Detection of Molecular Signatures of Homologous Recombination Deficiency in Bladder Cancer

Supplementary Methods

Judit Börcsök, Miklos Diossy, Zsofia Sztupinszki, Aurel Prosz, Viktoria Tisza, Sandor Spisak, Orsolya Rusz, Dag R. Stormoen, Helle Pappot, Istvan Csabai, Søren Brunak, Kent W. Mouw, and Zoltan Szallasi

Date of the latest update:
April 12, 2021

Contents

List of Figures

List of Tables

1 Analyzed Cohorts

In this study 533 whole genome (WGS) and whole exome sequenced (WES) **pretreatment** samples were analyzed from three urothelial bladder tumor cohorts (Supp. Table 1). Patients in the DFCI/MSKCC and Philadelphia cohorts received neoadjuvant cisplatin-based chemotherapy (NACC) and had available pre-chemotherapy tumor tissue.

Cohort	Sequencing type	Number of samples	Tissue source	Type	Therapy
TCGA	WGS	23	FF	MIBC	H
TCGA	WES	412	FF	MIBC	H
DFCI/MSKCC	WES	50	FF	MIBC	NACC
Philadelphia	WES	48	FFPE	MIBC	NACC

Supp. Table 1: Summary of the analyzed WGS and WES cohorts. Abbreviations: FF - fresh frozen; FFPE - formalin-fixed paraffin-embedded; MIBC - muscle-invasive bladder cancer; NACC - neoadjuvant cisplatin-based chemotherapy; H - heterogeneous therapy.

1.1 TCGA BLCA WGS

The WGS normal and tumor bam files were downloaded from the ICGC data portal (https://dcc.icgc.org/).

1.2 TCGA BLCA WES

The WES normal and tumor bam files, as well as the vcf files generated by MuTect2, were downloaded from the TCGA data portal (https://portal.gdc.cancer.gov/).

1.3 DFCI/MSKCC WES

The normal and tumor bam files were downloaded from The database of Genotypes and Phenotypes (dbGaP) upon request using the **phs000771.v2.p1** accession code (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000771.v2.p1).

1.4 Philadelphia WES

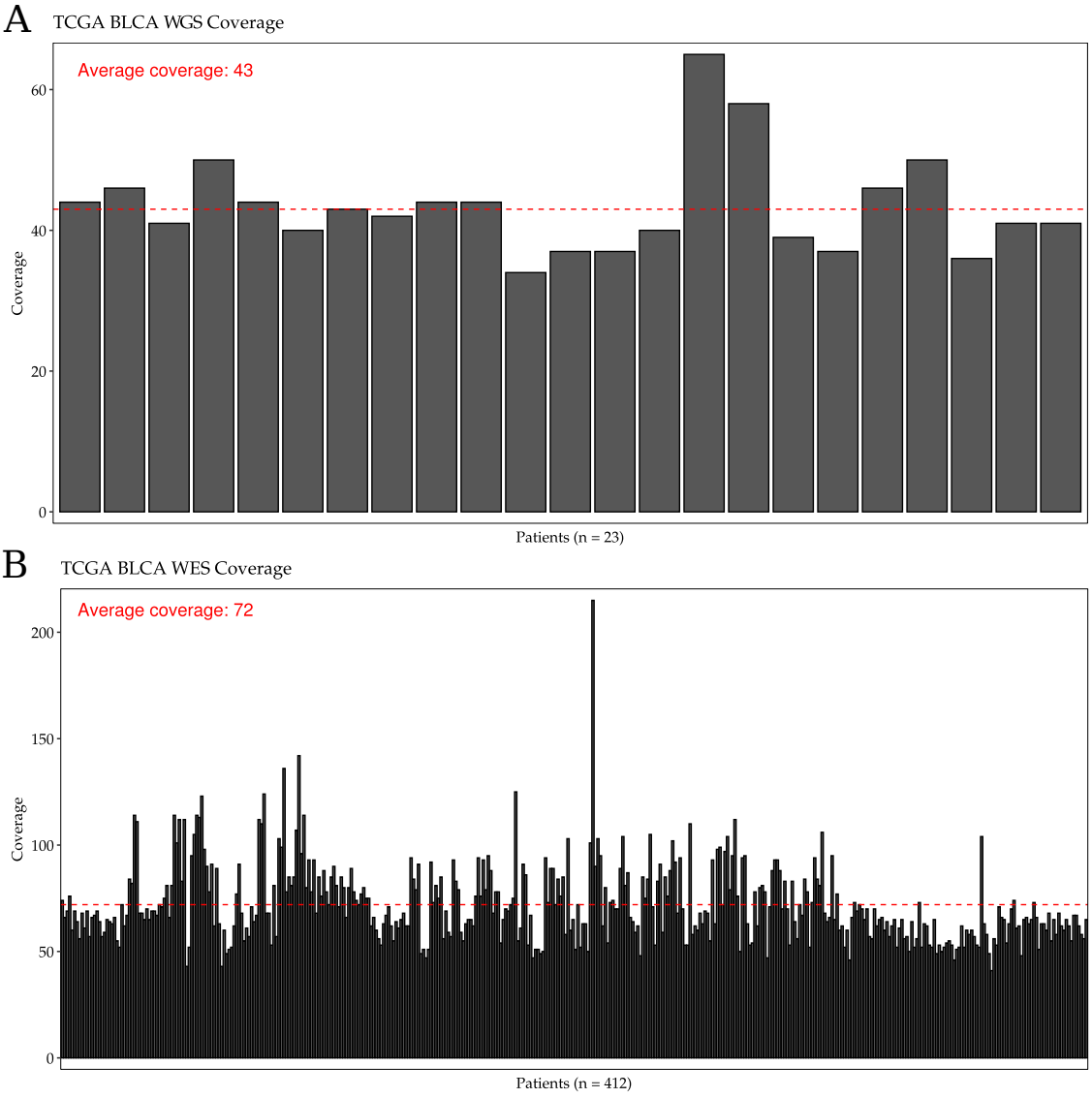
The normal and tumor bam files were downloaded from The database of Genotypes and Phenotypes (dbGaP) upon request using the **phs000771.v2.p1** accession code (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000771.v2.p1). Tumor samples in this cohort are derived from formalin-fixed paraffin-embedded (FFPE) tissues.

2 Coverage

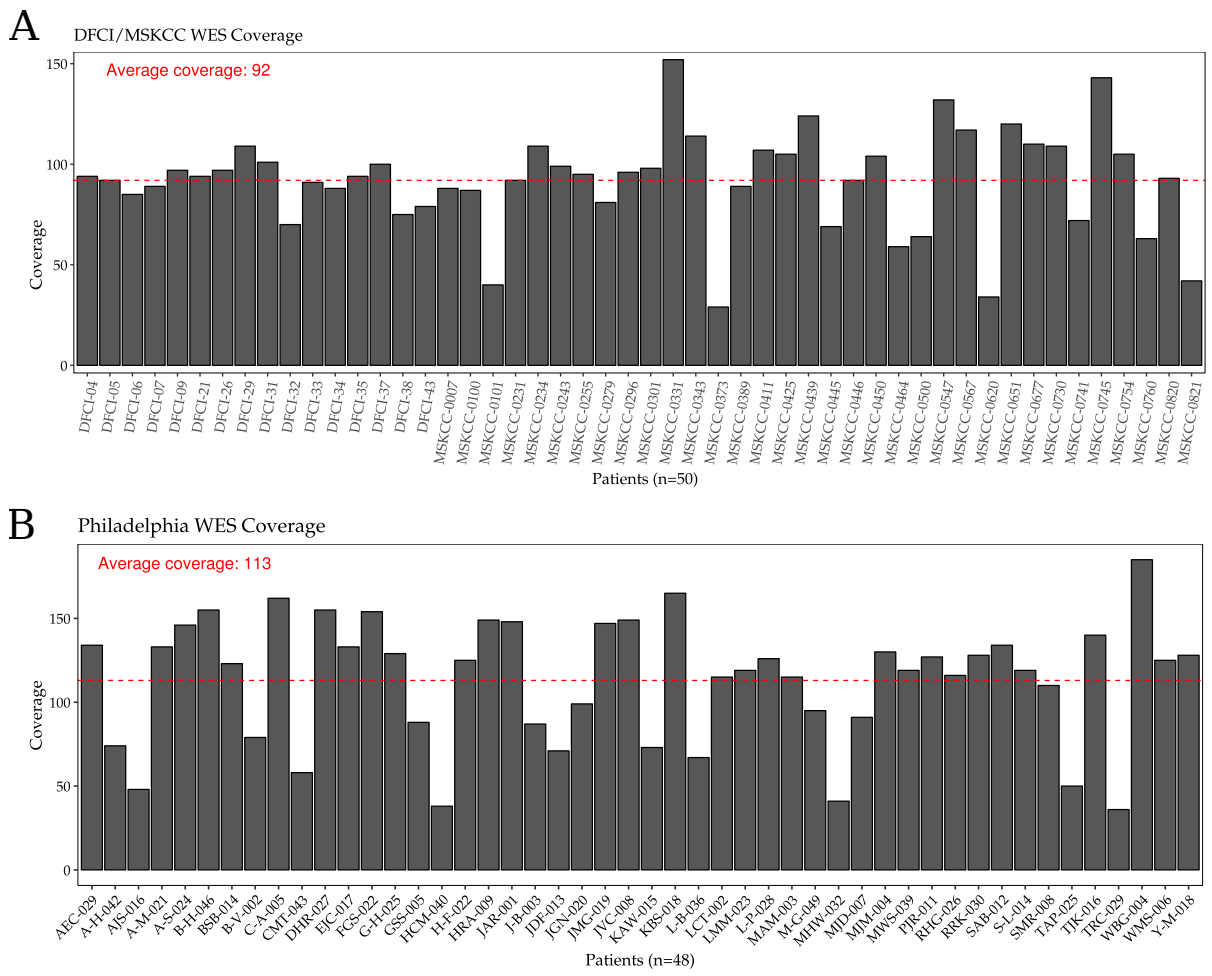
In order to determine the average coverage of the WGS and WES bam files in each cohort `samtools` [22] was used. The average coverage (D_{avg}) in a given cohort was calculated as follows

$$D_{avg} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M} d_{ij}, \tag{2.1}$$

where d_{ij} is the depth at the jth position within an exonic region of the genome of the ith sample, M is the number of exonic positions examined in the genome, and N is the number of samples in a given cohort. The average coverage of the analyzed WGS and WES samples from the TCGA BLCA cohort and the average coverage of the WES samples from the DFCI/MSKCC and Philadelphia cohorts were shown in Supp. Fig. 1 and Supp. Fig. 2, respectively.



Supp. Fig. 1: TCGA BLCA WGS and WES cohorts. The average coverage of the analyzed samples in the TCGA BLCA WGS (**A**) and TCGA BLCA WES (**B**) cohorts.



Supp. Fig. 2: DFCI/MSKCC and Philadelphia BLCA WES cohorts. The average coverage of samples in the DFCI/MSKCC (**A**) and Philadelphia (**B**) BLCA WES cohorts.

3 Mutation and copy number calling

3.1 Genotyping

Genotypes were determined according to the following scheme:

- germline variants were called via `GATK` [14] (version 3.8) `HaplotypeCaller` in key DNA damage response (DDR) genes (specifically focusing on HR-related genes) in WES and WGS samples,
- somatic point mutations and indels were called with `GATK` (version 3.8) `MuTect2` in WGS samples; whole exome vcf files generated by `MuTect2` were downloaded from the TCGA data portal.

The high fidelity of the reported germline and somatic variants was ensured by the application of additional hard filters (Supp. Table 2 and Supp. Table 3) in addition to the tools' default filters (FILTER == "PASS").

Cohort	TCGA WGS	TCGA WES	DFCI WES	MSKCC WES	Philadelphia WES
MQ	≥ 50	≥ 50	≥ 30	≥ 30	≥ 30
QUAL	≥ 20	≥ 20	≥ 10	≥ 10	≥ 10
DP	≥ 15	≥ 15	≥ 15	≥ 15	≥ 10

Supp. Table 2: Additional germline mutation filtering parameters applied in the cohorts.

Cohort	TCGA WGS	TCGA WES	DFCI WES	MSKCC WES
TLOD	≥ 6	≥ 6	≥ 6	≥ 6
NLOD	≥ 3	≥ 3	≥ 4	≥ 4
NORMAL.DEPTH	≥ 15	≥ 10	≥ 10	≥ 10
TUMOR.DEPTH	≥ 20	≥ 10	≥ 15	≥ 15
TUMOR.ALT	≥ 5	≥ 5	≥ 5	≥ 5
NORMAL.ALT	$= 0$	$= 0$	$= 0$	$= 0$
TUMOR.AF	≥ 0.05	≥ 0.05	≥ 0.03	≥ 0.03

Supp. Table 3: Additional somatic mutation filtering parameters applied in the cohorts.

The pathogenicity of the variants was assessed by `Intervar` [23] (version 2.0.1) which classifies variants into five categories: "Benign", "Likely Benign", "Uncertain significance", "Likely Pathogenic" and "Pathogenic". Mutations in exonic regions that were not synonymous SNVs and classified as "Pathogenic" or "Likely Pathogenic" were considered as deleterious. Variants with "Unknown Significance" were collected separately.

In the Philadelphia FFPE cohort somatic SNPs were identified by `MuTect` [10], with computational filtering of artifacts introduced by DNA oxidation during sequencing or FFPE-based DNA extraction using a filter-based method [11]. Damaging non-silent mutations in DDR genes were considered as deleterious.

3.2 Loss of heterozygosity

Sequenza

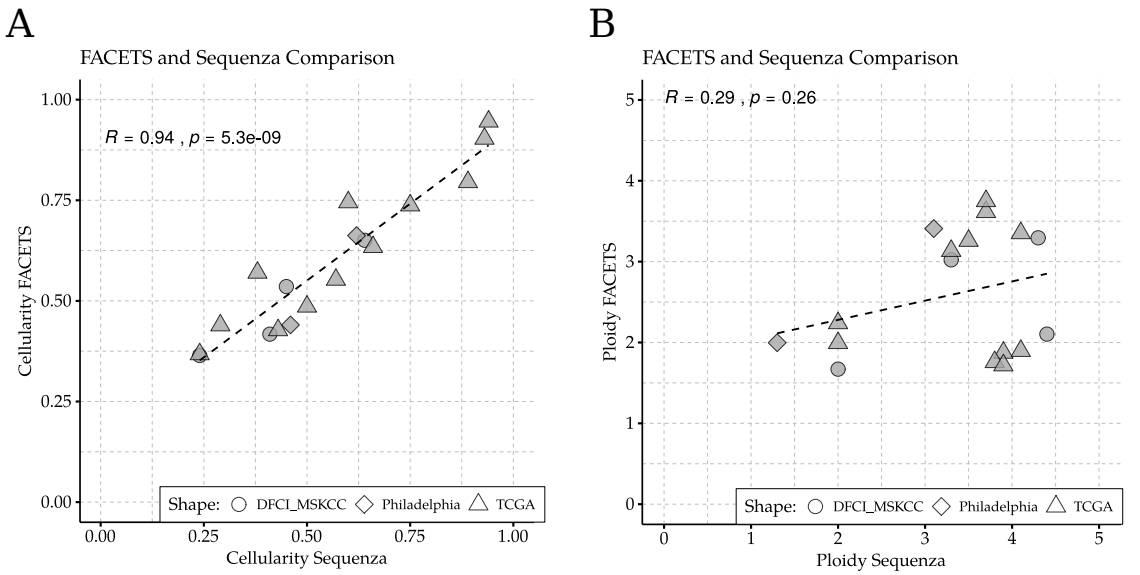
In order to estimate tumor cellularity and ploidy and to infer allele-specific copy number (ASCN) profiles `Sequenza` [17] was used. The fitted models were in the ploidy range of [1, 7] and cellularity range of [0, 1]. When the predictions of a fitted model were significantly different from the expected ploidy and cellularity values, an alternative solution was selected manually. If the copy numbers of either the A or B alleles dropped to zero within the coordinates of a gene, then an LOH event was registered. The summary of estimated LOH events in the

- DFCI/MSKCC WES cohort was shown in Supp. Fig. 5A;
- Philadelphia WES cohort was shown in Supp. Fig. 6A;
- TCGA WES cohort was presented in the top panel of Supp. Fig. 7.

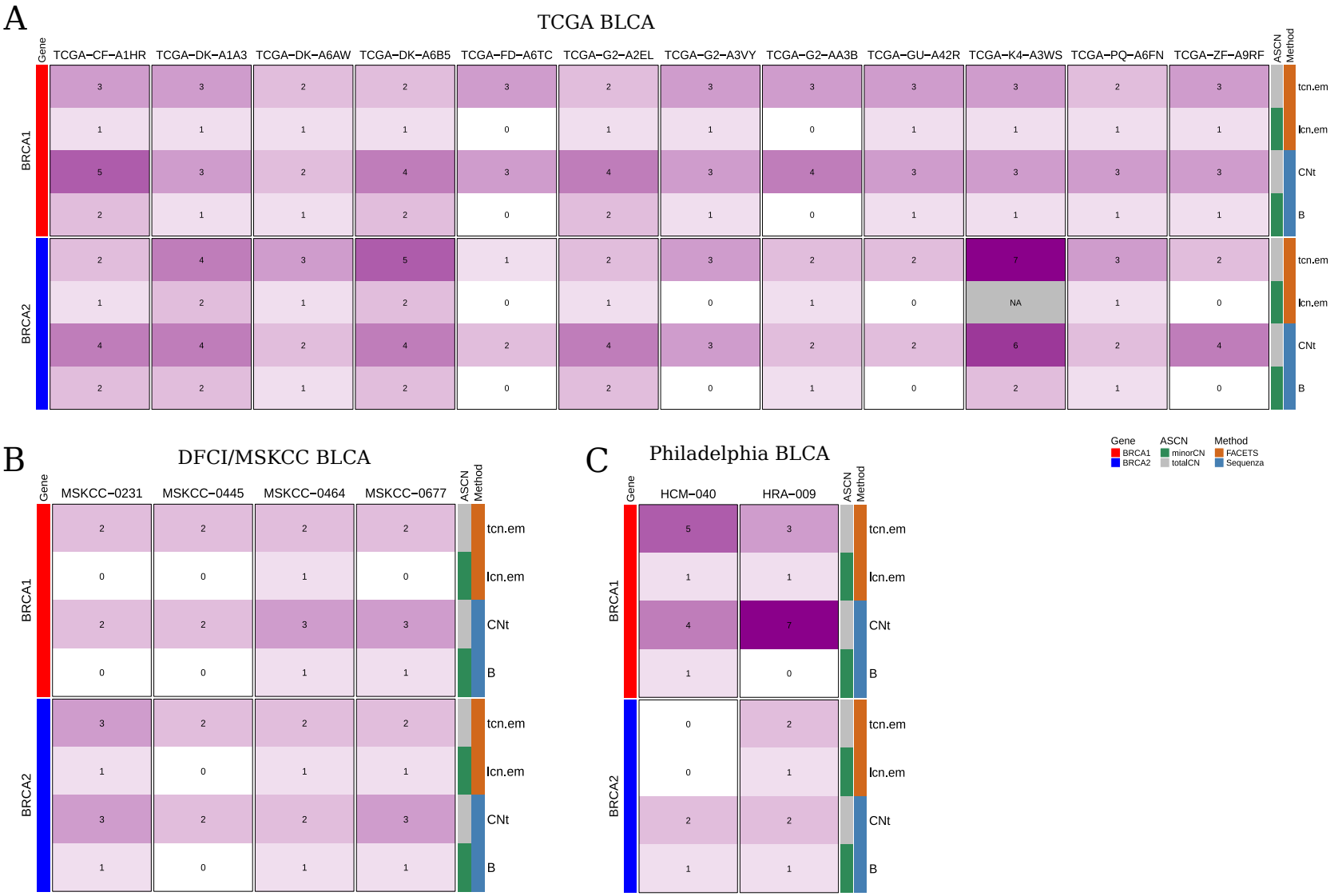
FACETS

The allele-specific copy number profiles of the samples with *BRCA1* or *BRCA2* pathogenic germline and/or somatic mutation(s) were estimated using a second tool called FACETS [31]. It is an open-source software applicable for next generation sequencing (NGS) data. The pipeline uses bam files as input files, and includes bam file post-processing, joint segmentation of total- and allele-specific read counts, and integer copy number calls corrected for tumor cellularity, ploidy and clonal heterogeneity.

