, and *CHEK2*, where the p-values were too small for R to calculate (Table 4). The signature models could tell each gene's most common cancer type apart from its other cancer types with an average AUC of 0.81 (range 0.64–0.99; Table 5).

ATM is the best example of how big this effect is. Its carriers show up in eight different cancer types, and those eight tissues were shaped by almost completely different exposures: smoking in lung, UV in skin, APOBEC in bladder, and aging in stomach. So even though every *ATM* carrier starts with the same broken gene, the tissues where their cancers show up have lived through totally different kinds of DNA damage. This fits the idea that a carrier's inherited mutation sets the risk, but the tissue's exposure history helps decide where that risk turns into cancer. One thing to keep in mind is that these comparisons used all TCGA tumors of each cancer type, so the next step is to check that carriers' own tumors follow the same pattern (Section 3.5).

Table 3

Average number of mutations per tumor from each signature, across *ATM*'s carrier cancer types (all TCGA tumors of each type).

Cancer type	SBS1 (aging)	SBS2 (APOBEC)	SBS4 (smoking)	SBS6 (repair)	SBS7a (UV)	SBS13 (APOBEC)
BLCA	18.6	104	0	0	0	101
BRCA	11.8	18.2	0	0.02	0.10	18.3
LUAD	8.2	19.5	164	0.60	0.08	21.9
PAAD	17.3	0.58	0	46.2	0.40	0.66
PRAD	12.2	0.11	0	14.8	0.12	0.05
SKCM	9.9	0.08	0	0	425	0.06
STAD	115	2.76	0	13.7	0	2.29

Table 4

Strongest signature difference for each gene (Kruskal–Wallis test, corrected p-values).

Gene	Signature	Adjusted p	Gene	Signature	Adjusted p
<i>ATM</i>	SBS2	$< 10^{-300}$	<i>GJB2</i>	SBS1	5.0×10^{-135}
<i>BRCA1</i>	SBS1	2.8×10^{-288}	<i>NF1</i>	SBS1	8.8×10^{-37}
<i>BRCA2</i>	SBS7a	$< 10^{-300}$	<i>PALB2</i>	SBS1	2.3×10^{-234}
<i>BRIP1</i>	SBS1	1.1×10^{-138}	<i>PMS2</i>	SBS1	8.0×10^{-102}
<i>BUB1B</i>	SBS1	1.2×10^{-96}	<i>VHL</i>	SBS1	2.2×10^{-26}
<i>CHEK2</i>	SBS7a	$< 10^{-300}$			

Table 5

How well signatures could tell each gene's most common cancer type apart from its other cancer types.

Gene	Main type	Tumors (n)	AUC	Gene	Main type	Tumors (n)	AUC
<i>ATM</i>	BRCA	3,318	0.715	<i>GJB2</i>	LUAD	1,240	0.823
<i>BRCA1</i>	OV	2,945	0.756	<i>NF1</i>	PCPG*	390	0.890
<i>BRCA2</i>	OV	3,415	0.699	<i>PALB2</i>	STAD	1,827	0.637
<i>BRIP1</i>	BRCA	1,716	0.800	<i>PMS2</i>	PRAD	958	0.931
<i>BUB1B</i>	GBM	789	0.989	<i>VHL</i>	PCPG	516	0.800
<i>CHEK2</i>	BLCA	2,259	0.903	Average			0.813

**NF1* had the same number of carriers in PCPG and SARC, so PCPG was used. These AUCs were measured on the same data used to build the models, so they are likely higher than they would be on new data.

3.4 BRCA2 shows the clearest chromatin link

The most interesting chromatin result came from *BRCA2*. Its repair pathway, homologous recombination, was more open in the tissues where *BRCA2* carriers got cancer than in the tissues where they didn't. This showed up in normal tissue from ENCODE (+ 1.67, the biggest difference of any gene) and again, even stronger, in actual tumors from TCGA (+ 3.26). Seeing the same direction in two completely separate datasets makes this the most convincing chromatin link I found (Fig. 3).

Across all 11 genes together, the picture was more mixed. When I looked only at each gene's own promoter, just 7 of 11 genes had any open DNA peaks, and only three (*BRCA1*, *BRCA2*, *CHEK2*) had data for both carrier and non-carrier tissues. There was no difference between the groups ($p = 0.99$). Looking at whole pathways gave me data for all 11 genes, but the results changed a lot depending on the version (Fig. 2). The three Reactome versions showed higher openness in carrier tissues ($p = 0.017$ – 0.026), but the mixed pathway version ($p = 0.090$), the tissue-adjusted version ($p = 0.56$), the tumor data ($p = 0.82$), and the combined data ($p = 0.12$) did not. None of the eight versions passed the stricter cutoff, so I treat the all-gene openness effect as a lead rather than a proven result. Showing all eight versions side by side (Fig. 2) turned out to be a finding of its own: it shows how much a chromatin result can depend on which pathway list and which tissues you pick.

