Structural variants at the *BRCA1/2* loci are a common source of homologous repair deficiency in high grade serous ovarian carcinoma

Supplementary Information

Scottish DNA sample preparation and quality control

Somatic DNA was extracted using the Qiagen DNeasy Blood and tissue kit (cat no 69504). The tissue was initially homogenised using a Qiagen Bioruptor, followed by the manufacturers recommended protocol (including RNase digestion step). Germline DNA was extracted from 1-3ml whole blood using the Qiagen FlexiGene kit (cat no 51206) following the manufacturers recommended protocol. The resulting DNA underwent quality control as follows: firstly, A260 and A280nm were measured on a Denovix DS-11 Fx to qualitatively illustrate A260/280nm and A260/230nm ratios as surrogate measures of DNA purity. A260/280 had to be 1.8 or greater and A260/230 had to be 2.0 or greater. Then, DNA was quantified using LifeTechnologies Qubit dsDNA BR kit (cat no Q32850) and we required a minimum of 50ul at 25ng/ul for WGS. Thirdly, DNA was diluted to 25ng/ul and a representative sample was loaded onto a 0.8% TAE gel, ran at 100v for 60mins and then imaged using a BioRad ChemiDoc imaging system to visualise the DNA quality.

Primary processing pipeline resources and versions

bamtools	2.4.0	https://doi.org/10.1093/bioinformatics/btr174
bcbio-nextgen	1.0.7	https://github.com/bcbio/bcbio-nextgen
bcftools	1.6	https://github.com/samtools/bcftools
bedtools	2.27.1	https://doi.org/10.1093/bioinformatics/btq033
biobambam2	2.0.87	https://gitlab.com/german.tischler/biobambam2
bwa	0.7.17	https://www.ncbi.nlm.nih.gov/pubmed/20080505
CLImaT		https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4155249/
cnvkit	0.9.2a0	https://www.ncbi.nlm.nih.gov/pubmed/27100738
facets		https://pubmed.ncbi.nlm.nih.gov/27270079/
fastqc	0.11.8	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
FeatureCounts	1.6.4	https://academic.oup.com/bioinformatics/article/30/7/923/232889
gatk4	4.0.0.0	https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1110s43 https://www.ncbi.nlm.nih.gov/pubmed?term=20644199
grabix	0.1.8	https://github.com/arq5x/grabix

manta	1.2.1	https://www.ncbi.nlm.nih.gov/pubmed/26647377
multiQC	1.7	https://academic.oup.com/bioinformatics/article/32/19/3047/2196507
mutect2	1.1.5	http://www.nature.com/nbt/journal/v31/n3/full/nbt.2514.html
picard	2.17.2	https://broadinstitute.github.io/picard/
Qualimap	2.2.2-dev	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4708105/
Qsignature	0.1	https://sourceforge.net/p/adamajava/wiki/qSignature/
Salmon quant	0.12.0	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5600148/
sambamba	0.6.8	A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins. Sambamba: fast processing of NGS alignment formats. Bioinformatics, 2015
samblaster	0.1.24	https://github.com/GregoryFaust/samblaster
samtools	1.6	https://github.com/samtools/samtools
Strelka2		https://www.nature.com/articles/s41592-018-0051-x
tximport	1.12.1	https://f1000research.com/articles/4-1521
vardict	2017.11.23	Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, and Dry JR. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016, pii: gkw227.
vardict-java	1.5.1	Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, and Dry JR. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016, pii: gkw227.
variant-effect-predictor	91_GRCh38	https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4
vcflib	1.0.0_rc1	https://github.com/vcflib/vcflib
VerifyBamId	1.1.3	https://www.ncbi.nlm.nih.gov/pubmed/23103226
vt	2015.11.10	http://bioinformatics.oxfordjournals.org/content/31/13/2202