The comparison of the cellularity values estimated by the two different methods showed a very strong ($R_{Pearson} = 0.94$) correlation (Supp. Fig. 3A). The comparison of Sequenza and FACETS ploidy estimates yielded a moderate ($R_{Pearson} = 0.29$) correlation coefficient (Supp. Fig. 3B). The CN profiles of *BRCA1* and *BRCA2* genes estimated by Sequenza and FACETS were reassuringly similar to each other (Supp.Fig. 4), and the application of the second method, FACETS, did not result in additional identification of *BRCA1/2*-deficient samples. Thus, the CN calls estimated by Sequenza were used further in the analysis.



Supp. Fig. 3: Correlation of cellularity (**A**) and ploidy (**B**) values estimated by FACETS (y axis) and Sequenza (x axis).



Supp. Fig. 4: Comparison of *BRCA1* and *BRCA2* CN profiles of samples with a *BRCA1* or a *BRCA2* pathogenic germline and/or somatic mutation(s) from the TCGA (**A**), DFCI/MSKCC (**B**) and Philadelphia (**C**) BLCA WES cohorts estimated by two different methods, `FACETS` and `Sequenza`. Abbreviations: *tcn.em* and *lcn.em*: total copy number and minor copy number estimated by `FACETS`; *CNt* and *B*: total copy number and minor copy number estimated by `Sequenza`. Note: due to the low coverage of the HCM-040 sample the ASCN calling was unreliable by `FACETS`.

3.3 Final genotypes

Genotyping was based on the presence of a pathogenic or likely pathogenic germline/somatic mutation and whether a loss of heterozygosity event occurred in a given HR-related gene.

- Wild type: no pathogenic or likely pathogenic germline or somatic mutation(s);
- Wild type with LOH: no pathogenic or likely pathogenic germline or somatic mutation(s), but an LOH event occurred;
- Heterozygote mutant: a pathogenic or likely pathogenic germline or somatic mutation is present, but no LOH;
- Heterozygote mutant with LOH: a pathogenic or likely pathogenic germline or somatic mutation is present and an LOH event occurred;
- Homozygote mutant: an identical germline or somatic mutation is present in both alleles;
- Compound heterozygote mutant: two different germline and/or somatic mutations are present in both alleles.

Summary of the final genotype of the samples in the

- DFCI/MSKCC WES cohort was shown in Supp. Fig. 5B;
- Philadelphia WES cohort was shown in Supp. Fig. 6B;
- TCGA WES cohort was presented in the bottom panel of Supp. Fig. 7.

TCGA BLCA									
Patient ID	Gene	Chr	POS	REF	ALT	Origin	ExonicFunc.refGene	Impact	Genotype
TCGA-G2-AA3B	*BRCA1*	17	43090981	G	C	somatic	stopgain	P	Heterozygote mutant with LOH
TCGA-G2-A3VY	*BRCA2*	13	32356472	C	T	germline	stopgain	P	Heterozygote mutant with LOH
TCGA-ZF-A9RF	*BRCA2*	13	32339511	A	AT	somatic	frameshift insertion	P	Heterozygote mutant with LOH
TCGA-K4-A3WS	*BRCA2*	13	32371076	C	T	somatic	stopgain	P	Compound heterozygote mutant
	BRCA2	13	32398437	C	G	germline	stopgain	VUS	
TCGA-XF-AAMX	*BRCA1*	17	43067615	C	T	somatic	nonsynonymous SNV	VUS	Heterozygote mutant with LOH
TCGA-BT-A3PH	*BRCA1*	17	43093817	C	T	somatic	nonsynonymous SNV	VUS	Heterozygote mutant with LOH
TCGA-BT-A20N	*BRCA2*	13	32398437	C	G	germline	stopgain	VUS	Heterozygote mutant with LOH
TCGA-4Z-AA7W	*BRCA2*	13	32337731	G	A	somatic	nonsynonymous SNV	VUS	Heterozygote mutant with LOH
TCGA-DK-AA6L	*BRCA2*	13	32363466	G	C	somatic	nonsynonymous SNV	VUS	Heterozygote mutant with LOH
DFCI/MSKCC BLCA									
MSKCC-0231	*BRCA1*	17	41209143	C	A	somatic	stopgain	P	Heterozygote mutant with LOH
MSKCC-0445	*BRCA2*	13	32915294	AG	A	germline	frameshift deletion	P	Compound heterozygote mutant
			32915299	TGG	T				
DFCI-32	*BRCA2*	13	32972413	G	A	somatic	nonsynonymous SNV	VUS	Heterozygote mutant with LOH
Philadelphia BLCA									
HRA-009	*BRCA1*	17	41244946	C	G	germline	missense mutation	D	Heterozygote mutant with LOH

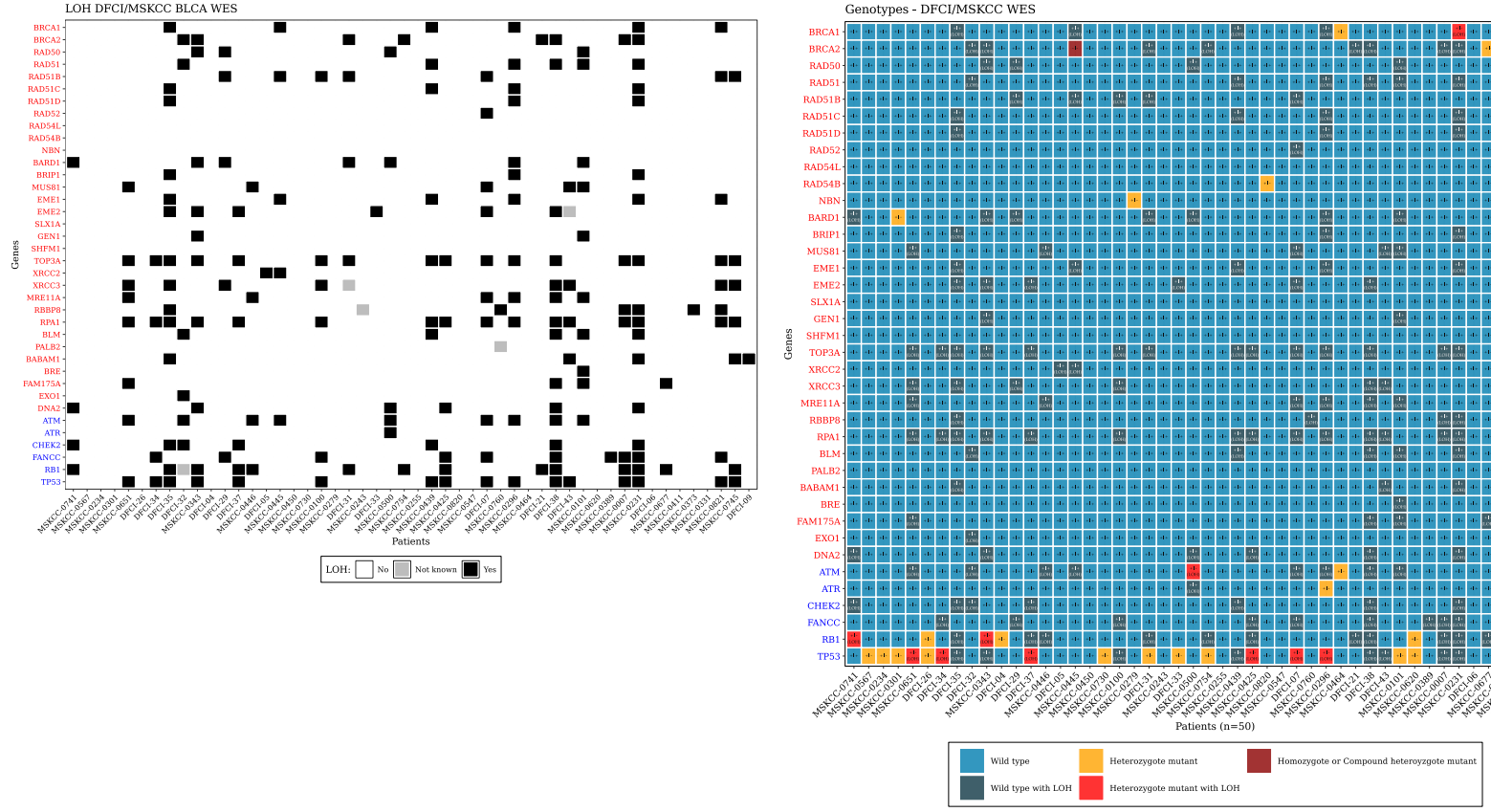
Supp. Table 4: The identified *BRCA1/2*-deficient samples in the three cohorts. Abbreviations: P - pathogenic variant, VUS - variant with uncertain significance, D - damaging variant.

TCGA BLCA									
Patient ID	Gene	Chr	POS	REF	ALT	Origin	ExonicFunc.refGene	Impact	Genotype
TCGA-ZF-AA58	*BARD1*	2	214781245	GTT	G	germline	frameshift deletion	P	Heterozygote mutant with LOH
TCGA-4Z-AA7W	*RBBP8*	18	22993313	C	T	germline	stopgain	P	Heterozygote mutant with LOH
TCGA-CF-A1HS	*RBBP8*	18	22993428	C	A	germline	stopgain	P	Heterozygote mutant with LOH

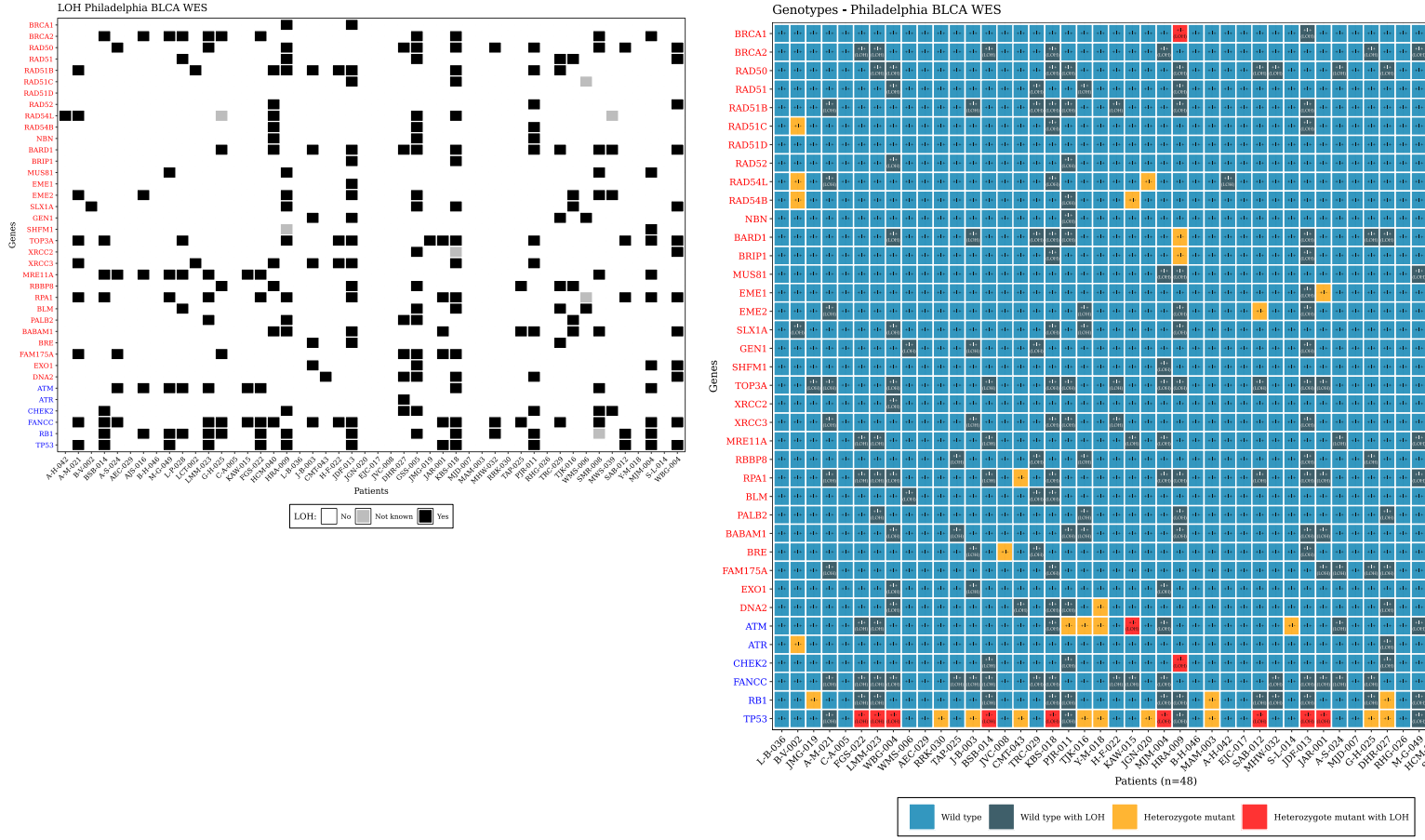
Supp. Table 5: Samples with a pathogenic or likely pathogenic mutation and an LOH event in other HR genes across the three cohorts. Abbreviations: P - pathogenic variant, VUS - variant with uncertain significance, D - damaging variant.

TCGA BLCA					
Gene	Wild-type	Wild-type with LOH	Heterozygote mutant	Heterozygote mutant with LOH	Homozygote or Compound heterozygote mutant
ATM	295	103	7	7	0
ATR	389	20	3	0	0
CHK2	311	98	2	0	1
RB1	258	107	14	33	0
TP53	122	161	21	107	1
DFCI/MSKCC BLCA					
ATM	40	8	1	1	0
ATR	48	1	1	0	0
CHK2	43	7	0	0	0
RB1	33	12	3	2	0
TP53	22	10	12	6	0
Philadelphia BLCA					
ATM	34	8	4	2	0
ATR	45	1	2	0	0
CHK2	40	6	1	1	0
RB1	29	15	4	0	0
TP53	22	4	13	9	0

Supp. Table 6: Frequency of pathogenic or likely pathogenic mutations in DNA damage checkpoint genes in the three cohorts.

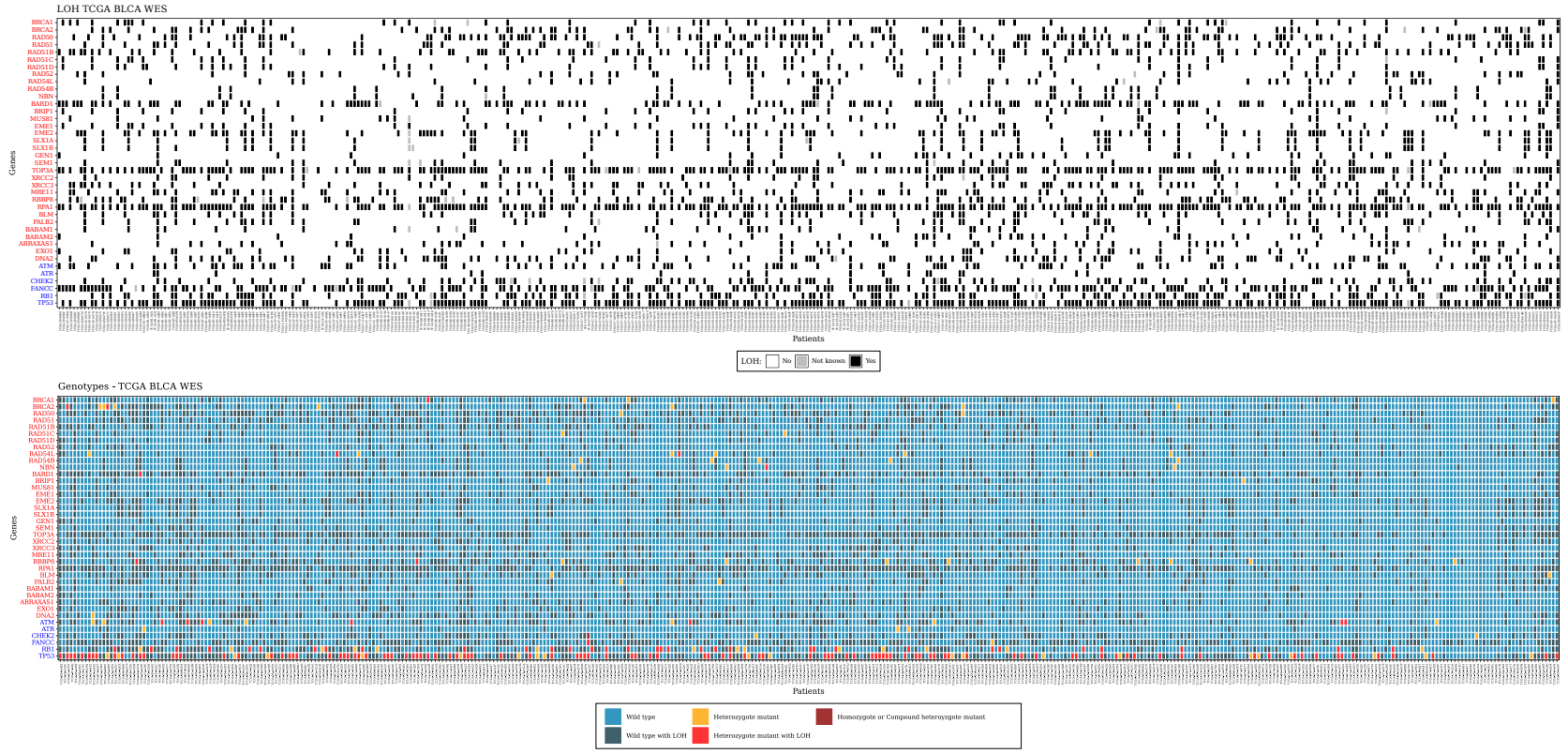


Supp. Fig. 5: DFCI/MSKCC BLCA WES. **A:** The estimated occurrence of LOH events in the examined DDR genes. **B:** Final genotype of the samples.