Looking gene by gene explains why (Fig. 3A). In the Reactome 2 kb version, seven of the 11 genes had more open pathways in carrier tissues. The biggest differences were in *BRCA2* (+ 1.67), *GJB2* (+ 0.82), *CHEK2* (+ 0.78), and *ATM* (+ 0.68). *PMS2* (– 1.00), *BUB1B* (– 0.34), *VHL* (– 0.31), and *BRCA1* (– 0.16) went the opposite way. In the tumor data (Fig. 3B), *BRCA2* was again higher in carrier tissues (+ 3.26), while *CHEK2* and *BRCA1* were lower. *NF1* and *PMS2* did not have enough data. So chromatin openness seems to matter a lot for some genes (especially *BRCA2*) and not for others, which is why averaging all 11 genes together weakens the signal.

3.5 Phase 2: Breaking BRCA1 or BRCA2 leaves the HR-failure scar in breast cancer

After cleaning, the MC3 file had 9,818 tumors, and 7,991 of them had enough mutations to fit signatures. Many of my catalog pairs turned out to have very few tumors with a truncating mutation in the gene (Supplementary Table S3). For example, *GJB2* had none in any of its cancer types, and *BRIP1*, *BUB1B*, and *CHEK2* had at most two per cancer type. That left 108 testable gene–cancer–signature combinations.

Two of my 11 predictions came out clearly significant (Table 6; Fig. 4). Breast tumors with a broken *BRCA2* had a median of 56% of their mutations coming from SBS3, compared to 12% in other breast tumors. After adjusting for mutation burden, age, and sex, the mutated tumors had about 29 percentage

points more SBS3 ($p = 3.6 \times 10^{-6}$; corrected $p = 2.5 \times 10^{-5}$). Breast tumors with a broken *BRCA1* showed the same thing: a median of 44% vs. 12%, about 19 points more after adjustment ($p = 4.9 \times 10^{-4}$; corrected $p = 0.0017$). These were also the only two results out of all 108 exploratory tests that passed correction (corrected $p = 3.8 \times 10^{-4}$ and 0.027). Looking at Fig. 4, 9 of the 10 *BRCA2*-mutated breast tumors had more SBS3 than the typical breast tumor, and most of them were above 40%.

In ovarian cancer, both genes went the same direction but the difference was smaller and not significant (*BRCA1*: +9 points, $p = 0.058$; *BRCA2*: +12 points, $p = 0.080$). I think this makes sense biologically, because ovarian cancers in TCGA are mostly high-grade serous tumors, and about half of those already have HR problems from other causes [29]. The “not mutated” ovarian tumors already had a median of 42% SBS3, so there was less room for the mutated ones to stand out (Supplementary Figure S3).

Table 6

The 11 predictions made before running Phase 2 (truncating mutations; adjusted for mutation burden, age, and sex).

Gene	Cancer	Signature	Mutated / not	Median share (mutated vs. not)	Adjusted difference	Adjusted p	Corrected p
<i>BRCA2</i>	BRCA	SBS3	10 / 850	0.56 vs. 0.12	+ 0.29	3.6×10^{-6}	2.5×10^{-5}
<i>BRCA1</i>	BRCA	SBS3	12 / 848	0.44 vs. 0.12	+ 0.19	4.9×10^{-4}	0.0017
<i>BRCA1</i>	OV	SBS3	16 / 365	0.50 vs. 0.43	+ 0.09	0.058	0.14
<i>BRCA2</i>	OV	SBS3	7 / 374	0.56 vs. 0.42	+ 0.12	0.080	0.14
<i>ATM</i>	LUAD	SBS4	17 / 478	0.59 vs. 0.48	+ 0.05	0.27	0.37
<i>ATM</i>	LUSC	SBS4	6 / 471	0.41 vs. 0.35	+ 0.00	1.00	1.00
<i>PMS2</i>	UCEC	MMR	12 / 514	0.58 vs. 0.26	+ 0.04	0.60	0.70
<i>BRIP1</i>	OV	SBS3	0 / 381	not testable			
<i>CHEK2</i>	LUAD	SBS4	0 / 495	not testable			
<i>PALB2</i>	BRCA	SBS3	1 / 859	not testable			
<i>PMS2</i>	PRAD	MMR	0 / 344	not testable			

“Corrected p” uses Benjamini–Hochberg across the testable predictions. A prediction was not testable when fewer than three tumors of that cancer type had a truncating mutation in the gene.

3.6 The other predictions: right direction, not enough tumors

The *ATM* and smoking prediction pointed the way I expected. Lung adenocarcinomas with a broken *ATM* had a median of 59% of their mutations from smoking (SBS4), compared to 48% in other lung adenocarcinomas (Supplementary Figure S3). But once I adjusted for mutation burden, the difference shrank to 5 points and wasn't significant ($p = 0.27$). This is probably because heavy smokers have more mutations overall, so they are more likely to have an *ATM* mutation by chance, and the adjustment takes that into account. In lung squamous cell cancer, there were only six *ATM*-mutated tumors and no difference.

PMS2 is an interesting case. Endometrial tumors with a broken *PMS2* had much more mismatch-repair damage (median 58% vs. 26%), and the simple Wilcoxon test was significant ($p = 0.0099$). But after adjusting for mutation burden, the effect disappeared ($p = 0.60$). I think the adjustment is actually the wrong choice for this one gene. When mismatch repair breaks, the tumor ends up with a huge number of mutations, so mutation burden is a result of losing *PMS2* and not a separate factor. Adjusting for it basically removes the effect I'm trying to measure. So I'd call this result suggestive, not proven either way. The other four predictions (*BRIP1*, *CHEK2*, *PALB2* in breast, and *PMS2* in prostate) couldn't be tested because there were zero or one mutated tumors.