Formal definition of BRCA1/2 mutational categories

General *BRCA1/2* mutational category	***BRCA1/2* mutational category**	**Formal definition**
Deletion ●	Deletion of both *BRCA1* and *BRCA2* / Deletion at both genes	• Deletion of at least one exon at BOTH *BRCA1* and *BRCA2* • Absence of other SVs at *BRCA1/2* • Absence of GSM or SSM at either gene (Figure 3c,d)
●	Deletion of one of *BRCA1/2*	• Deletion of at least one exon at only one of *BRCA1* or *BRCA2* • Absence of other SVs at gene with deletion
● *(Figures3c,d)*	Single deletion at *BRCA1*	• Deletion of at least one exon at *BRCA1* only • Absence of other SVs at *BRCA1* • Absence of GSM or SSM at either gene (Figure 3c,d)
● *(Figures3c,d)*	Single deletion at *BRCA2*	• Deletion of at least one exon at *BRCA2* only • Absence of other SVs at *BRCA2* • Absence of GSM or SSM at either gene (Figure 3c,d)
●	Intronic deletion at *BRCA1/2*	• Deletion not involving any exons at *BRCA1/2* • Absence of other SVs at either gene
Duplication ●	Duplication spanning *BRCA1/2*	• Duplication spanning at least one of *BRCA1* or *BRCA2* • Absence of other SVs at the duplicated gene
●	*BRCA2* duplication	• Duplication spanning *BRCA2* • Absence of other SVs at *BRCA2* • Absence of GSM/SSM/deletion/inversion at either gene (Figure 3c,d)
●	Intragenic exonic duplication at *BRCA1/2*	• Duplication spanning at least one exon but not the entirety of one of *BRCA1* or *BRCA2* • Absence of other SVs at either gene
Inversion ●	Inversion spanning *BRCA1/2*	• Inversion spanning at least one of *BRCA1* or *BRCA2* • Absence of other SVs at the inverted gene
●	*BRCA1* inversion	• Inversion spanning *BRCA1* • Absence of other SVs at *BRCA1* • Absence of GSM/SSM/deletion at either gene (Figure 3c,d)
Complex SVs ●	Complex combination of SVs including deletion of *BRCA1/2*	• The presence of more than one SV including at least one deletion at at least one of *BRCA1/2*
●	Complex combination of SVs at *BRCA1/2* without deletion	• The presence of more than one SV excluding deletion at at least one of *BRCA1/2*

Further details regarding implementation of HRDetect

The HRDetect[1] algorithm is a logistic regression model with the probability of HR deficiency defined as '*BRCA*-ness' as the outcome. The variables that make up the linear predictor represent genomic signatures that have been shown to correlate well with *BRCA1/2* mutation status. They include: the proportion of indels with microhomology at the breakpoints; the contribution of COSMIC SNV signatures 3 and 8 to the mutational profile of the tumour; the contribution of rearrangement signatures 3 and 5 to structural variation in the tumour; and the value of an earlier predictor of HR deficiency, the HRDIndex[2], which combines levels of genome-wide medium length runs of loss of heterozygosity (LOH), telomere allelic imbalance (TAI) and large state transitions (LST). We based our implementation on a Snakemake pipeline made publicly available by Zhao et al[3], with some modifications to ensure accurate recapitulation of the original method. As some of the AOCS cohort included here were also used in the validation of HRDetect in the original publication we were able to compare our implementation for the same patients with that of the authors.

Zhao's pipeline makes use of the R package HRDtools in order to determine the value of the HRIndex. We used the same method determined by Zhao et al with the exception that we took the mean of the three inputs (LOH, TAI and LST) instead of the sum to reflect the original HRDetect approach. In addition, we redefined microhomology at indel breakpoints as an overlap between the deletion and the flanking region that is less than the full length of the deletion. In order to determine the contribution of each of the signatures to the mutational profile of each tumour we implemented three different methods: deconstructSigs[4], SignIT[5] and SigProfilerSingleSample[6]. Ultimately, we chose to use deconstructSigs as its estimates were the most strongly correlated with the results from the original HRDetect paper. After determining the value of each of the components of the linear predictor for each sample, each of these input variables were standardized using the corresponding mean and standard deviation for the variable in question in the dataset that was used to determine the weights in the original model published by Davies et al[1].