Supp. Fig. 6: Philadelphia BLCA WES. **A:** The estimated occurrence of LOH events in the tested DDR-related genes. **B:** Final genotype of the samples.

Online figure. Zoom in for details.



Supp. Fig. 7: TCGA BLCA WES. **Top panel**: The estimated occurrence of LOH events in the tested DDR-related genes. **Bottom panel**: Final genotype of the samples.

4 Mutational signature extraction

4.1 Single base substitution signatures

Single base substitution (SBS) signatures were extracted with the help of the `deconstructSigs` R package [29] which determines the linear combination of pre-defined signatures [4] that most accurately reconstructs the mutational profile of a single tumor sample.

The selected signatures, the linear combination of which could lead to the final mutational catalog, were confined to those, that were reported to be present in bladder and breast carcinoma according to the Catalogue of Somatic Mutations in Cancer (COSMIC) (https://cancer.sanger.ac.uk/signatures/signatures_v2/, BLCA: Signature 1, 2, 5, 10, 13, BRCA: Signature 1, 2, 3, 5, 6, 8, 10, 13, 17, 18, 20, 26, 30). In addition, Signature 4 was also extracted, because Signature 4 is associated with exposure to tobacco carcinogens [2] and it is well known that smoking is a strong risk factor for bladder cancer [18].

After evaluation of a sample's signature composition, its mutational catalog was reconstructed, and the cosine of the angles between the 96-dimensional original and reconstructed vectors was calculated (cosine similarity). In the cohorts, cosine similarities were high, mean cosine similarity ≥ 0.92 (Supp. Table 7), between the original and the reconstructed mutational profiles.

Cohort	Mean cosine similarity
TCGA WGS	0.99
TCGA WES	0.93
DFCI/MSKCC	0.92
Philadelphia WES	0.94

Supp. Table 7: Mean cosine similarity between original and reconstructed mutational catalogs in the analyzed cohorts.

One sample from the DFCI/MSKCC cohort (DFCI-43) had an extremely low cosine similaritiy, probably due to the low number of mutations after filtering, thus the results regarding this sample should not be considered reliable. Furthermore, from the 412 TCGA BLCA WES samples 17 were excluded from the final set of samples according to the following criteria:

- containing fewer than 50 somatic mutations after filtering;
- MSI samples identified by Bonneville *et al.* [7].

The extracted single base substitution signatures (COSMIC v2) from the

- TCGA WGS cohort were shown in Supp. Fig. 8A;
- DFCI/MSKCC WES cohort were plotted in Supp. Fig. 10A;
- Philadelphia WES cohort were shown in Supp. Fig. 11A;
- TCGA WES cohort were presented in Supp. Fig. 12.

The new version of single base substitution signatures (COSMIC v3; https://cancer.sanger.ac.uk/signatures/sbs/) was also extracted and the distributions of the number of mutations attributed to a certain signature were compared in regard of HR status. However, in order to calculate the HRDetect scores, the COSMIC v3 signatures were not used, since the HRDetect model was trained using the COSMIC v2 signatures [12].

4.2 Doublet base substitution signatures

Doublet base substitution (DBS) signatures (11) were characterized by Alexandrov *et al.* [3] using methods based on non-negative matrix factorization (NMF). The identified matrix of DBS signatures (**P**) was downloaded using the link below.

- Separate extraction of DBS signatures from all PCAWG whole genome samples together: https://www.synapse.org/#!Synapse:syn12025148;

Doublet base substitutions in each sample were classified into a 78-dimesional DBS catalog (**M**) with the help of the ICAMS R package [8]. The **M** and the **P** matrices were used in a non-negative least-squares problem to estimate the matrix of exposures to mutational processes (**E**).

$$\min_{\mathbf{E_i}} \|\mathbf{PE_i} - \mathbf{M_i}\|^2, \quad \text{subject to} \quad \mathbf{E_i} \geq 0, \text{ for all } i = 1, \ldots, N \tag{4.1}$$

where i is a given sample.

The extracted DBS signatures from the

- TCGA WGS cohort were shown in Supp. Fig. 8B;
- DFCI/MSKCC WES cohort were plotted in Supp. Fig. 10B;
- Philadelphia WES cohort were shown in Supp. Fig. 11B;
- TCGA WES cohort were presented in Supp. Fig. 13.

4.3 Indel signatures

Similarly to doublet base substitution signatures, small insertion and deletion (ID) signatures (17) were also characterized [3] and the identified matrix of ID signatures (**P**) was published.

- Separate extraction of ID signatures from all PCAWG whole genome samples together: https://www.synapse.org/#!Synapse:syn12025148;

Insertions and deletions in each sample were classified into an 83-dimesional indel catalog (**M**) with the help of the ICAMS R package [8]. The **M** and the **P** matrices were used in a non-negative least-squares problem to estimate the matrix of exposures to mutational processes (**E**) (Supp. Eq. 4.1).

The extracted ID signatures from the

- TCGA WGS cohort were shown in Supp. Fig. 8C;
- DFCI/MSKCC WES cohort were plotted in Supp. Fig. 10C;
- Philadelphia WES cohort were shown in Supp. Fig. 11C;
- TCGA WES cohort were presented in Supp. Fig. 14.

4.4 Classification of deletions

It has been shown previously, that cancer cells exhibiting homologous recombination deficiency, have unique characteristics in their indel profiles. Specimens with biallelic *BRCA1/2* mutations have significantly more deletions that are longer than 10 bp than *BRCA1/2* wild-type tumors, and they also tend to have more deletions than insertions [13]. It has also been found, that these deletions mostly arise due to the activity of the Microhomology Mediated End Joining (MMEJ) or the Single Strand Annealing (SSA) DNA repair pathways, and thus the relative ratio of microhomology-mediated deletions among them is significantly higher than in HR-competent cases [36].

In general, deletions were classified into three groups: (1) complete repetitions; when the complete deleted sequence is repeated after the deletion in the reference genome, (2) microhomology-mediated deletions; when only the first n nucleotides of the deleted sequence is repeated after the deletion, and (3) unique deletions; when the sequence following the deletion has no resemblance to the deleted series of nucleotides. However, since the repetition of the first 1-2 nucleotides could occur by pure chance (with 0.25 and 0.0625 probabilities, respectively, assuming that all 4 nucleotides can occur with the same probability), when investigating the effects of the MMEJ/SSA pathway, it is considered a good practice to work with the $n \geq 3$ microhomologies in WGS and $n \geq 2$ microhomologies in WES samples.

The deletion profile (the relative ratio of the three deletion classes) of each sample in the

- TCGA WGS cohort was shown in Supp. Fig. 9A;
- DFCI/MSKCC WES cohort was plotted in Supp. Fig. 10D;
- Philadelphia WES cohort was shown in Supp. Fig. 11D;
- TCGA WES cohort was presented in Supp. Fig. 15.

5 Structural variant calling

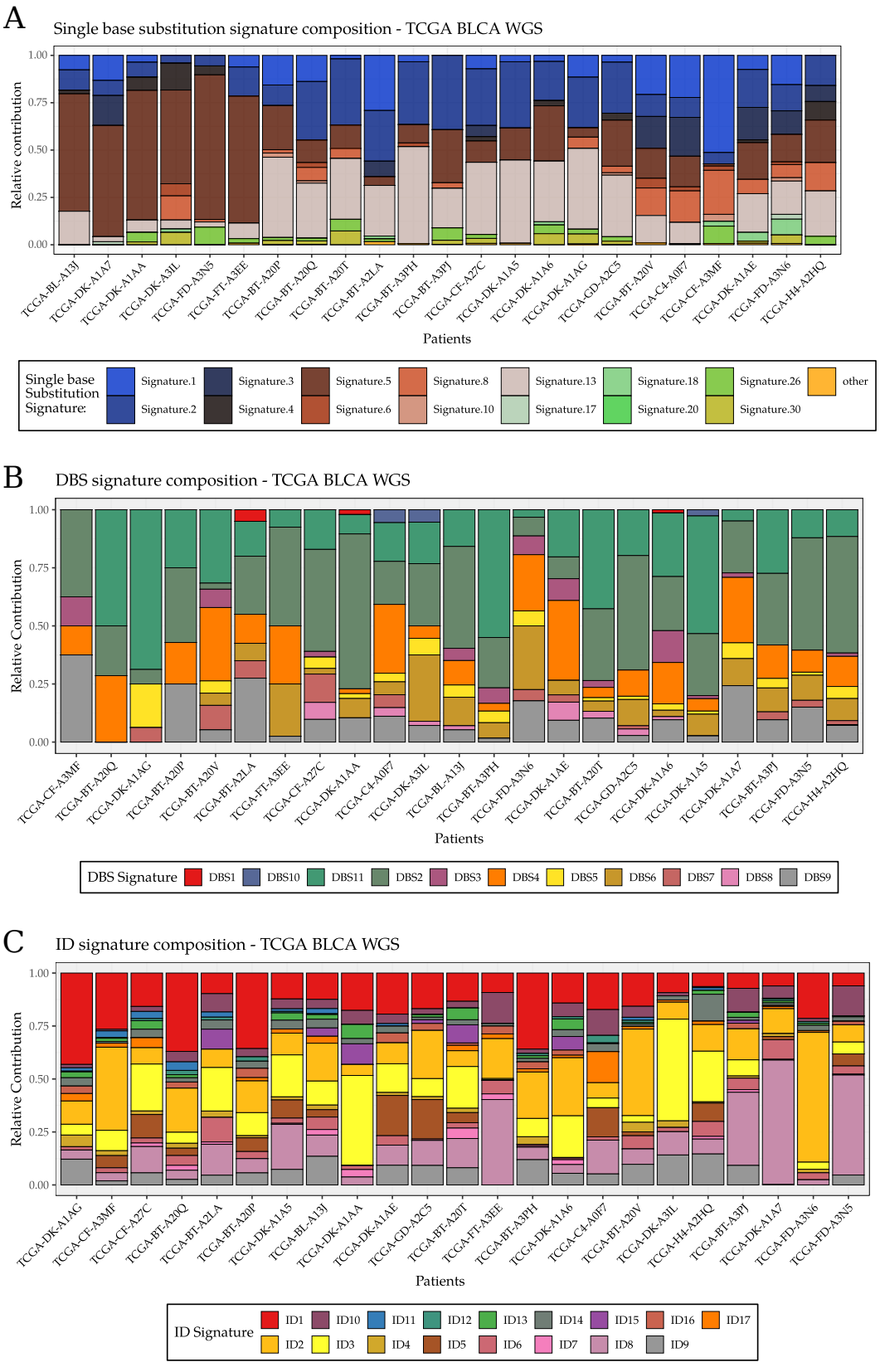
Structural variants (SVs) were called using `BRASS` (version 6.0.0) (`https://github.com/cancerit/BRASS`). In the analysis only those variants were taken into consideration, which were supported by at least 6 read pairs that were successfully *de novo* assembled by `velvet` [38].

5.1 Rearrangement signatures

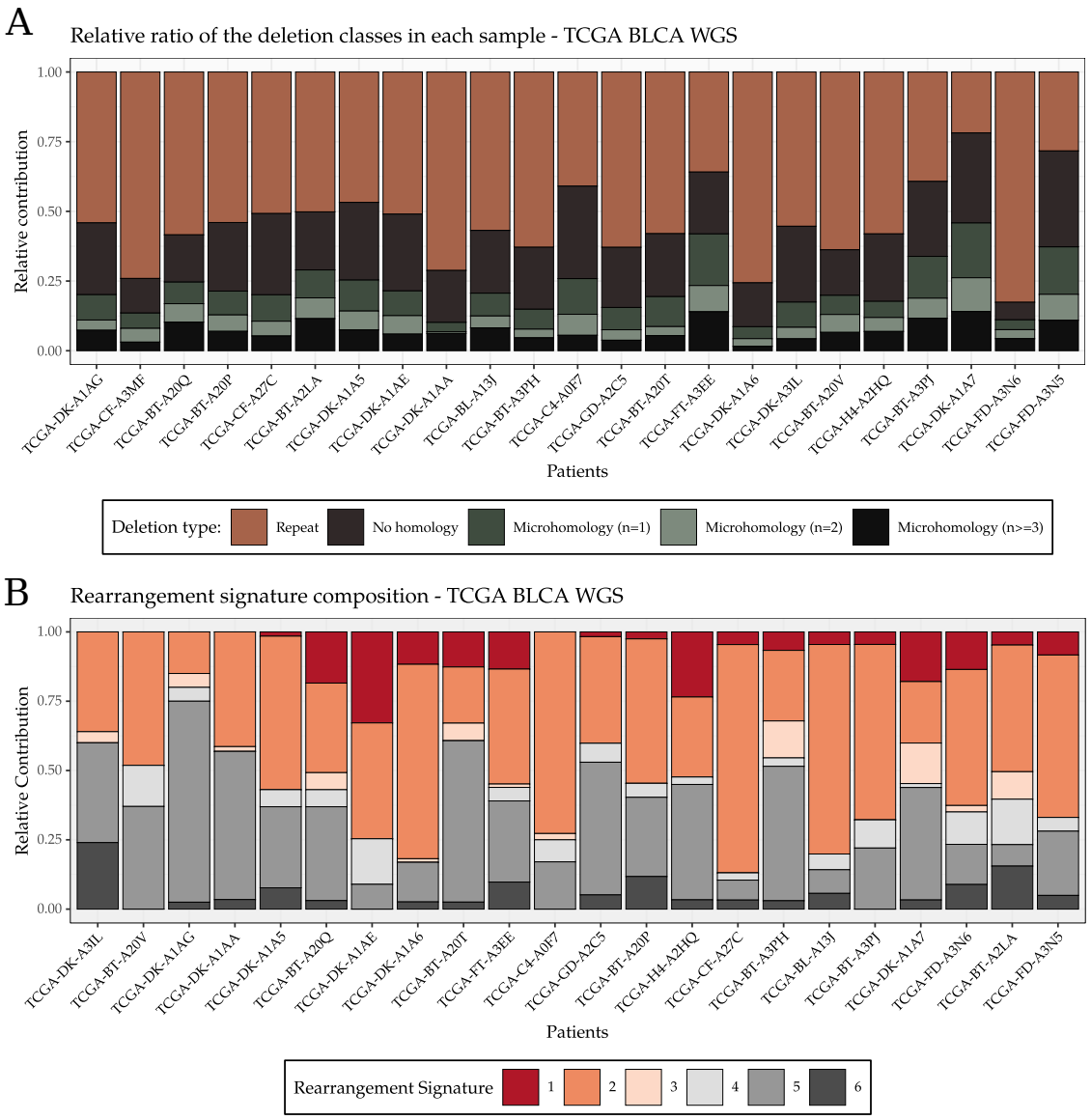
The resulting structural variants in each sample were mapped to a 32-dimensional rearrangement signature (RS) catalog described in breast cancer (**M**) [26]. The previously identified matrix of rearrangement signatures (**P**) was downloaded from the following link

- `https://static-content.springer.com/esm/art%3A10.1038%2Fnature17676/MediaObjects/41586_2016_BFnature17676_MOESM47_ESM.zip`

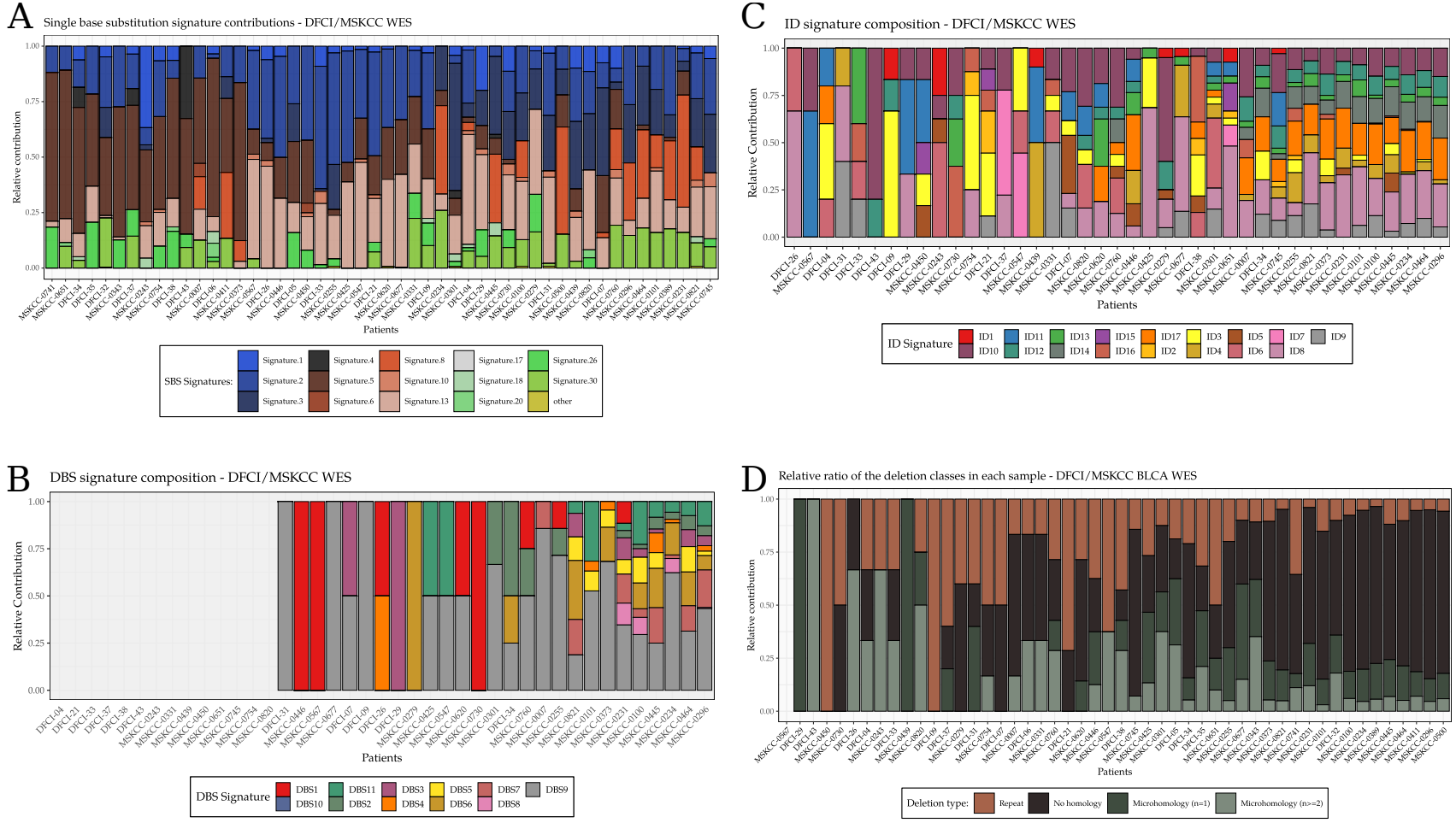
As previously, the **M** and the **P** matrices were used in a non-negative least-squares problem to estimate the matrix of exposures to mutational processes (**E**) (Supp. Eq. 4.1).