I also noticed that many of the unadjusted Wilcoxon tests were significant in stomach and endometrial cancer (for example, *BRCA2* in stomach cancer and higher MMR damage), but none of them held up after adjusting for mutation burden. Those are the cancer types with the most hypermutated tumors, which is exactly the problem the adjustment is designed to catch. This is why I used the adjusted model as my main result.

3.7 The SBS3 effect depends on the tissue

This was the most important result of Phase 2. When I looked at every cancer type with at least three HR-mutated tumors, the extra SBS3 showed up in the carrier tissues and mostly nowhere else (Fig. 5; Supplementary Figures S1 and S2). With hypermutated tumors removed, HR-mutated tumors had more SBS3 in breast (+ 22 points, $p = 3.6 \times 10^{-8}$), prostate (+ 20 points, $p = 0.004$), and ovarian cancer (+ 9 points, $p = 0.017$). Pancreatic cancer had too few mutated tumors to test. In the other tissues, like lung, bladder, colon, endometrium, and head and neck, the effect was close to zero, even though several of those had more mutated tumors than prostate did. The one other tissue with a clear effect was cervical cancer (+ 14 points, $p = 1.1 \times 10^{-4}$), which I discuss below.

The interaction model tested this formally (Table 7; Fig. 6). With hypermutated tumors removed, breaking an HR gene added about 18 points of SBS3 in carrier tissues but only 0.9 points in other tissues ($p = 0.66$ for the other tissues on their own). The difference between those two effects was very significant (interaction $p = 2.8 \times 10^{-9}$). *BRCA2* alone showed the cleanest pattern: +26 points in carrier tissues and

basically exactly zero elsewhere (-0.04 points; interaction $p = 8.6 \times 10^{-10}$). With all tumors included, the interaction was even stronger (HR group $p = 5.4 \times 10^{-21}$; *BRCA2* $p = 1.5 \times 10^{-17}$).

BRCA1 was different. With all tumors included, it also looked tissue-specific ($+15$ points in carrier tissues vs. $+0.2$ elsewhere; interaction $p = 1.4 \times 10^{-4}$). But after removing hypermutated tumors, *BRCA1*-mutated tumors in other tissues also had more SBS3 ($+13$ points, $p = 0.004$), the same as in carrier tissues, so the interaction went away ($p = 0.99$). Only 12 non-hypermutated *BRCA1*-mutated tumors were in other tissues, mostly lung squamous and bladder, so this estimate is shaky. But I'm reporting it as is: the tissue-specific pattern is strong for *BRCA2* and for the HR group as a whole, but I can't say the same for *BRCA1* on its own.

Table 7

Tissue-specificity and pooled breast + ovarian results (adjusted for mutation burden and age; the interaction model also adjusts for cancer type).

Group	Tumors	Mutated in carrier / other tissues	Effect in carrier tissues	Effect in other tissues	Interaction p	Pooled breast + ovarian effect (95% CI)	Pooled p
HR genes	All	56 / 182	+ 0.19	-0.03	5.4×10^{-21}	+ 0.17 (0.11 to 0.22)	2.8×10^{-9}
HR genes	Non-hyper.	54 / 62	+ 0.18	+ 0.01	2.8×10^{-9}	+ 0.15 (0.10 to 0.20)	3.6×10^{-8}
<i>BRCA2</i>	All	24 / 105	+ 0.26	-0.04	1.5×10^{-17}	+ 0.24 (0.14 to 0.33)	4.4×10^{-7}
<i>BRCA2</i>	Non-hyper.	22 / 31	+ 0.26	0.00	8.6×10^{-10}	+ 0.24 (0.15 to 0.33)	1.8×10^{-7}
<i>BRCA1</i>	All	28 / 38	+ 0.15	+ 0.00	1.4×10^{-4}	+ 0.13 (0.06 to 0.20)	3.6×10^{-4}
<i>BRCA1</i>	Non-hyper.	28 / 12	+ 0.13	+ 0.13	0.99	+ 0.11 (0.04 to 0.18)	0.0026

Pooling breast and ovarian tumors together gave the HR group 49 mutated tumors, about twice as many as either gene alone, and a clear overall effect of $+17$ points of SBS3 (95% CI 11 to 22; $p = 2.8 \times 10^{-9}$). All three groups stayed significant after removing hypermutated tumors (Table 7).

Two tissues are worth mentioning on their own. Cervical cancer was the only non-carrier tissue with a clear SBS3 effect. Interestingly, cervical cancer is one of *BRCA1*'s carrier cancer types in my Aim 1 catalog (Table 2), even though it's not one of the four classic ones I picked. Cervical tumors also have a lot of APOBEC damage, which can sometimes blur with SBS3 when fitting signatures, so I think this needs a closer look before reading much into it. Second, in the all-tumor analysis, three glioblastomas with HR mutations had much lower SBS3 than other glioblastomas (-49 points). All three were hypermutated and disappeared when I removed hypermutated tumors. Hypermutated glioblastomas are

usually tumors that came back after treatment with the chemotherapy drug temozolomide, which leaves its own signature (SBS11) that I didn't include in my fit. So this is most likely a fitting artifact and not biology.