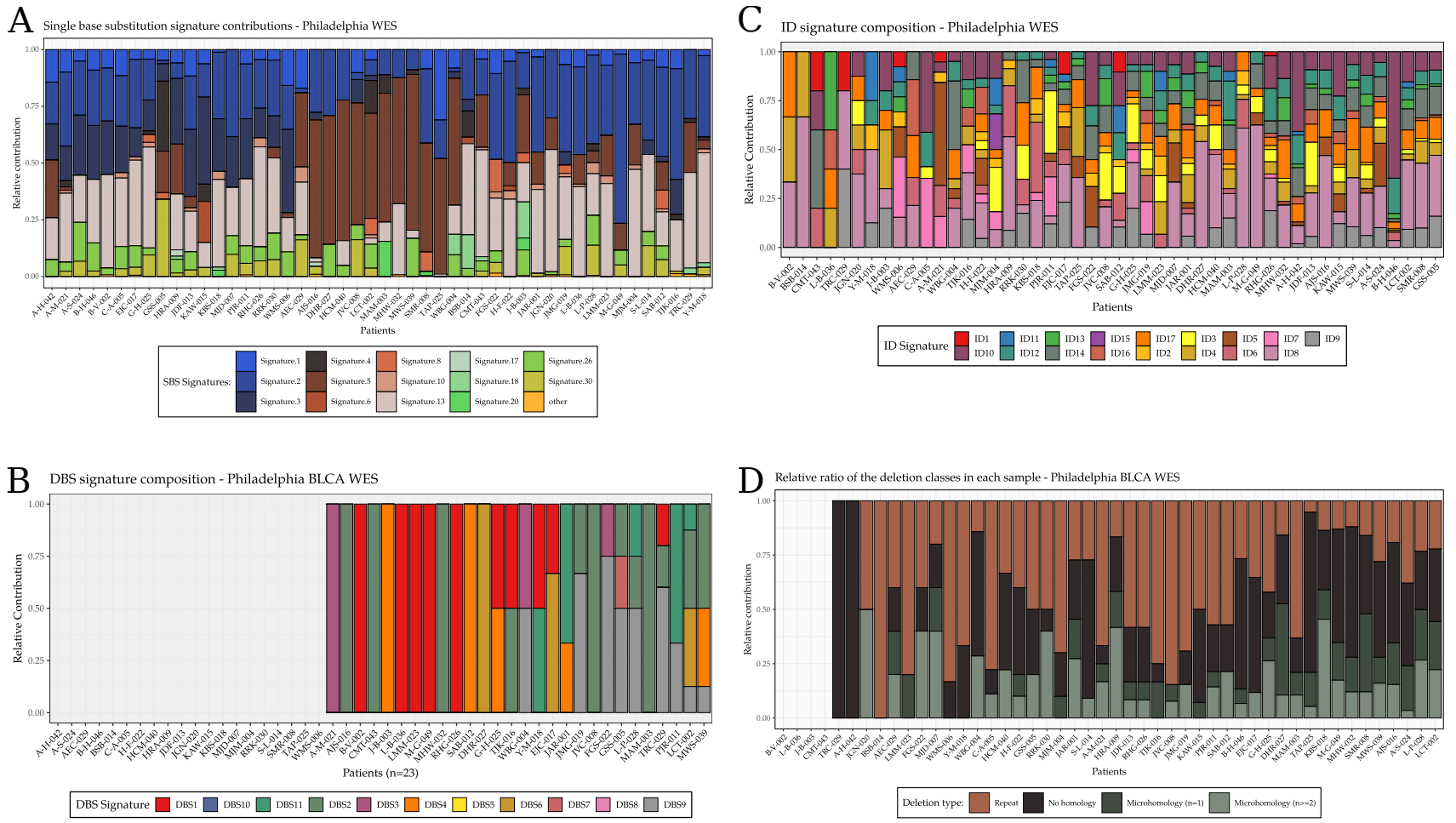
Supp. Fig. 8: TCGA BLCA WGS: SBS signature (COSMIC v2) (**A**), DBS signature (**B**), ID signature (**C**) composition of the samples.



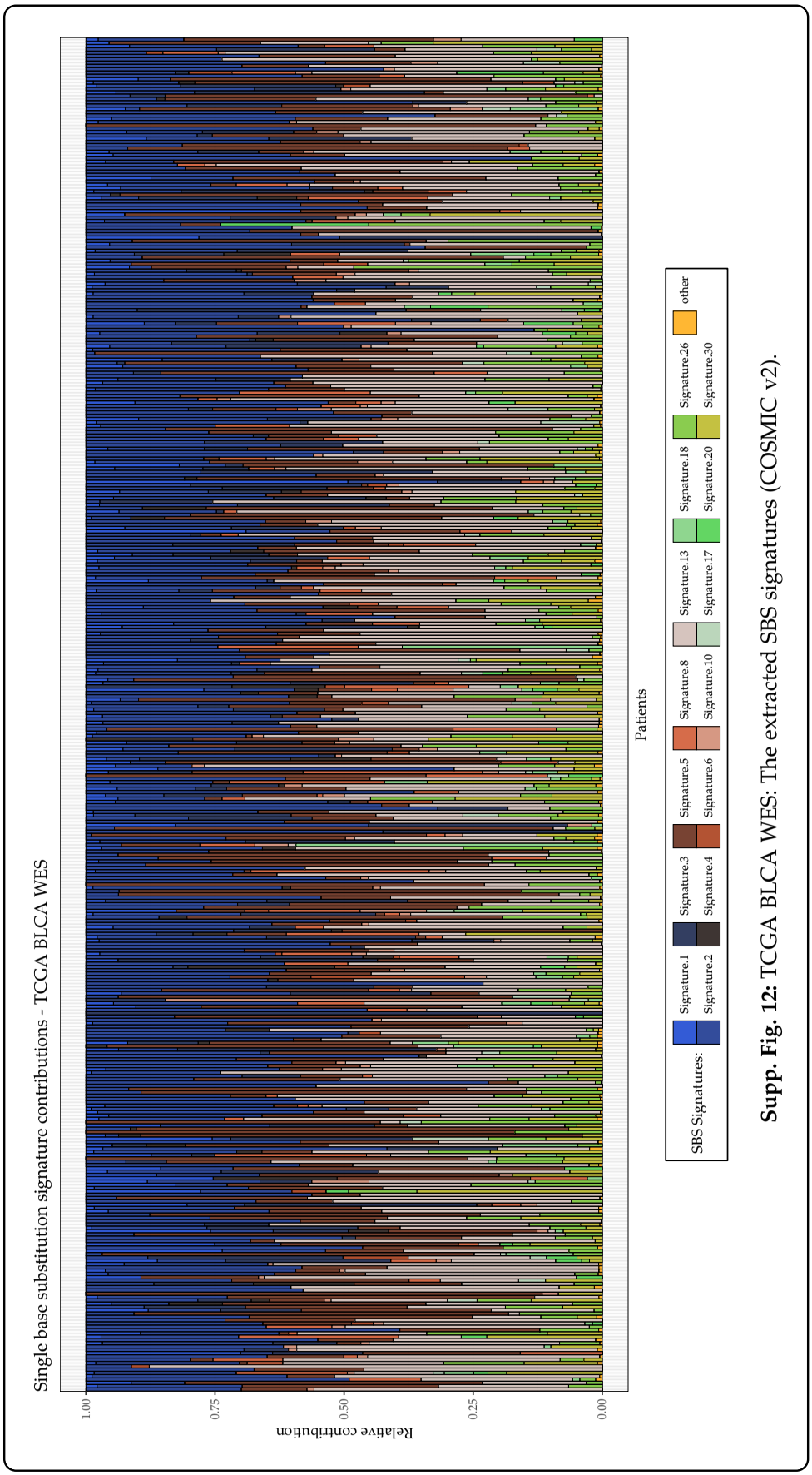
Supp. Fig. 9: TCGA BLCA WGS: Classification of deletions in each sample (**A**). RS composition of the samples (**B**).



Supp. Fig. 10: DFCI/MSKCC BLCA WES: SBS signature (COSMIC v2) composition (**A**), DBS signature composition (**B**), ID signature composition (**C**) and deletion profiles of the samples (**D**).



Supp. Fig. 11: Philadelphia BLCA WES: SBS signature (COSMIC v2) composition (**A**), DBS signature composition (**B**), ID signature composition (**C**) and deletion profiles of the samples (**D**).

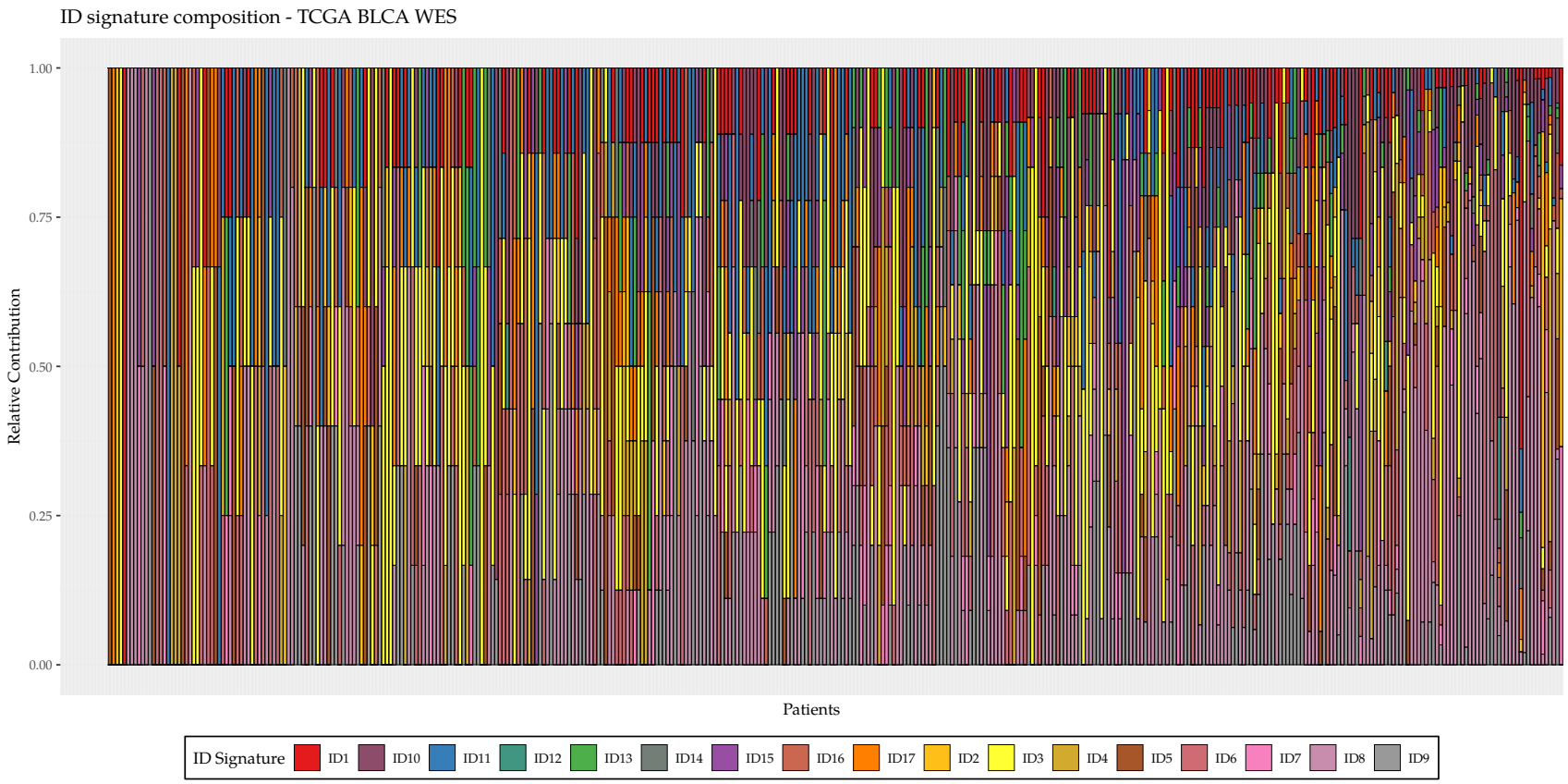


Supp. Fig. 12: TCGA BLCA WES: The extracted SBS signatures (COSMIC v2).

Online figure. Zoom in for details.

Supp. Fig. 13: TCGA BLCA WES: The extracted DBS signatures.

Online figure. Zoom in for details.



Supp. Fig. 14: TCGA BLCA WES. ID signature composition of each sample.

Online figure. Zoom in for details.

Supp. Fig. 15: TCGA BLCA WES. Classification of deletions in each sample.

6 Genomic scar scores

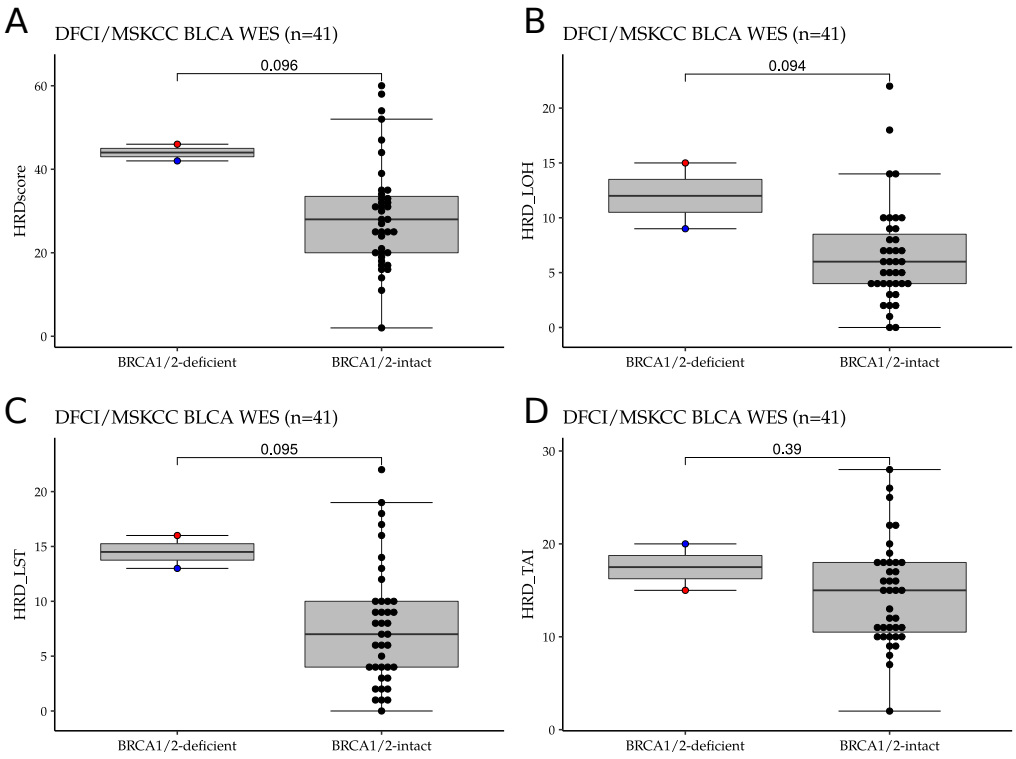
Three independent DNA-based measures of genomic instability using single nucleotide polymorphism (SNP) arrays were developed on the basis of

- homologous recombination deficiency (HRD)-associated loss of heterozygosity (HRD-LOH) [1],
- telomeric allelic imbalance (HRD-TAI) [6], and
- large-scale state transition (HRD-LST) [27].

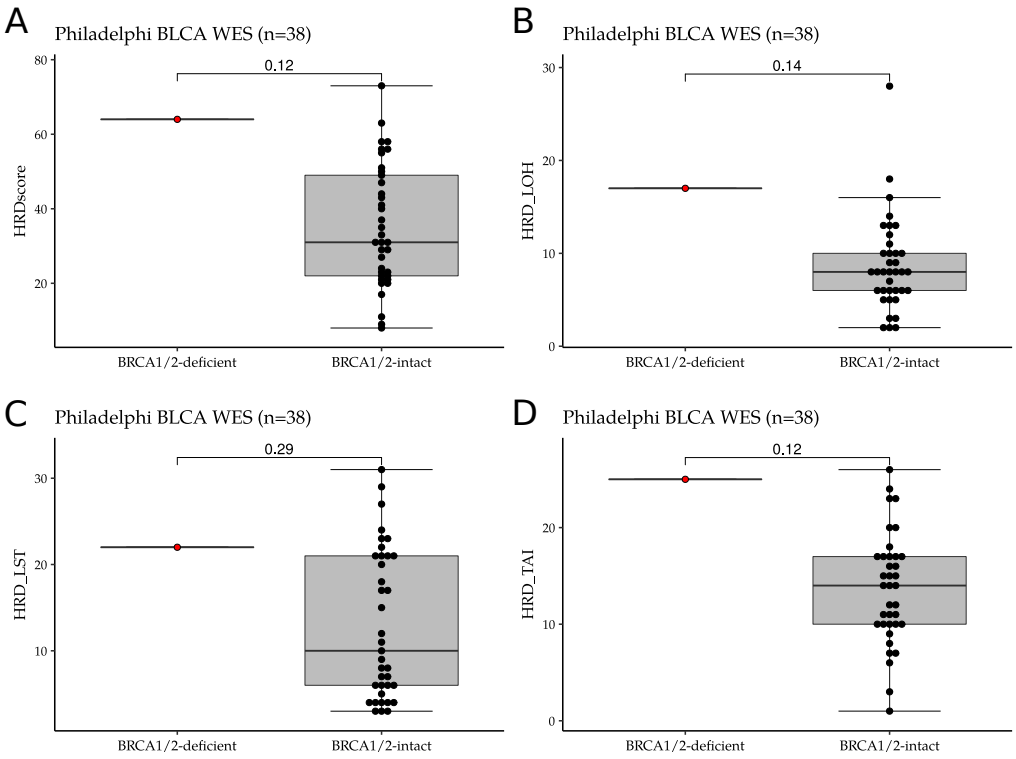
The **HRD-LOH score** is the number of 15 Mb exceeding LOH regions that do not cover the whole chromosome [1]. The **HRD-TAI score** is defined as the number of allelic imbalances (AIs) that extend to the telomeric ends of a chromosome without crossing its centromere [6]. The **HRD-LST score** is the number of chromosomal breaks between adjacent regions of at least 10 Mb with a distance between them not larger than 3 Mb [27].

All three individual scores were highly correlated with defects in *BRCA1/2* and other HR pathway genes in breast or ovarian cancer, and were associated with sensitivity to platinum agents [1] [6] [27] [34]. The aggregated form of these three measures are often referred to as the HRD score [35] [34]. Although these genomic scar scores were developed on SNP arrays our group has previously demonstrated their applicability using WES- and WGS-derived CN profiles [33].

As it was described above, `Sequenza` was used to estimate the copy number profile of the samples in the three bladder cancer cohorts. The three genomic scar scores were calculated for each sample with the help of the `scarHRD` [33] R package and presented in Fig. 2 for the TCGA, in Supp. Fig 16 for the DFCI/MSKCC, and in Supp. Fig. 17 for the Philadelphia BLCA WES cohorts. The comparison of the sum of the genomic scar scores (= HRD score) of WES and WGS samples from the TCGA BLCA cohort showed a very strong ($R_{Pearson} = 0.93$) correlation as it was shown in Supp. Fig 27.



Supp. Fig. 16: DFCI/MSKCC BLCA WES. Distribution of genomic scar scores. The identified *BRCA1/2*-deficient samples had elevated genomic scar scores, although the sample size was too small to reach statistical significance. On the boxplots the midline represents the median, the two edges of the box represent the lower and upper IQR, the upper whisker $= \min(\max(x), Q3 + 1.5 \times \times IQR)$ and the lower whisker $= \max(\min(x), Q1 - 1.5 \times IQR)$. *ERCC2* mutant samples were excluded from the plots (n = 41). P-values were calculated by the Wilcoxon rank-sum test and no mathematical correction was made for multiple comparisons.



Supp. Fig. 17: Philadelphia BLCA WES. Distribution of genomic scar scores. The identified *BRCA1*-deficient sample had elevated genomic scar scores, although the sample size was too small to reach statistical significance. On the boxplots the midline represents the median, the two edges of the box represent the lower and upper IQR, the upper whisker $= \min(\max(x), Q3 + 1.5 \times \times IQR)$ and the lower whisker $= \max(\min(x), Q1 - 1.5 \times IQR)$. *ERCC2* mutant samples were excluded from the plots (n = 38). P-values were calculated by the Wilcoxon rank-sum test and no mathematical correction was made for multiple comparisons.

7 HRDetect

A lasso logistic regression model, called HRDetect [12], was trained on 560 breast cancer WGS samples to identify tumors that exhibit the signs of *BRCA1/2* deficiency. The weights of the genomic features contributing to the model were shown in Supp. Table 8.

Predictor	Weight
Proportion of deletions with microhomology	2.398
Substitution signature 3	1.611
Rearrangement signature RS3	1.153
Rearrangement signature RS5	0.847
HRD index	0.667
Substitution signature 8	0.091
Intercept	-3.364

Supp. Table 8: The weights of the HRDetect model [12]. The genomic feature called "HRD index" corresponds to the "HRD-LOH score" term used here.

7.1 Data transformation

In order to reduce right skewness of the data and to ensure that the distributions of the features more resemble to Gaussian curves, the input variables (x_i) were log-transformed, according to the following formula:

$$x_i' = \ln(x_i + 1). \tag{7.1}$$

The constant shift was added to keep the $x_i = 0$ values away from $-\infty$.

The log-transformed data were standardized (each feature had a mean of 0 and a standard deviation of 1) to make the variables comparable to one another.

$$x_i'' = \frac{x_i' - \mathbb{E}[x_i']}{\sigma(x_i')} \tag{7.2}$$

7.2 HRDetect WGS

The TCGA database contains only 23 WGS bladder cancer cases; therefore, we could not train a new logistic regression model. Instead, we used the weights of the original HRDetect model to calculate the scores of the WGS bladder cancer samples (Supp. Table 8).

7.3 HRDetect WES

In order to calculate the HRDetect scores of the BLCA WES samples, an alternative HRDetect model was used which was retrained on 560 artificially derived (from WGS samples) breast cancer WES samples [15].

8 DNA methylation and RNA expression analysis

TCGA BLCA WES

A systematic screen for hypermethylation of DDR genes in the TCGA dataset revealed that 39 DDR genes were hypermethylted in 32 tumor types [25]. Interestingly, most exclusively and frequently methylated gene (37%) in the TCGA BLCA cohort was *RBBP8* (encoding the protein CtIP) [25]. In addition, a significantly negative correlation between *RBBP8* methylation and *RBBP8* mRNA expression was only demonstrated for bladder cancer [25].

CtIP is a major HR repair factor; it has a key role in DNA double-strand break (DSB) end resection. Furthermore, it was shown that besides its fundamental role in DSB resection, CtIP is a critical regulator of DNA replication fork integrity upon replication stress [28] [37].

DNA methylation data measured by the Illumina HumanMethylation450 platform were downloaded from the Xena platform (`https://tcga.xenahubs.net`) [20]. Using the *cpgCollapse* function from the `minfi` R package [5], promoter-associated CpG loci were divided into clusters with a default maximum gap of 500 bp and a maximum cluster gap of 1500 bp. The mean β value of the loci within each cluster was calculated and used as a single methylation estimate per cluster. In the TCGA BLCA cohort, methylation clusters of *RBBP8* showed high correlation (see Supp. Fig. 9), thus the mean methylation of the clusters was used to determine the methylation status of *RBBP8* in a given sample (Supp. Eq. 8).

$$\beta_{mean} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{N}\sum_{i=1}^{N} c_{ij}, \tag{8.1}$$

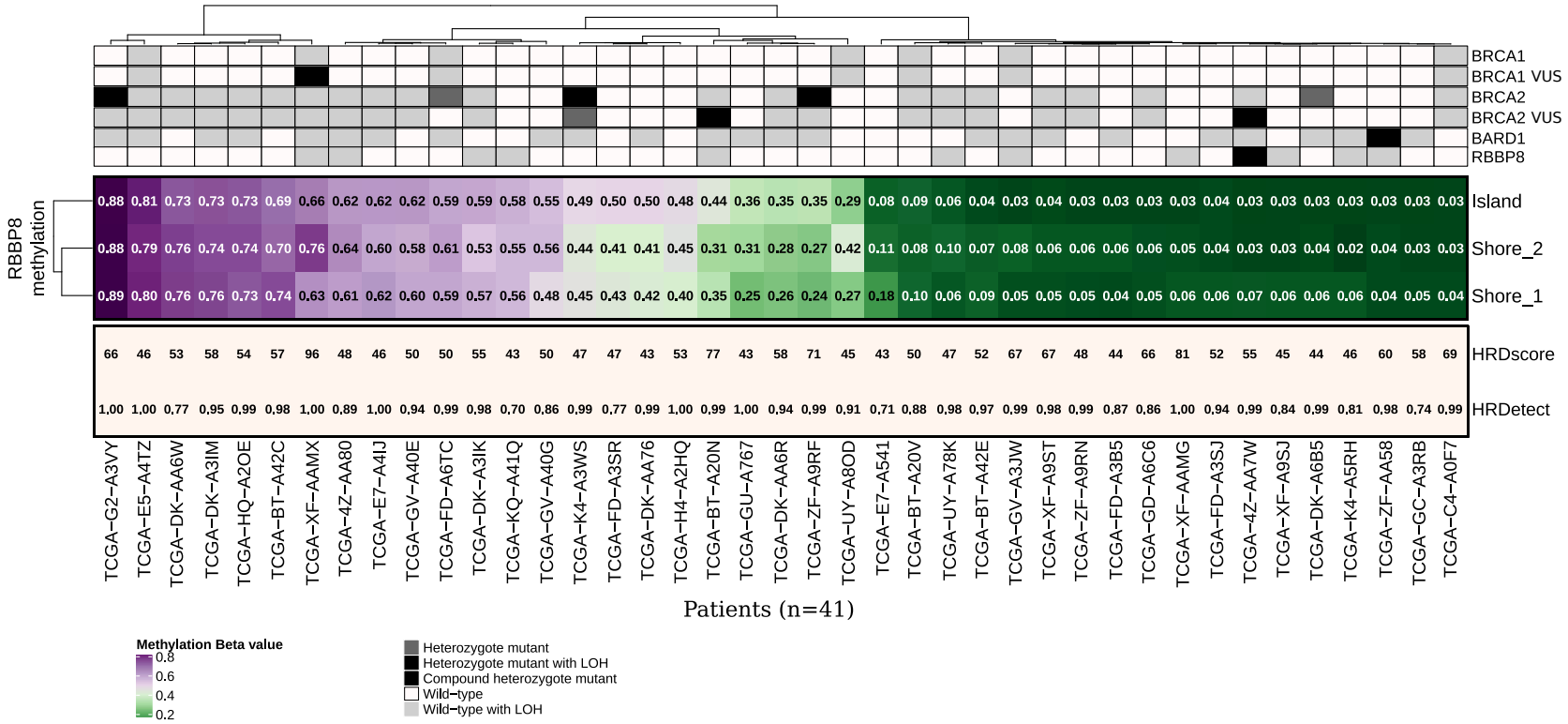
where c_{ij} is the methylation value at the ith CpG position within the jth CpG cluster, N is the number of CpG loci in a given CpG cluster, and M is the number of CpG clusters in a given gene. Based on their methylation status samples were divided into two groups using 0.5 as a cut-off value (Supp. Table 9).
Methylation data were not available for the DFCI/MSKCC and Philadelphia BLCA WES cohorts.

	$\beta_{mean} \leq 0.5$	$\beta_{mean} > 0.5$	Total
HRDetect ≥ 0.7	53	20	73
HRDetect < 0.7	268	54	322
Total	321	74	395
HRDscore ≥ 42	61	26	87
HRDscore < 42	260	48	308
Total	321	74	395

Supp. Table 9: TCGA BLCA WES. Number of samples with low ($\beta_{mean} \leq 0.5$) and high ($\beta_{mean} > 0.5$) methylation values.

RNA expression data were downloaded from the Xena platform (https://gdc.xenahubs.net) [20]. The Fragments Per Kilobase of transcript per Million mapped reads (FPKM) normalization method was used and the data were then log-transformed according to the following formula

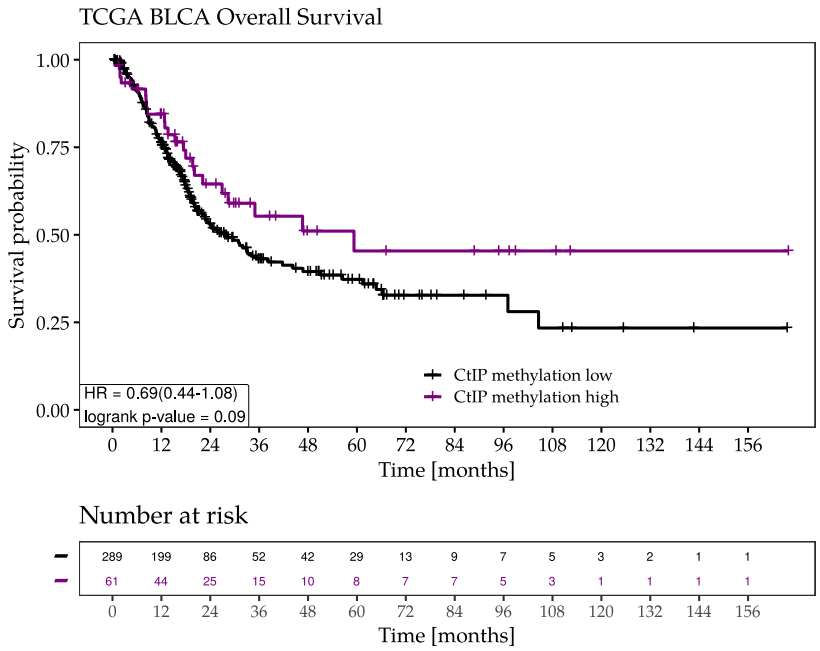
$$y = \log_2 (FPKM + 1)\,. \tag{8.2}$$



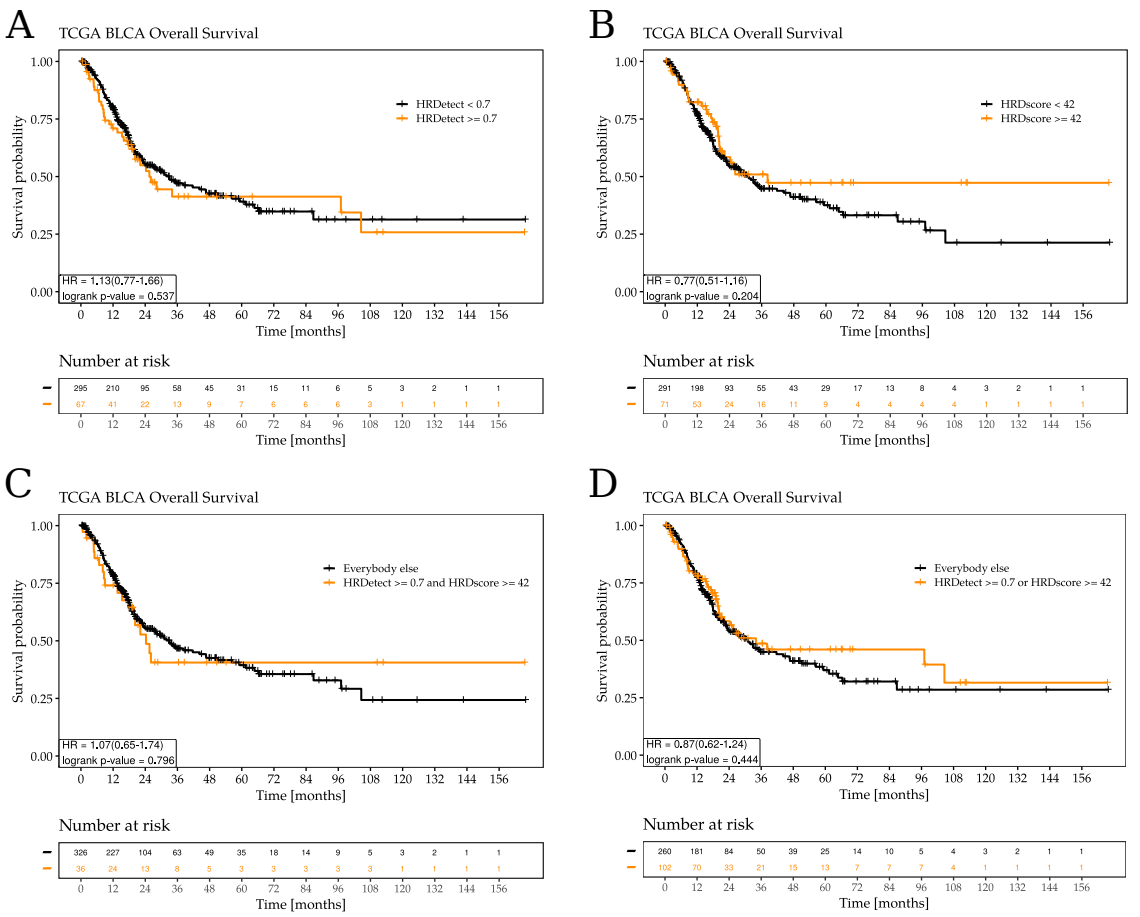
Supp. Fig. 18: TCGA BLCA WES. The heatmap shows the mean methylation values of the CpG clusters within *RBBP8* in the 41 samples with high HRD score ($\geq$ 42) and HRDetect score ($\geq$ 0.7).