3.8 Sensitivity check: counting missense mutations

When I also counted missense mutations as “broken,” the groups got bigger, but the breast cancer signals got weaker (*BRCA2*: 27 mutated tumors, + 6 points, $p = 0.13$; *BRCA1*: 24 tumors, + 6 points, $p = 0.16$). The ovarian results got slightly stronger (*BRCA2* $p = 0.023$; *BRCA1* $p = 0.046$). The missense version also produced a couple of strange results in the opposite direction, like less MMR damage in *BRCA1*-missense endometrial tumors, which were almost all hypermutated. Most missense mutations in big genes like *BRCA1* and *BRCA2* don't actually break the protein [14], so mixing them in adds tumors whose HR is still working. I think this weakening supports using truncating mutations as the main definition, since it's what you'd expect if the effect is really coming from losing the gene.

3.9 Putting the two phases together

Phase 1 and Phase 2 point at the same thing for *BRCA2*, using completely different data. In Phase 1, the HR pathway around *BRCA2* was more open in the tissues where *BRCA2* carriers get cancer, in both ENCODE normal tissues (+ 1.67) and TCGA tumors (+ 3.26). In Phase 2, losing *BRCA2* left the HR-failure scar mainly in carrier tissues and almost nowhere else. Both results fit the same idea: some tissues lean on HR more heavily, so losing *BRCA2* matters much more there. I want to be clear that I didn't directly test whether the amount of openness in each tissue predicts the size of the SBS3 effect in that tissue. The two phases also defined carrier tissues a little differently (Phase 1 used the full Aim 1 catalog, and Phase 2 used the four classic *BRCA* cancers). Connecting them tissue by tissue is the most important next step.

4. Discussion

I started this project with a simple question: why do people who inherit the same broken cancer gene end up with different cancers? Across two phases, I found four main things. First, the cancer types that carriers of the same gene develop come from tissues with very different exposure histories, and this was true for all 11 genes. Second, the *BRCA2* HR pathway is more open in *BRCA2* carrier tissues, in both normal tissue and tumors. Third, when *BRCA1* or *BRCA2* is broken in a breast tumor, the tumor carries a strong HR-failure scar (SBS3). Fourth, and most important, that scar depends on the tissue. For *BRCA2* and for the HR genes as a group, it shows up in breast, ovarian, and prostate tumors and is basically absent in other tissues, even when the gene is just as broken.

The same broken gene doesn't do the same thing everywhere. The tissue-specificity result is the part I think matters most. If losing *BRCA2* always caused HR failure, you would expect to see the SBS3 scar in every tumor with a broken *BRCA2*, no matter where it grew. That isn't what happened. In lung, bladder, colon, and head and neck tumors, a truncating *BRCA2* mutation came with no extra SBS3 at all. This matches a clinical study that found *BRCA1/2* changes in non-classic tumor types often don't show the

usual signs of HR loss [25]. There are a few possible reasons. In other tissues, the *BRCA2* mutation might be a passenger that happened after the tumor already formed, the second copy of the gene might still be working [26], or those tissues might just not depend on HR as much. My Phase 1 results support the last idea for *BRCA2*, because its HR pathway was more open in the tissues where the scar shows up. What I think my project adds is the chromatin side: connecting where the pathway is open to where losing the gene actually leaves a mark.

Why this could matter for patients. Tumors with HR failure respond to PARP inhibitors, a type of drug that takes advantage of the broken repair [27]. My results suggest that finding a *BRCA2* mutation in a tumor from a non-classic tissue may not mean the tumor actually has HR failure. Checking the SBS3 scar directly, and not just whether the gene is mutated, might be a better way to decide who would benefit, which is also what larger clinical studies have suggested [25]. I'm a high school student and not a doctor, so I don't want to overstate this, but it's the direction the data points.

Exposure is the big pattern, but it's harder to pin to one gene. Phase 1 showed that exposure history separates carriers' cancer types very well. But in Phase 2, the one exposure-specific prediction I could test, *ATM* and smoking in lung cancer, only pointed the right way without reaching significance once I adjusted for mutation burden. With 17 mutated tumors, I probably didn't have enough to detect a modest effect. I think the most honest summary is that exposure explains which tissues carriers' cancers come from in general, while the repair-gene effects I could measure are more tissue-specific and gene-specific.

Not every gene works the same way. In Phase 1, *PMS2*'s mismatch repair pathway was actually less open in its carrier tissues, which is the opposite of *BRCA2*. In Phase 2, *PMS2*-mutated endometrial tumors had much more mismatch-repair damage before adjustment, but it's hard to separate the gene effect from the fact that losing mismatch repair creates a huge number of mutations. *BRCA1* also behaved differently from *BRCA2* once hypermutated tumors were removed. So different predisposition genes probably shape where cancer shows up through different routes, which is also why pooling all 11 genes together in Phase 1 weakened the chromatin signal.