9 Survival analysis

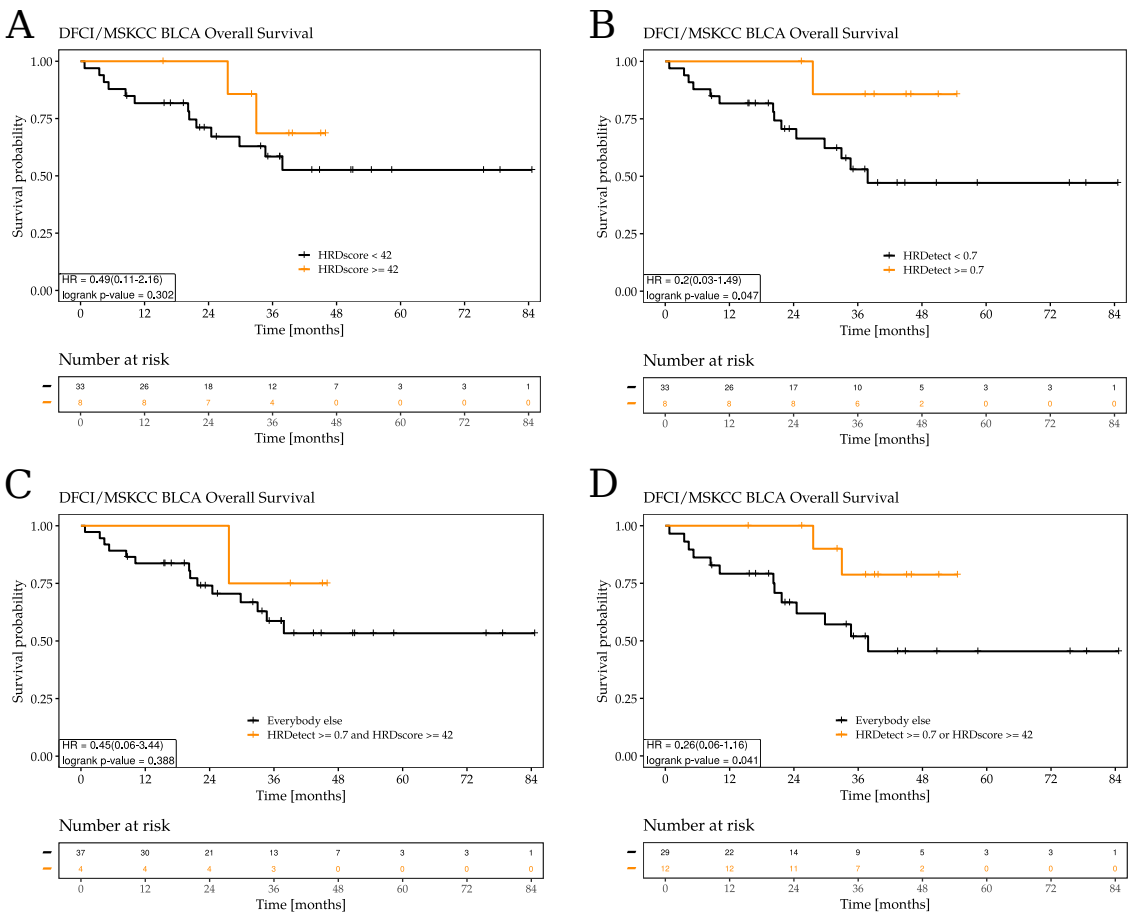
Survival analysis was carried out using the `survival` and `survminer` R packages. High CtIP methylation ($>$ 0.5) was associated with better overall survival in the TCGA BLCA cohort (Supp. Fig. 19). HRDetect score or the combination of HRDetect score and HRD score was not associated with better overall survival, although samples with high HRD score ($\geq$ 42) demonstrated a strong trend towards improved survival in the TCGA BLCA cohort (Supp. Fig. 20). The estimated Kaplan-Meier curves of patients who received neoadjuvant cisplatin-based chemotherapy were shown in Supp. Fig. 21 for the DFCI/MSKCC cohort and in Supp. Fig. 22 for the Philadelphia cohort. In the DFCI/MSKCC cohort, patients with high HRDetect score ($\geq$ 0.7) and patients with either high HRDetect score ($\geq$ 0.7) or HRD score ($\geq$ 42) demonstrated significantly better survival compared with patients having low HRDetect score and low HRDetect or low HRD score, respectively (Supp. Fig. 21). In the Philadelphia cohort, no association with improved survival was observed (Supp. Fig. 22). P-values were calculated using the log-rank test.



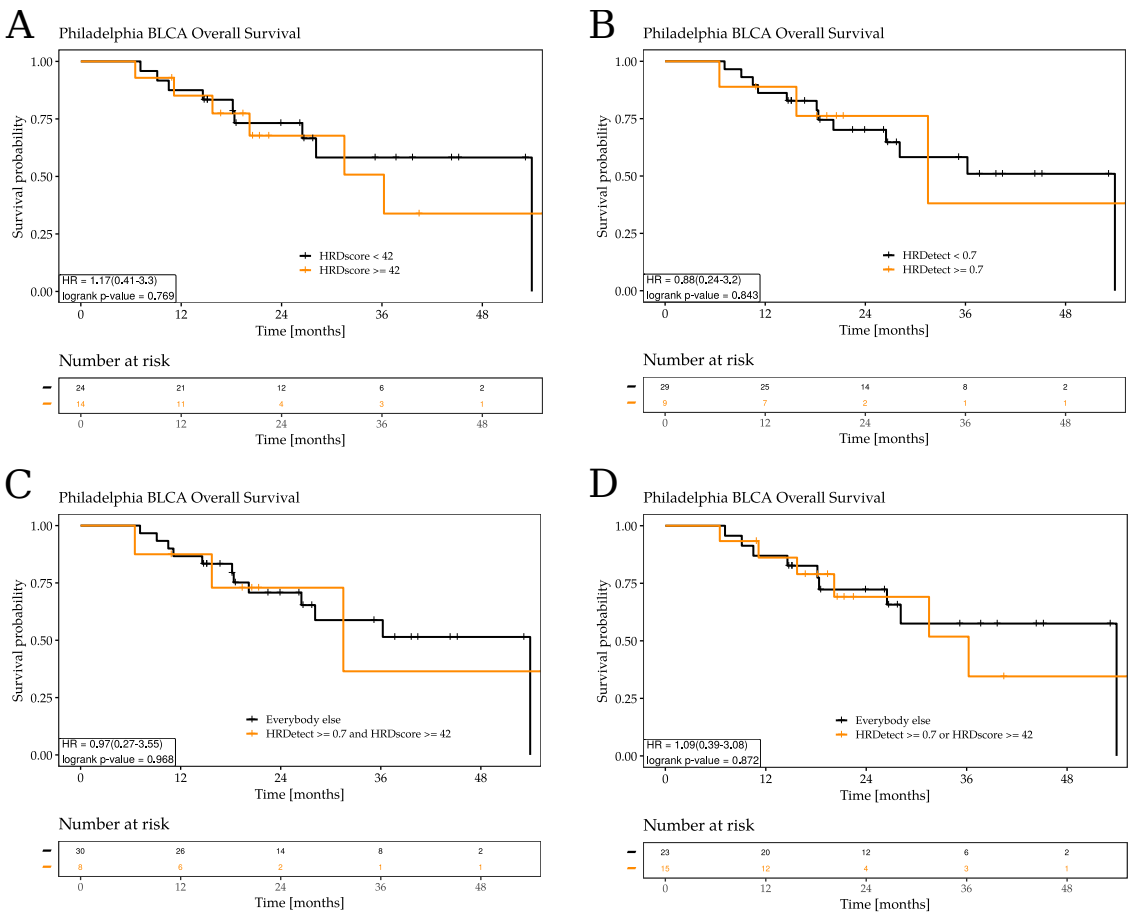
Supp. Fig. 19: TCGA BLCA WES. High CtIP methylation ($>$ 0.5) was associated with better overall survival. In order to see the effect of *RBBP8* methylation on overall survival, *BRCA1/2*-deficient samples, *ERCC2* mutant samples and samples with high ERCC2mut score ($\geq$ 0.7) were excluded from the plot.



Supp. Fig. 20: TCGA BLCA WES. HRDetect score (**A**) or the combinations of HRDetect score and HRD score (**C**, **D**) was not associated with better overall survival, although samples with high HRD score ($\geq$ 42) (**B**) demonstrated a strong trend towards improved overall survival. *ERCC2* mutant samples and samples with high ERCC2mut score ($\geq$ 0.7) were excluded from the plots.



Supp. Fig. 21: DFCI/MSKCC BLCA WES. Patients with high HRDetect score ($\geq$ 0.7) (**B**) and patients with either high HRDetect score ($\geq$ 0.7) or HRD score ($\geq$ 42) (**D**) demonstrated significantly better survival compared with patients having low HRDetect score and low HRDetect or low HRD score, respectively. *ERCC2* mutant samples were excluded from the plots.

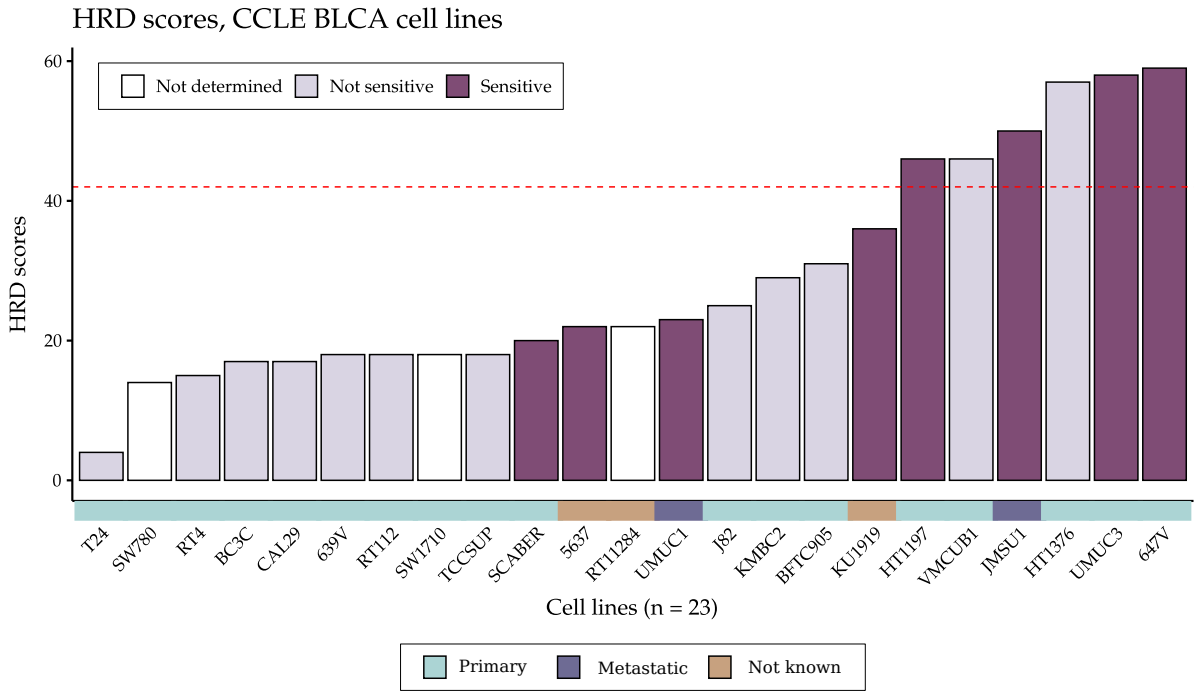


Supp. Fig. 22: Philadelphia BLCA WES. No association was observed between (**A**) high HRD score ($\geq$ 42), or (**B**) high HRDetect score ($\geq$ 0.7), or (**C-D**) the combinations of the two measures and improved overall survival. *ERCC2* mutant samples were excluded from the plots.

10 BLCA cell line data

10.1 HRDscore

The CCLE mutation and CN calls [9] [24] and ABSOLUTE copy number analysis results [19] were downloaded from the DepMap (`https://depmap.org/portal/download/`) portal and the `scarHRD` [33] R package was used to calculate the HRD score of BLCA cell lines.



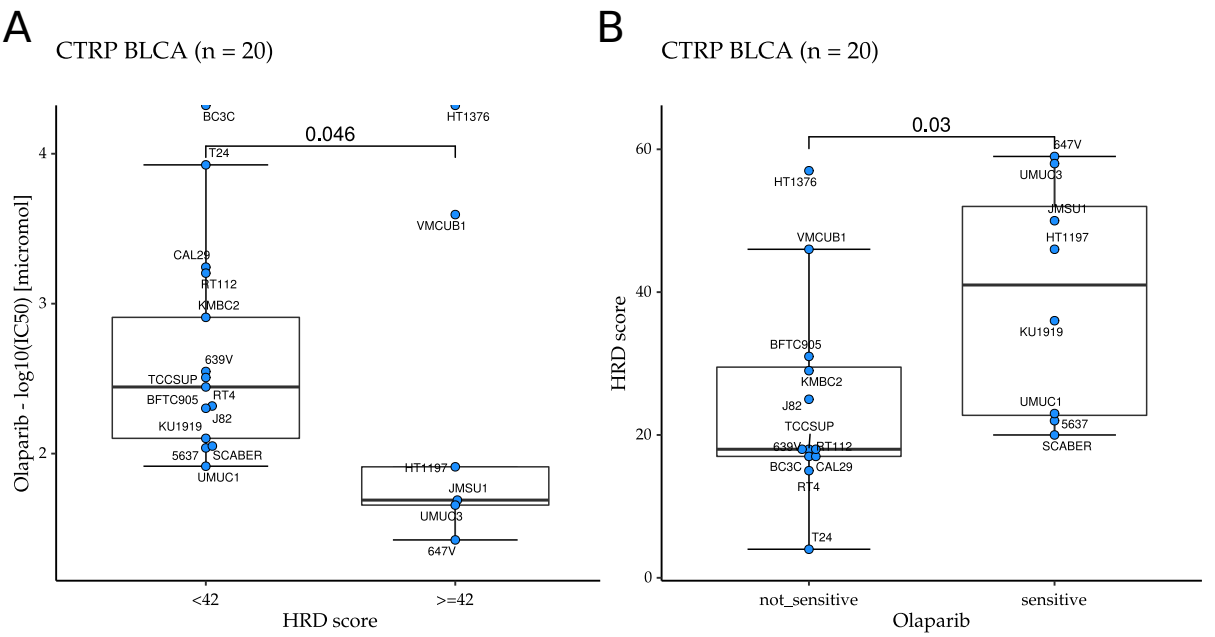
Supp. Fig. 23: HRD score of BLCA cell lines. From the 23 BLCA cell lines 6 had an HRD score ≥ 42.

Broad ID	ACH-000753
Cell line name	JMSU1
Hugo symbol	*PALB2*
NCBI build	37
Genome change	g.chr16:23646980_23646981insT
Codon change	c.(886-888)atgfs
Protein change	p.M296fs
Is deleterious	Yes
Is LOH	Yes

Supp. Table 10: We found a deleterious *PALB2* mutation accompanied by LOH in JMSU1 metastatic BLCA cell line which had a high HRD score (≥ 42).

10.2 Olaparib sensitivity

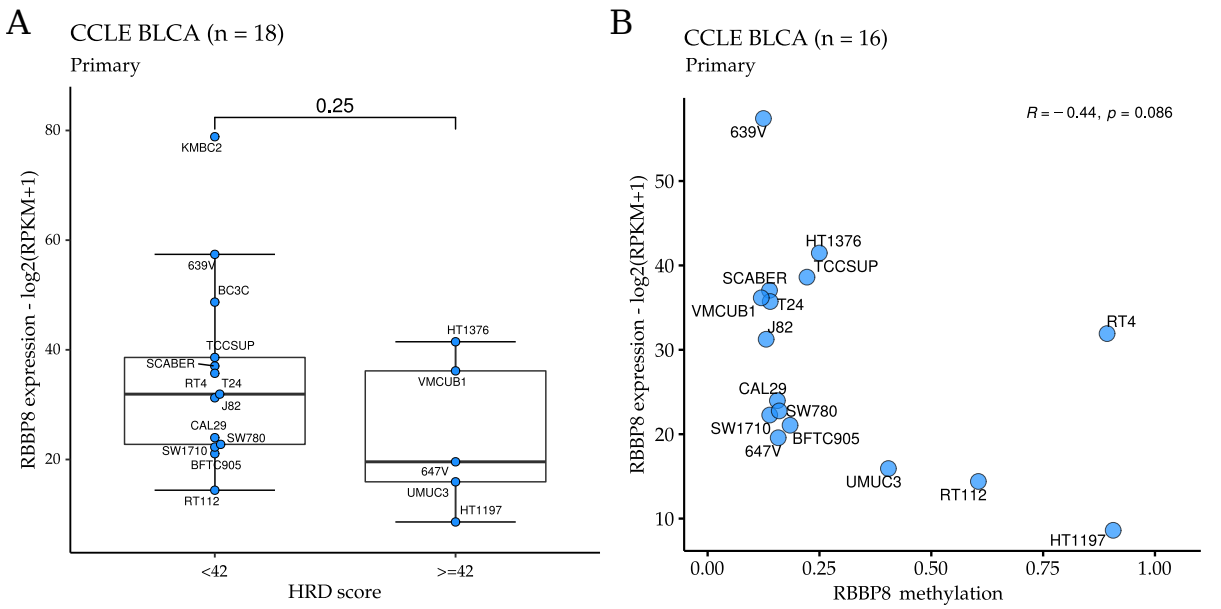
The `PharmacoGx` [32] R package was used to download and analyze CTRPv2 (Cancer Therapeutics Response Portal version 2) cell line sensitivity data [30].



Supp. Fig. 24: Olaparib sensitivity data of BLCA cell lines from the CTRPv2 database. **A.** BLCA cell lines with a high HRD score (≥ 42) had significantly lower IC_{50} values for olaparib compared to BLCA cell lines with a low HRD score (< 42). The only exception was the cell line called VMCUB1, which had a high IC_{50} value for olaparib in the CTRPv2 data set. However, we think that it was due an incorrect fitting of the dose-response curve. Supporting this suspicion, we checked the IC_{50} value of VMCUB1 for olaparib in the GDSC1 and GDSC2 data sets and in there the values were fallen into the sensitive range. Two additional cell lines (BC3C and HT1376) had an infinite recomputed IC_{50} value in the CTRPv2 data set. The published IC_{50} values for the HT1376 cell line were inconsistent between the GDSC1 and GDSC2 data sets. The y axis shows the logarithm with base 10 of IC_{50} values. **B.** Based on the dose-response curves, cell lines were divided into two categories: "sensitive" - showing at least 50% growth inhibition in the tested dose range, and "not_sensitive" - when this value was not reached within the tested dose range. In the sensitive group the HRD score was significantly higher compared to the nonsensitive group. On the boxplots the midline represents the median, the two edges of the box represent the lower and upper IQR, the upper whisker $= \min(\max(x), Q3 + 1.5 \times IQR)$ and the lower whisker $= \max(\min(x), Q1 - 1.5 \times IQR)$. P-values were calculated by the Wilcoxon rank-sum test.