What I learned about doing this kind of analysis. Three choices made a big difference to what I would have concluded. If I hadn't adjusted for mutation burden, I would have reported a bunch of "significant" results in stomach and endometrial cancer that were really just about hypermutation. If I hadn't removed hypermutated tumors, I would have said *BRCA1* was clearly tissue-specific, which it may not be. And if I had only reported the best of my eight chromatin versions in Phase 1, the all-gene openness result would have looked real when it didn't pass correction. I tried to report every version I ran so readers can see how much each choice matters.

5. Limitations

1. **Tumor mutations are a stand-in for inherited ones.** Phase 2 used tumors where the gene is broken by a somatic mutation, not inherited carriers. A somatic mutation might hit only one copy of the gene or happen late in the tumor's life, while inherited carriers start with one broken copy in every

cell. Some “not mutated” tumors could also belong to inherited carriers. The results should be repeated in true carriers.

2. **Only one copy may be broken.** I didn't check whether the second copy of each gene was also lost (loss of heterozygosity). Tumors where both copies are gone would probably show a stronger SBS3 effect [26].
3. **Small groups.** Many groups had fewer than 10 mutated tumors, and several predictions (*BRIP1*, *CHEK2*, *PALB2* in breast, *PMS2* in prostate) couldn't be tested at all. Pancreatic cancer, one of the four carrier tissues, had too few HR-mutated tumors to test on its own.
4. **The tissue follow-up came after seeing Phase 2.** I picked the four carrier tissues before running the follow-up, based on clinical genetics, but after seeing the breast result. It's a follow-up test, not a fully pre-planned one.
5. **Signature fitting on exomes.** SBS3 is a flat signature that can be hard to separate from other flat signatures, especially in tumors with few mutations, and I used genome-based reference signatures on exome data. The glioblastoma result shows that missing signatures (like SBS11 from temozolomide) can create artifacts.
6. **The two phases weren't connected tissue by tissue.** I didn't directly test whether tissues with more HR-pathway openness have bigger SBS3 effects, and the two phases defined carrier tissues slightly differently.
7. **Phase 1 limits still apply.** The Phase 1 signature comparisons used all TCGA tumors of each type, eight versions of the openness analysis were tested and the all-gene result didn't pass correction, some tissues were stand-ins, and TCGA is mostly adults, which is why *TP53* dropped out.

6. Future Directions

1. Repeat Phase 2 in true inherited carriers by linking each carrier to their own tumor through TCGA controlled access (with my mentor), and test carriers against non-carriers with the same cancer type.
2. Connect the two phases tissue by tissue: test whether each tissue's HR-pathway openness predicts how big the SBS3 effect of a broken *BRCA2* is in that tissue.
3. Add loss-of-heterozygosity data to separate tumors that lost both copies of an HR gene from tumors with only one hit, and check the SBS3 results against published HRD scores for TCGA [28].
4. Retest *PMS2* without adjusting for mutation burden in a way that fits how mismatch repair works, and test *ATM* and smoking in a larger lung cancer cohort.
5. Finish Aim 4 with cross-validated models where each tumor gets openness data from its own tumor (Corces et al. [10]), so exposure and chromatin can be compared head to head.

7. Conclusion

People who inherit a mutation in the same cancer gene develop cancers in tissues with very different exposure histories, and this held for all 11 genes I studied. For *BRCA2*, the tissue matters in two connected ways: its HR pathway is more open in the tissues where carriers get cancer, and when *BRCA2*

is broken, the HR-failure scar shows up in those same tissues (+ 26 points of SBS3) and basically nowhere else. Together, these results suggest that which cancer a carrier develops depends not just on the gene they inherit, but on how much each tissue relies on the pathway that gene belongs to. All of this came from public data, and the pipeline can be reused as soon as the inherited-carrier data becomes available.

Declarations

Data and Code Availability

All data sources are public and are listed in Table 1. The Phase 2 analysis scripts (Phase2_Somatic_Proxy_Analysis.R and Phase2b_Tissue_Specificity.R), which download the public data and reproduce every Phase 2 table and figure, are available on request, along with the Phase 1 scripts and output files.

References

1. Huang KL, Mashl RJ, Wu Y et al (2018) Pathogenic germline variants in 10,389 adult cancers. *Cell* 173(2):355–370e14
2. Knudson AG (1971) Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci USA* 68(4):820–823
3. Malkin D, Li FP, Strong LC et al (1990) Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* 250(4985):1233–1238
4. Lynch HT, de la Chapelle A (2003) Hereditary colorectal cancer. *N Engl J Med* 348(10):919–932
5. Kuchenbaecker KB, Hopper JL, Barnes DR et al (2017) Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA* 317(23):2402–2416
6. Alexandrov LB, Kim J, Haradhvala NJ et al (2020) The repertoire of mutational signatures in human cancer. *Nature* 578(7793):94–101
7. Polak P, Karlić R, Koren A et al (2015) Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* 518(7539):360–364
8. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10(12):1213–1218
9. ENCODE Project Consortium, Moore JE, Purcaro MJ et al (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 583(7818):699–710
10. Corces MR, Granja JM, Shams S et al (2018) The chromatin accessibility landscape of primary human cancers. *Science* 362(6413):eaav1898
11. Landrum MJ, Lee JM, Benson M et al (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 46(D1):D1062–D1067

12. Liberzon A, Subramanian A, Pinchback R et al (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27(12):1739–1740
13. Gillespie M, Jassal B, Stephan R et al (2022) The reactome pathway knowledgebase 2022. *Nucleic Acids Res* 50(D1):D687–D692
14. Richards S, Aziz N, Bale S et al (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 17(5):405–424
15. Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. *J Am Stat Assoc* 47(260):583–621
16. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 57(1):289–300
17. Robin X, Turck N, Hainard A et al (2011) pROC: an open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinformatics* 12:77
18. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22(11):1359–1366
19. Ellrott K, Bailey MH, Saksena G et al (2018) Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst* 6(3):271–281e7
20. Liu J, Lichtenberg T, Hoadley KA et al (2018) An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173(2):400–416e11
21. Tate JG, Bamford S, Jubb HC et al (2019) COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* 47(D1):D941–D947
22. Manders F, Brandsma AM, de Kanter J et al (2022) MutationalPatterns: the one stop shop for the analysis of mutational processes. *BMC Genomics* 23:134
23. Nik-Zainal S, Davies H, Staaf J et al (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 534(7605):47–54
24. Polak P, Kim J, Braunstein LZ et al (2017) A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat Genet* 49(10):1476–1486
25. Jonsson P, Bandlamudi C, Cheng ML et al (2019) Tumour lineage shapes BRCA-mediated phenotypes. *Nature* 571(7766):576–579
26. Maxwell KN, Wubbenhorst B, Wenz BM et al (2017) BRCA locus-specific loss of heterozygosity in germline BRCA1 and BRCA2 carriers. *Nat Commun* 8:319
27. Lord CJ, Ashworth A (2017) PARP inhibitors: synthetic lethality in the clinic. *Science* 355(6330):1152–1158
28. Knijnenburg TA, Wang L, Zimmermann MT et al (2018) Genomic and molecular landscape of DNA damage repair deficiency across The Cancer Genome Atlas. *Cell Rep* 23(1):239–254e6

Figures

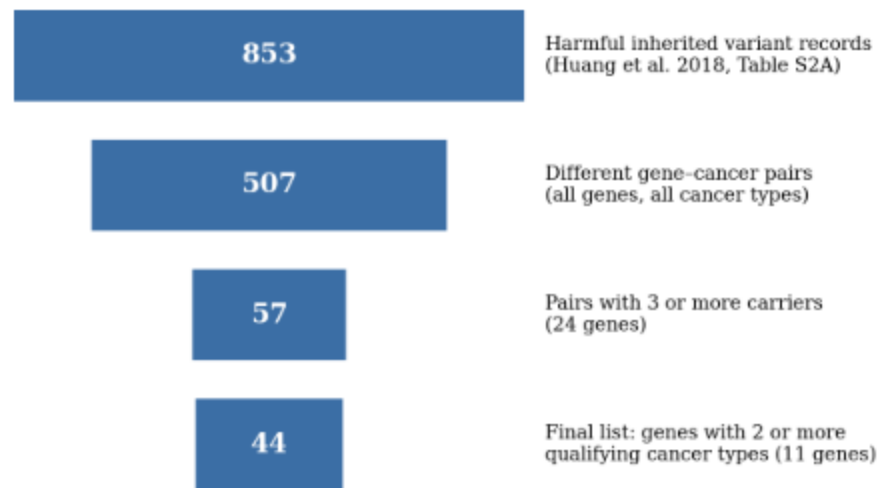


Figure 1

How the carrier list was built. The top number counts variant records, and the lower numbers count gene–cancer pairs.