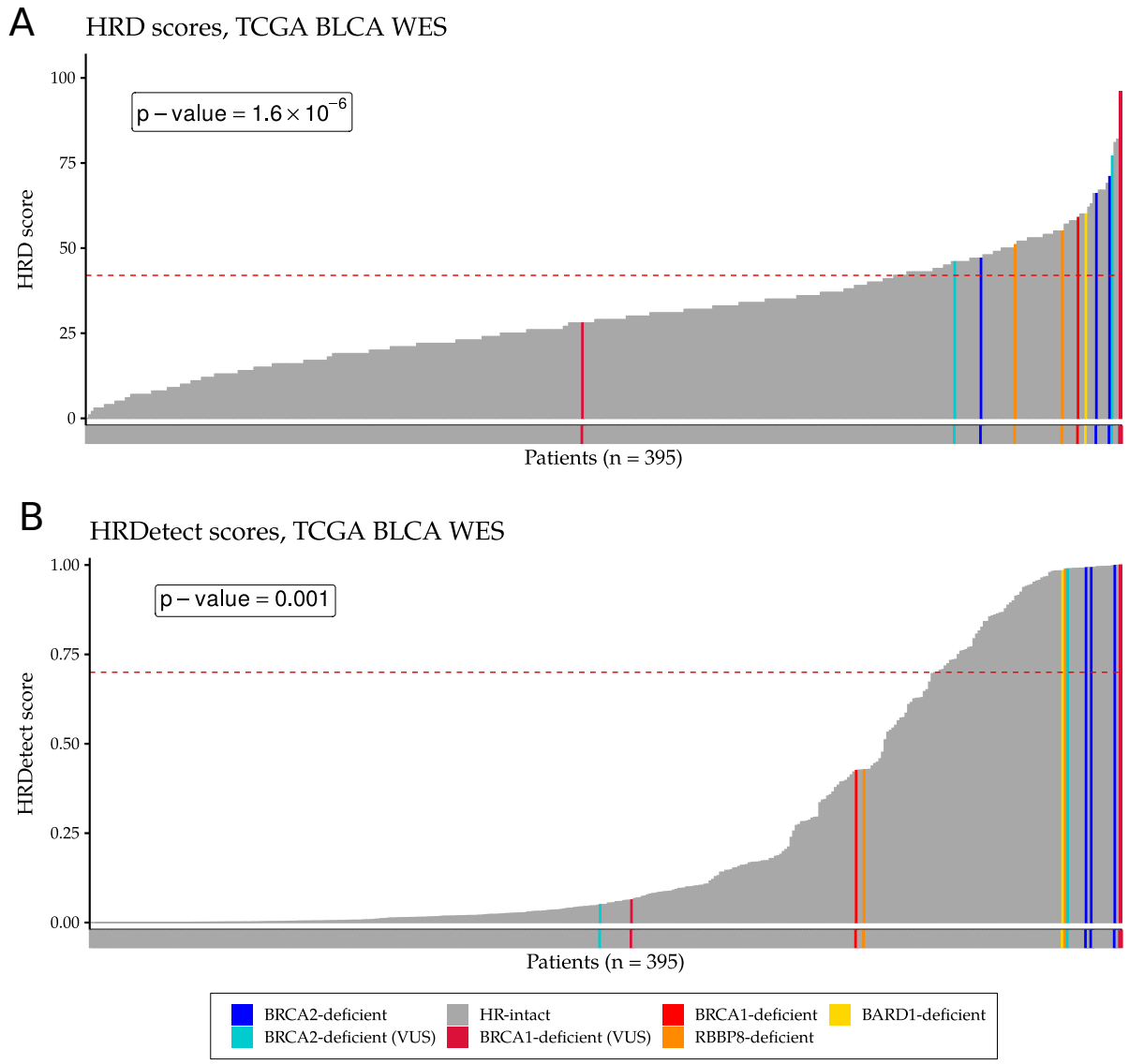
10.3 *RBBP8* expression and methylation

The CCLE RNAseq gene expression data [19] were downloaded from the DepMap (`https://depmap.org/portal/download/`) portal. CCLE DNA methylation raw data measured by the Illumina HumanMethylation450 platform [21] were downloaded from the GEO database [16] using the GSE68379 accession number. The `minfi` R package was used for the analysis in a similar way described in Section 8.

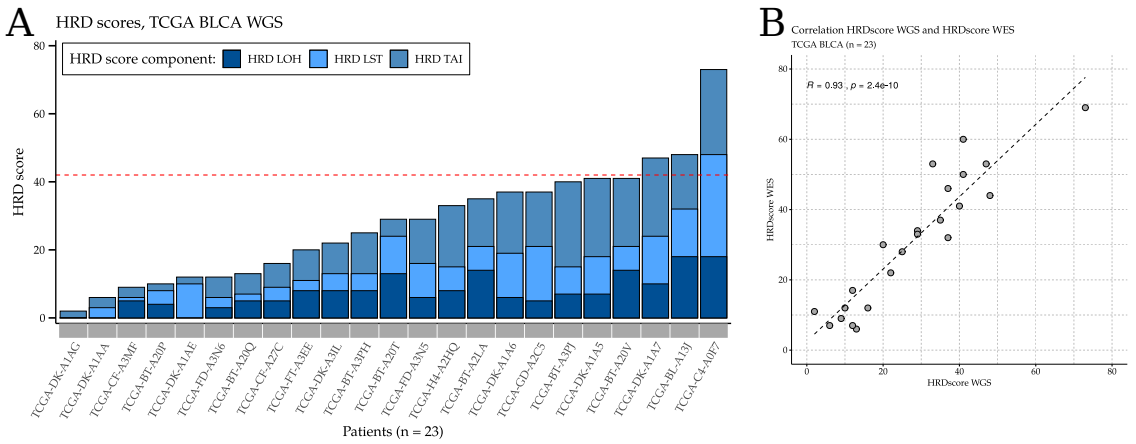


Supp. Fig. 25: BLCA cell lines *RBBP8* expression and methylation analysis. **A.** BLCA cell lines derived form primary tissue sites with high HRD score (≥ 42) showed a trend toward lower *RBBP8* mRNA expression compared to the primary cell lines with low HRD score (< 42). The y axis shows the logarithm with base 2 of the RPKM (Reads Per Kilobase Million) normalized mRNA expression values. P-value was calculated by the Wilcoxon rank-sum test. On the boxplots the midline represents the median, the two edges of the box represent the lower and upper IQR, the upper whisker $= \min(\max(x), Q3 + 1.5 \times IQR)$ and the lower whisker $= \max(\min(x), Q1 - 1.5 \times IQR)$. **B.** A moderate negative correlation ($R_{Pearson} = -0.44$) was observed between *RBBP8* mRNA expression and *RBBP8* promoter methylation. The x axis shows the DNA methylation β values, and the y axis shows the logarithm with base 2 of the RPKM normalized mRNA expression values.

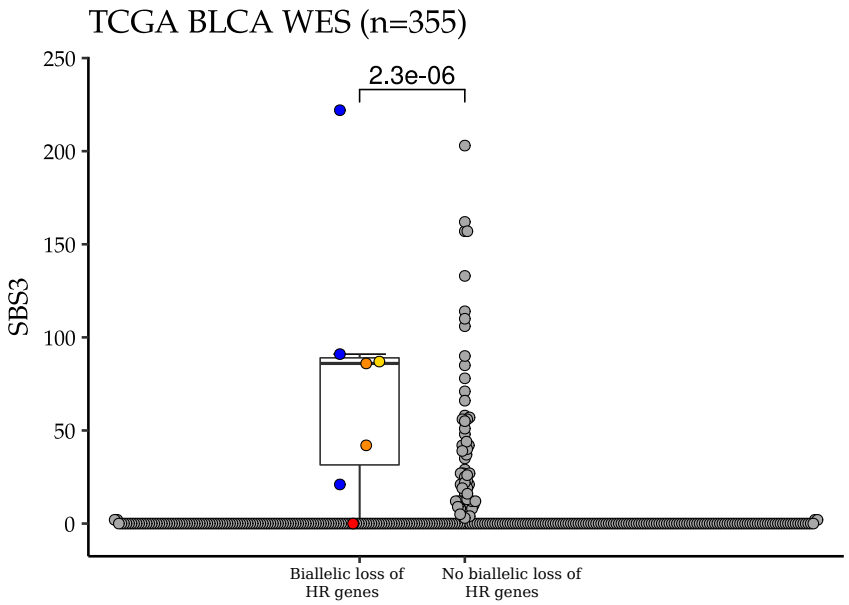
11 Additional supplementary figures (TCGA BLCA)



Supp. Fig. 26: TCGA BLCA WES. HRD (**A**) and HRDetect (**B**) scores of the samples. In addition to the identified HR-deficient samples, *BRCA1/2* variants with uncertain significance with accompanying LOH were colored as well. P-values were calculated by the Fisher exact test.

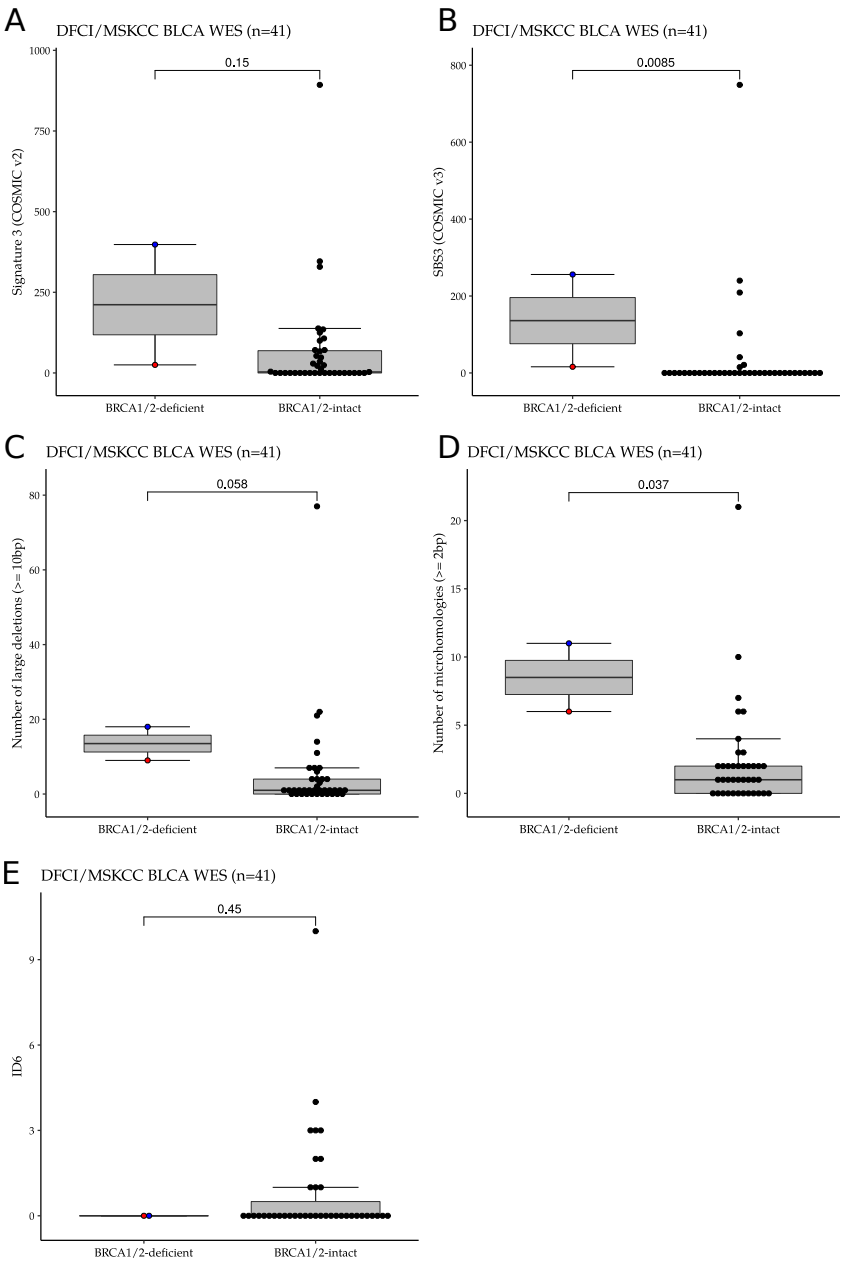


Supp. Fig. 27: TCGA BLCA WGS and WES correlation. **A:** HRD score of the samples broken down into components. **B:** Correlation of HRD score calculated using the WGS and the corresponding WES samples from the TCGA BLCA cohort ($R_{Pearson} = 0.93$).

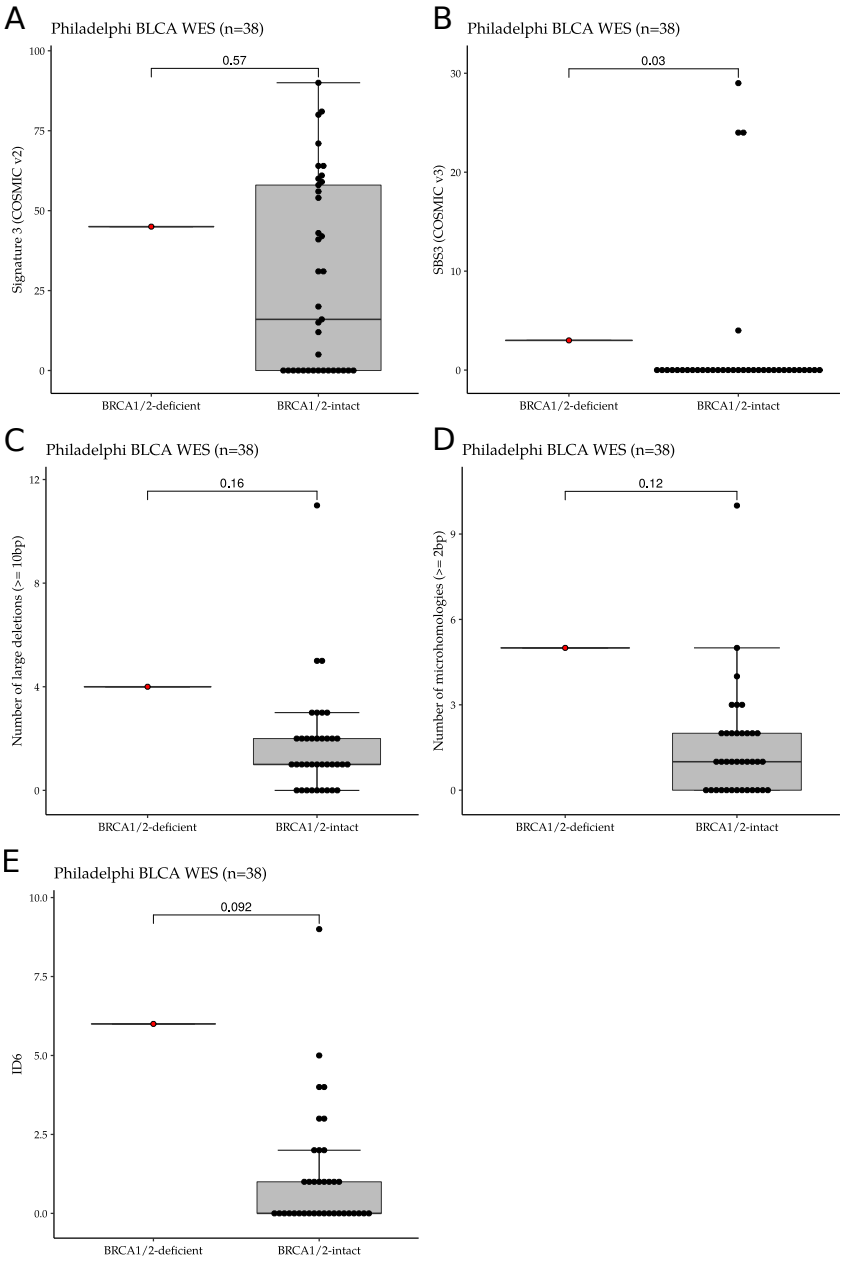


Supp. Fig. 28: TCGA BLCA WES. Signatures significantly elevated in *BRCA1/2*-deficient samples in the TCGA BLCA WES cohort: SBS3 (COSMIC v3). On the boxplots the midline represents the median, the two edges of the box represent the lower and upper interquartile range (IQR), the upper whisker $=$ $\min(\max(x), Q3 + 1.5 \times IQR)$ and the lower whisker $=$ $\max(\min(x), Q1 - 1.5 \times IQR)$. *ERCC2* mutant samples or samples with high *ERCC2mut* score (≥ 0.7) were excluded from the plots (n = 355). P-values were calculated by the Wilcoxon rank-sum test and no mathematical correction was applied for multiple comparisons.

12 Signatures associated with *BRCA1/2* deficiency (DFCI/MSKCC and Philadelphia BLCA)



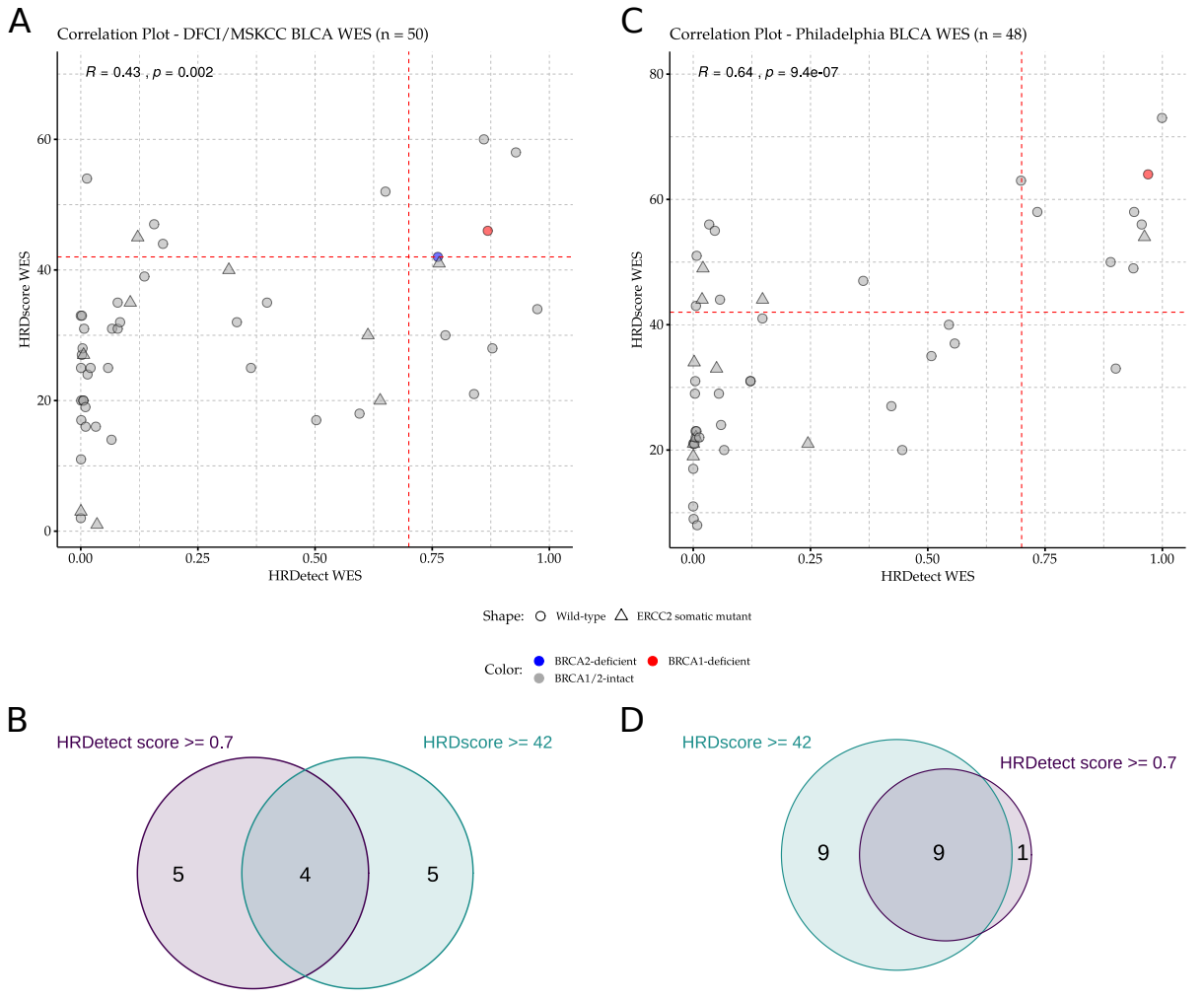
Supp. Fig. 29: DFCI/MSKCC BLCA WES. Distribution of signatures associated with BRCA1/2 deficiency in the DFCI/MSKCC BLCA WES cohort: signature 3 (COSMIC v2) (**A**), SBS3 (COSMIC v3) (**B**), number of large deletions ($\geq$ 10bp) (**C**), number of microhomologies ($\geq$ 2 bp) (**D**), and ID6 (**D**). On the boxplots the midline represents the median, the two edges of the box represent the lower and upper IQR, the upper whisker $= \min(\max(x), Q3 + 1.5 \times IQR)$ and the lower whisker $= \max(\min(x), Q1 - 1.5 \times IQR)$. *ERCC2* mutant samples were excluded from the plots (n = 41). P-values were calculated by the Wilcoxon rank-sum test and no mathematical correction was applied for multiple comparisons.