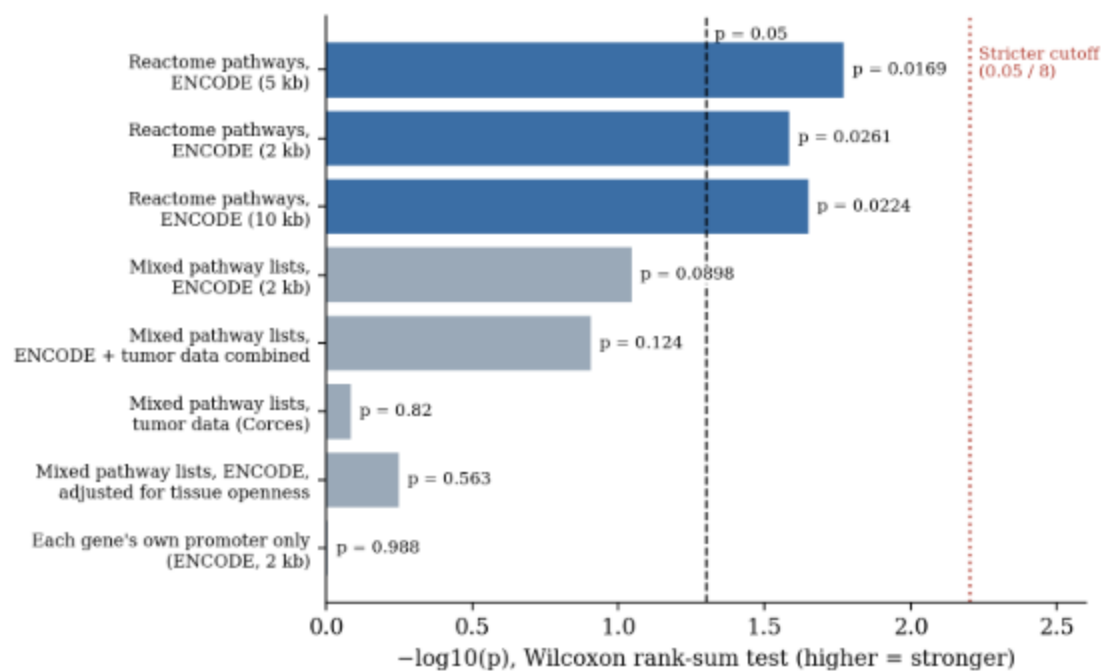


Figure 2

Results for all eight versions of the DNA openness analysis. Blue bars had $p < 0.05$, but none passed the stricter cutoff for testing eight versions (red dotted line).

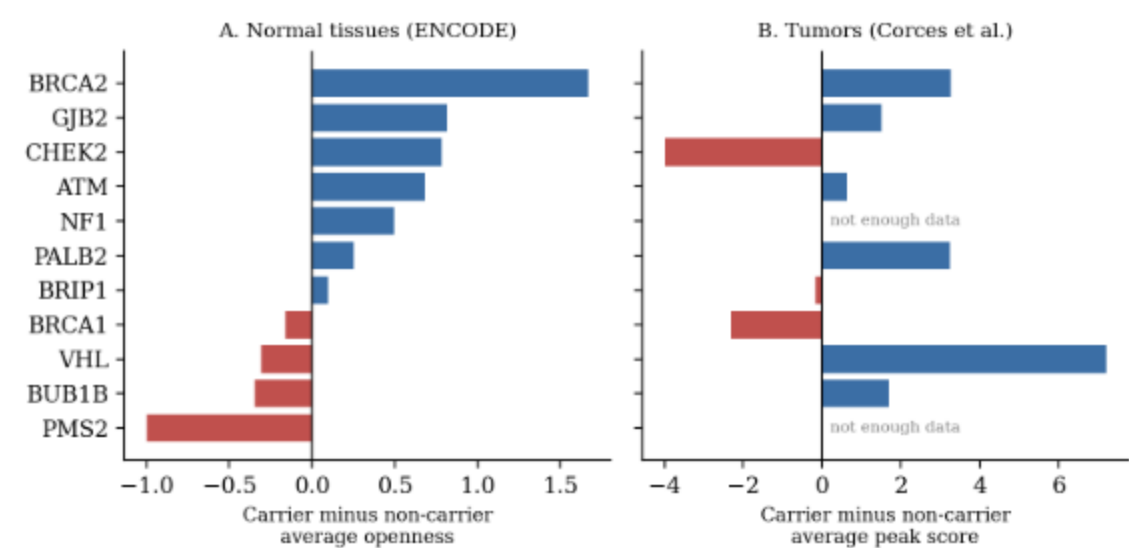


Figure 3

Difference in pathway openness between carrier and non-carrier tissues for each gene, in (A) normal tissues from ENCODE (Reactome, 2 kb) and (B) tumors from TCGA. Blue means more open in carrier tissues, and red means less open.

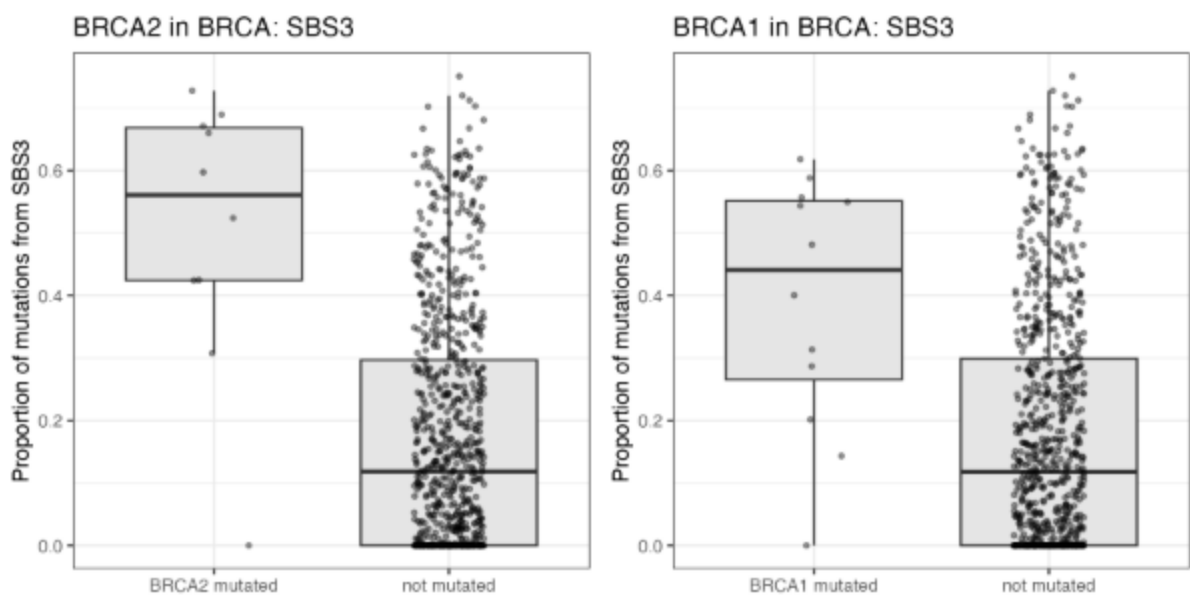


Figure 4

Share of each tumor's mutations that come from SBS3 (the HR-failure signature) in breast cancer, comparing tumors with a truncating mutation in *BRCA2* (left, $n = 10$) or *BRCA1* (right, $n = 12$) to all other breast tumors. Each dot is one tumor.

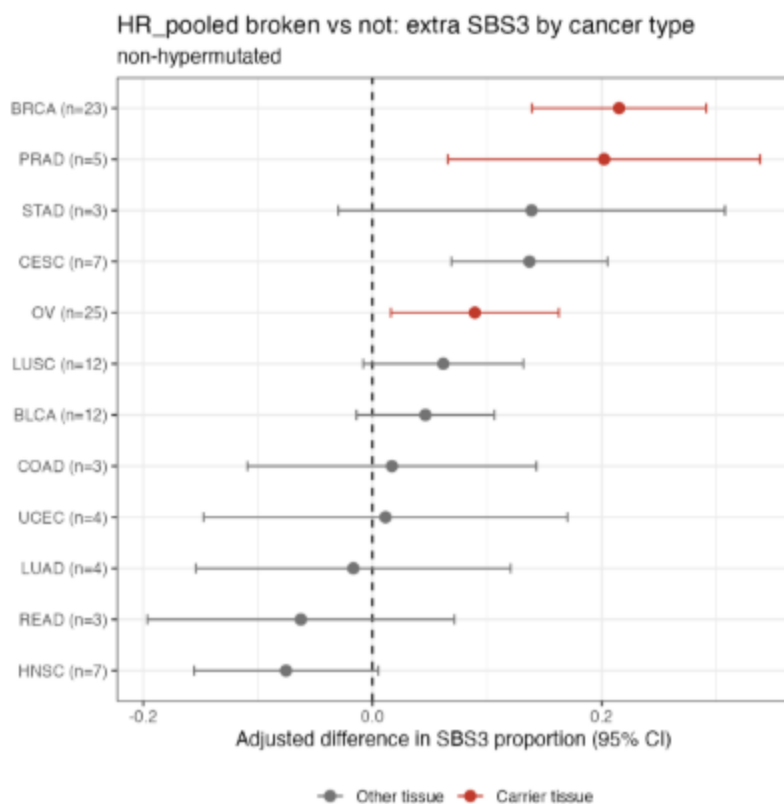


Figure 5

Extra SBS3 in tumors with a truncating mutation in any of *BRCA1*, *BRCA2*, *PALB2*, or *BRIP1*, by cancer type (hypermutated tumors removed). Points show the adjusted difference in SBS3 share, and lines show 95% confidence intervals. Red marks the carrier tissues, and n is the number of mutated tumors.

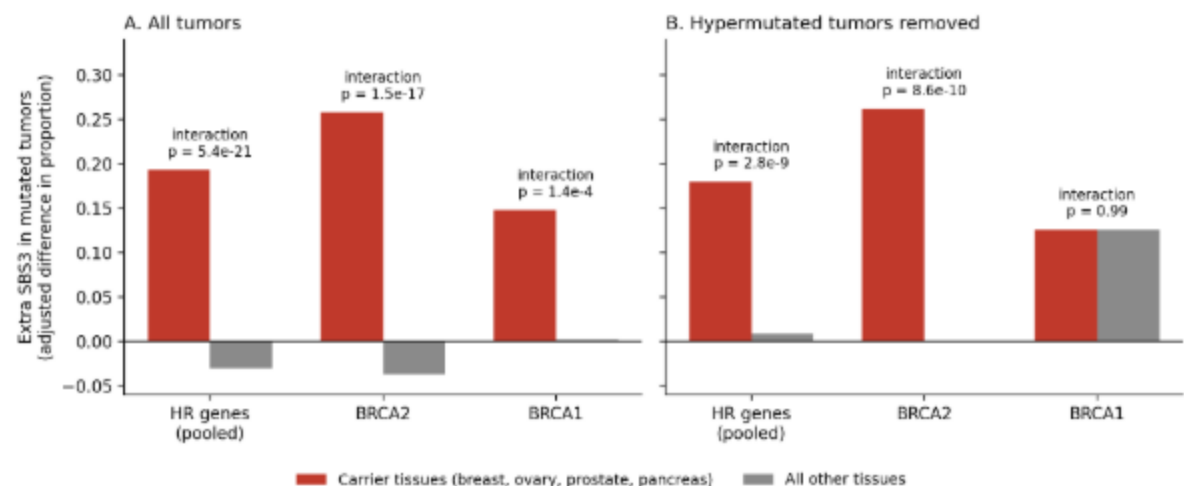


Figure 6

Extra SBS3 from a broken HR gene in carrier tissues (red) vs. all other tissues (grey), from the interaction model. (A) All tumors. (B) Tumors with 500 or more mutations removed. The interaction p-value tests whether the red and grey effects are different.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterial.docx](#)