Supp. Fig. 30: Philadelphia BLCA WES. Distribution of signatures associated with BRCA1/2 deficiency in the Philadelphia BLCA WES cohort: signature 3 (COSMIC v2) (**A**), SBS3 (COSMIC v3) (**B**), number of large deletions ($\geq$ 10bp) (**C**), number of microhomologies ($\geq$ 2 bp) (**D**), and ID6 (**D**). On the boxplots the midline represents the median, the two edges of the box represent the lower and upper IQR, the upper whisker $= \min(\max(x), Q3 + 1.5 \times IQR)$ and the lower whisker $= \max(\min(x), Q1 - 1.5 \times IQR)$. *ERCC2* mutant samples were excluded from the plots (n = 38). P-values were calculated by the Wilcoxon rank-sum test and no mathematical correction was applied for multiple comparisons.

13 HRD score and HRDetect correlation (DFCI/MSKCC and Philadelphia BLCA)

Moderate correlations were observed between HRD scores and HRDetect scores in all three cohorts (TCGA: Fig. 4A, DFCI/MSKCC: 31A, Philadelphia: 31C). The number of samples in each cohort with an HRD score ≥ 42 and an HRDetect score ≥ 0.7 was presented by Venn-diagrams (TCGA: Fig. 4B, DFCI/MSKCC: 31B, Philadelphia: 31D).



Supp. Fig. 31: DFCI/MSKCC and Philadelphia BLCA WES. **A.** Comparison of HRD and HRDetect scores of the samples showed a moderate correlation, $R_{Pearson} = 0.43$, in the DFCI/MSKCC BLCA WES cohort. **B.** As it is visualized by a Venn-diagram, 4 samples had an HRD score ≥ 42 and an HRDetect score ≥ 0.7, including the BRCA1-deficient and the BRCA2-deficient samples. **C.** Comparison of HRD and HRDetect scores of the samples showed a moderate correlation, $R_{Pearson} = 0.64$, in the Philadelphia BLCA WES cohort. **D.** As it is visualized by a Venn-diagram, 9 samples had an HRD score ≥ 42 and an HRDetect score ≥ 0.7, including the BRCA1-deficient sample.

References

[1] V Abkevich, K M Timms, B T Hennessy, J Potter, M S Carey, L A Meyer, K Smith-McCune, R Broaddus, K H Lu, J Chen, T V Tran, D Williams, D Iliev, S Jammulapati, L M FitzGerald, T Krivak, J A DeLoia, A Gutin, G B Mills, and J S Lanchbury. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer*, 107(10):1776–1782, nov 2012.

[2] Ludmil B Alexandrov, Young Seok Ju, Kerstin Haase, Peter Van Loo, Iñigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, Akihiro Fujimoto, Hidewaki Nakagawa, Tatsuhiro Shibata, and Others. Mutational signatures associated with tobacco smoking in human cancer. *Science (80-.).*, 354(6312):618–622, 2016.

[3] Ludmil B Alexandrov, Jaegil Kim, Nicholas J Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R Covington, Dmitry A Gordenin, Erik N Bergstrom, and Others. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2020.

[4] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel A J R Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P Butler, Carlos Caldas, Helen R Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Marcin Imielinsk, Natalie Jäger, David T W Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C Munshi, Hiromi Nakamura, Paul A Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V Pearson, Xose S Puente, Keiran Raine, Manasa Ramakrishna, Andrea L Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N Schumacher, Paul N Span, Jon W Teague, Yasushi Totoki, Andrew N J Tutt, Rafael Valdés-Mas, Marit M van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R Yates, Australian Pancreatic Cancer Genome Initiative, I.C.G.C. Breast Cancer Consortium, I.C.G.C. M M M L.-Seq Consortium, I C G C PedBrain, Jessica Zucman-Rossi, P Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M Pfister, Peter J Campbell, and Michael R Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, aug 2013.

[5] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael A Irizarry. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369, may 2014.

[6] Nicolai J Birkbak, Zhigang C Wang, Ji-Young Kim, Aron C Eklund, Qiyuan Li, Ruiyang Tian, Christian Bowman-Colin, Yang Li, April Greene-Colozzi, J Dirk Iglehart, Nadine Tung, Paula D Ryan, Judy E Garber, Daniel P Silver, Zoltan Szallasi, and Andrea L Richardson. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov.*, 2(4):366–375, apr 2012.

[7] Russell Bonneville, Melanie A Krook, Esko A Kautto, Jharna Miya, Michele R Wing, Hui-Zi Chen, Julie W Reeser, Lianbo Yu, and Sameek Roychowdhury. Landscape of microsatellite instability across 39 cancer types. *JCO Precis. Oncol.*, 1:1–15, 2017.

[8] Arnoud Boot, Mi Ni Huang, Alvin W T Ng, Szu-Chi Ho, Jing Quan Lim, Yoshiiku Kawakami, Kazuaki Chayama, Bin Tean Teh, Hidewaki Nakagawa, and Steven G Rozen. In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Res.*, 28(5):654–665, 2018.

[9] Broad DepMap. DepMap Achilles 18Q3 public, 2018.

[10] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, 31(3):213–219, mar 2013.

[11] Maura Costello, Trevor J Pugh, Timothy J Fennell, Chip Stewart, Lee Lichtenstein, James C Meldrim, Jennifer L Fostel, Dennis C Friedrich, Danielle Perrin, Danielle Dionne, Sharon Kim, Stacey B Gabriel, Eric S

Lander, Sheila Fisher, and Gad Getz. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.*, 41(6):e67, apr 2013.

[12] Helen Davies, Dominik Glodzik, Sandro Morganella, Lucy R Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, Sancha Martin, Sandrine Boyault, Anieta M Sieuwerts, and Others. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.*, 23(4):517, 2017.

[13] Brennan Decker, Danielle M Karyadi, Brian W Davis, Eric Karlins, Lori S Tillmans, Janet L Stanford, Stephen N Thibodeau, and Elaine A Ostrander. Biallelic BRCA2 mutations shape the somatic mutational landscape of aggressive prostate tumors. *Am. J. Hum. Genet.*, 98(5):818–829, 2016.

[14] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43(5):491–498, may 2011.

[15] Miklos Diossy, Lilla Reiniger, Zsofia Sztupinszki, Marcin Krzystanek, Kirsten M Timms, Chris Neff, Cara Solimeno, Dmitry Pruss, Aron C Eklund, Erika Toth, and Others. Breast cancer brain metastases show increased levels of genomic aberration-based homologous recombination deficiency scores relative to their corresponding primary tumors. *Ann. Oncol.*, 29(9):1948–1954, 2018.

[16] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–10, jan 2002.

[17] F Favero, T Joshi, A M Marquard, N J Birkbak, M Krzystanek, Q Li, Z Szallasi, and A C Eklund. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, 26(1):64–70, jan 2015.

[18] Neal D Freedman, Debra T Silverman, Albert R Hollenbeck, Arthur Schatzkin, and Christian C Abnet. Association between smoking and risk of bladder cancer among men and women. *JAMA*, 306(7):737–745, aug 2011.

[19] Mahmoud Ghandi, Franklin W Huang, Judit Jané-Valbuena, Gregory V Kryukov, Christopher C Lo, E Robert McDonald, Jordi Barretina, Ellen T Gelfand, Craig M Bielski, Haoxin Li, Kevin Hu, Alexander Y Andreev-Drakhlin, Jaegil Kim, Julian M Hess, Brian J Haas, François Aguet, Barbara A Weir, Michael V Rothberg, Brenton R Paolella, Michael S Lawrence, Rehan Akbani, Yiling Lu, Hong L Tiv, Prafulla C Gokhale, Antoine de Weck, Ali Amin Mansour, Coyin Oh, Juliann Shih, Kevin Hadi, Yanay Rosen, Jonathan Bistline, Kavitha Venkatesan, Anupama Reddy, Dmitriy Sonkin, Manway Liu, Joseph Lehar, Joshua M Korn, Dale A Porter, Michael D Jones, Javad Golji, Giordano Caponigro, Jordan E Taylor, Caitlin M Dunning, Amanda L Creech, Allison C Warren, James M McFarland, Mahdi Zamanighomi, Audrey Kauffmann, Nicolas Stransky, Marcin Imielinski, Yosef E Maruvka, Andrew D Cherniack, Aviad Tsherniak, Francisca Vazquez, Jacob D Jaffe, Andrew A Lane, David M Weinstock, Cory M Johannessen, Michael P Morrissey, Frank Stegmeier, Robert Schlegel, William C Hahn, Gad Getz, Gordon B Mills, Jesse S Boehm, Todd R Golub, Levi A Garraway, and William R Sellers. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, 569(7757):503–508, may 2019.

[20] Mary J Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N Brooks, Jingchun Zhu, and David Haussler. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.*, 38(6):675–678, 2020.

[21] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, Thomas Cokelaer, Patricia Greninger, Ewald van Dyk, Han Chang, Heshani de Silva, Holger Heyn, Xianming Deng, Regina K Egan, Qingsong Liu, Tatiana Mironenko, Xeni Mitropoulos, Laura Richardson, Jinhua Wang, Tinghu Zhang, Sebastian Moran, Sergi Sayols, Maryam Soleimani, David Tamborero, Nuria Lopez-Bigas, Petra Ross-Macdonald,

Manel Esteller, Nathanael S Gray, Daniel A Haber, Michael R Stratton, Cyril H Benes, Lodewyk F A Wessels, Julio Saez-Rodriguez, Ultan McDermott, and Mathew J Garnett. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 166(3):740–754, jul 2016.

[22] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, aug 2009.

[23] Quan Li and Kai Wang. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am. J. Hum. Genet.*, 100(2):267–280, 2017.

[24] Robin M Meyers, Jordan G Bryan, James M McFarland, Barbara A Weir, Ann E Sizemore, Han Xu, Neekesh V Dharia, Phillip G Montgomery, Glenn S Cowley, Sasha Pantel, Amy Goodale, Yenarae Lee, Levi D Ali, Guozhi Jiang, Rakela Lubonja, William F Harrington, Matthew Strickland, Ting Wu, Derek C Hawes, Victor A Zhivich, Meghan R Wyatt, Zohra Kalani, Jaime J Chang, Michael Okamoto, Kimberly Stegmaier, Todd R Golub, Jesse S Boehm, Francisca Vazquez, David E Root, William C Hahn, and Aviad Tsherniak. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.*, 49(12):1779–1784, dec 2017.

[25] Jolein Mijnes, Jürgen Veeck, Nadine T Gaisa, Eduard Burghardt, Tim C de Ruijter, Sonja Gostek, Edgar Dahl, David Pfister, Sebastian C Schmid, Ruth Knüchel, and Michael Rose. Promoter methylation of DNA damage repair (DDR) genes in human tumor entities: RBBP8/CtIP is almost exclusively methylated in bladder cancer. *Clin. Epigenetics*, 10:15, 2018.

[26] Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B Alexandrov, Sancha Martin, David C Wedge, Peter Van Loo, Young Seok Ju, Marcel Smid, Arie B Brinkman, Sandro Morganella, Miriam R Aure, Ole Christian Lingjærde, Anita Langerød, Markus Ringnér, Sung-Min Ahn, Sandrine Boyault, Jane E Brock, Annegien Broeks, Adam Butler, Christine Desmedt, Luc Dirix, Serge Dronov, Aquila Fatima, John A Foekens, Moritz Gerstung, Gerrit K J Hooijer, Se Jin Jang, David R Jones, Hyung-Yong Kim, Tari A King, Savitri Krishnamurthy, Hee Jin Lee, Jeong-Yeon Lee, Yilong Li, Stuart McLaren, Andrew Menzies, Ville Mustonen, Sarah O'Meara, Iris Pauporté, Xavier Pivot, Colin A Purdie, Keiran Raine, Kamna Ramakrishnan, F Germán Rodríguez-González, Gilles Romieu, Anieta M Sieuwerts, Peter T Simpson, Rebecca Shepherd, Lucy Stebbings, Olafur A Stefansson, Jon Teague, Stefania Tommasi, Isabelle Treilleux, Gert G den Eynden, Peter Vermeulen, Anne Vincent-Salomon, Lucy Yates, Carlos Caldas, Laura van't Veer, Andrew Tutt, Stian Knappskog, Benita Kiat Tee Tan, Jos Jonkers, Åke Borg, Naoto T Ueno, Christos Sotiriou, Alain Viari, P Andrew Futreal, Peter J Campbell, Paul N Span, Steven Van Laere, Sunil R Lakhani, Jorunn E Eyfjord, Alastair M Thompson, Ewan Birney, Hendrik G Stunnenberg, Marc J van de Vijver, John W M Martens, Anne-Lise Børresen-Dale, Andrea L Richardson, Gu Kong, Gilles Thomas, and Michael R Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, jun 2016.

[27] Tatiana Popova, Elodie Manié, Guillaume Rieunier, Virginie Caux-Moncoutier, Carole Tirapo, Thierry Dubois, Olivier Delattre, Brigitte Sigal-Zafrani, Marc Bollet, Michel Longy, Claude Houdayer, Xavier Sastre-Garau, Anne Vincent-Salomon, Dominique Stoppa-Lyonnet, and Marc-Henri Stern. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.*, 72(21):5454–5462, nov 2012.

[28] Sara Przetocka, Antonio Porro, Hella A Bolck, Christina Walker, Aleksandra Lezaja, Anika Trenner, Christine von Aesch, Sarah-Felicitas Himmels, Alan D D'Andrea, Raphael Ceccaldi, Matthias Altmeyer, and Alessandro A Sartori. CtIP-Mediated Fork Protection Synergizes with BRCA1 to Suppress Genomic Instability upon DNA Replication Stress. *Mol. Cell*, 72(3):568–582.e6, 2018.

[29] Rachel Rosenthal, Nicholas McGranahan, Javier Herrero, Barry S Taylor, and Charles Swanton. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.*, 17(1):1–11, 2016.

[30] Brinton Seashore-Ludlow, Matthew G Rees, Jaime H Cheah, Murat Cokol, Edmund V Price, Matthew E Coletti, Victor Jones, Nicole E Bodycombe, Christian K Soule, Joshua Gould, Benjamin Alexander, Ava Li, Philip Montgomery, Mathias J Wawer, Nurdan Kuru, Joanne D Kotz, C Suk-Yee Hon, Benito Munoz, Ted Liefeld, Vlado Dančík, Joshua A Bittker, Michelle Palmer, James E Bradner, Alykhan F Shamji, Paul A Clemons, and Stuart L Schreiber. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov.*, 5(11):1210–23, nov 2015.

[31] Ronglai Shen and Venkatraman E Seshan. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.*, 44(16):e131—-e131, 2016.

[32] Petr Smirnov, Zhaleh Safikhani, Nehme El-Hachem, Dong Wang, Adrian She, Catharina Olsen, Mark Freeman, Heather Selby, Deena M A Gendoo, Patrick Grossmann, Andrew H Beck, Hugo J W L Aerts, Mathieu Lupien, Anna Goldenberg, and Benjamin Haibe-Kains. PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics*, 32(8):1244–1246, apr 2016.

[33] Zsofia Sztupinszki, Miklos Diossy, Marcin Krzystanek, Lilla Reiniger, István Csabai, Francesco Favero, Nicolai J Birkbak, Aron C Eklund, Ali Syed, and Zoltan Szallasi. Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. *NPJ breast cancer*, 4(1):1–4, 2018.

[34] Melinda L Telli, Kirsten M Timms, Julia Reid, Bryan Hennessy, Gordon B Mills, Kristin C Jensen, Zoltan Szallasi, William T Barry, Eric P Winer, Nadine M Tung, Steven J Isakoff, Paula D Ryan, April Greene-Colozzi, Alexander Gutin, Zaina Sangale, Diana Iliev, Chris Neff, Victor Abkevich, Joshua T Jones, Jerry S Lanchbury, Anne-Renee Hartman, Judy E Garber, James M Ford, Daniel P Silver, and Andrea L Richardson. Homologous Recombination Deficiency (HRD) Score Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients with Triple-Negative Breast Cancer. *Clin. Cancer Res.*, 22(15):3764–73, 2016.

[35] Kirsten M Timms, Victor Abkevich, Elisha Hughes, Chris Neff, Julia Reid, Brian Morris, Saritha Kalva, Jennifer Potter, Thanh V Tran, Jian Chen, Diana Iliev, Zaina Sangale, Eliso Tikishvili, Michael Perry, Andrey Zharkikh, Alexander Gutin, and Jerry S Lanchbury. Association of BRCA1/2 defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes. *Breast Cancer Res.*, 16(6):475, dec 2014.

[36] J Zámborszky, B Szikriszt, J Z Gervai, O Pipek, Á Póti, M Krzystanek, D Ribli, J M Szalai-Gindl, I Csabai, Z Szallasi, C Swanton, A L Richardson, and D Szüts. Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene*, 36(35):5085–5086, 2017.

[37] Reihaneh Zarrizi, Martin R Higgs, Karolin Voßgröne, Maria Rossing, Birgitte Bertelsen, Muthiah Bose, Arne Nedergaard Kousholt, Heike Rösner, Bent Ejlertsen, Grant S Stewart, and Others. Germline RBBP8 variants associated with early-onset breast cancer compromise replication fork stability. *J. Clin. Invest.*, 130(8), 2020.

[38] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, 18(5):821–829, may 2008.