Scottish RNA sample preparation, quality control and sequencing

Somatic RNA was extracted from the resulting RNA sample using the Qiagen Qiasymphony RNA protcol (cat no 931636). The tissue was initially homogenised using a Qiagen Bioruptor, followed by the manufacturers recommended protocol (including DNase digestion). The resulting RNA the underwent quality control as follows: firstly, A260 and A280nm were measured on a Denovix DS-11 Fx to qualitatively illustrate A260/280nm and A260/230nm ratios as measures of RNA purity. A260/280 had to be 2.0 and A260/230 had to be 2.0-2.2. Then RNA was quantified using LifeTechnologies Qubit RNA BR kit (cat no Q10210). RNAseq was carried out by the Edinburgh Clinical Research Facility on an Illumina NExtSeq500 as detailed below.

Total RNA samples were assessed on the Agilent Bioanalyser (Agilent Technologies, #G2939AA) with the RNA 6000 Nano Kit (#5067-1512) for quality and integrity of total RNA, and then quantified using the Qubit 2.0 Fluorometer (Thermo Fisher Scientific Inc, #Q32866) and the Qubit RNA HS assay kit (#Q32855). Libraries were prepared from total-RNA sample using the NEBNext Ultra 2 Directional RNA library prep kit for Illumina (#E7760S) with the NEBNext rRNA Depletion kit (#E6310) according to the provided protocol. 400ng of total-RNA was then added to the ribosomal RNA (rRNA) depletion reaction using the NEBNext rRNA depletion kit (Human/mouse/rat) (#E6310). This step uses specific probes that bind to the rRNA in order to cleave it. rRNA-depleted RNA was then DNase treated and purified using Agencourt RNAClean XP beads (Beckman Coulter Inc, #66514). RNA was then fragmented using random primers before undergoing first strand and second strand synthesis to create cDNA. cDNA was end repaired before ligation of sequencing adapters, and libraries were enriched by PCR using the NEBNext Multiplex oligos for Illumina set 1 and 2 (#E7500). Final libraries had an average peak size of 271bp. Libraries were quantified by fluorometry using the Qubit dsDNA HS assay and assessed for quality and fragment size using the Agilent Bioanalyser with the DNA HS Kit (#5067-4626). Sequencing was performed using the NextSeq 500/550 High-Output v2 (150 cycle) Kit (# FC- 404-2002) on the NextSeq 550 platform (Illumina Inc, #SY-415-1002). Libraries were combined in an equimolar pool based on the library quantification results and run across 5 High-Output Flow Cell v2.5.

Identifying differentially expressed genes in the presence of HRD

Transcriptomic signatures have previously been generated[7–10] to identify HRD tumours; however, most have used suboptimal proxies such as mutation rate to predict HRD or have been based upon expression in HR deficient cell lines or samples that are not from HGSOC patients[7–10]. Exploiting our novel combined cohort with matched genomic and transcriptomic data, we identified a list of differentially expressed (DE) genes between HR deficient and HR proficient HGSOC tumours, encompassing 306 protein coding genes (Supplementary Table 4). For the samples with RNAseq information, we defined a conservative HR deficient group which included the samples with pathogenic short variants at *BRCA1/2* either in the germline or in the tumour (N=50). The contrasting HR proficient group of tumours, consisted of samples without damaging *BRCA1/2* short variants or *BRCA1* promoter hypermethylation or damaging short variants at HR genes as defined by KEGG and a quiet mutational profile defined by absence of the HRD related rearrangement signatures (N=47). This is consistent with the definition of HR proficiency used to train HRDetect. We used DESeq2 to compare the expression of all protein coding genes between the two groups and identified those genes that were differentially expressed. Cohort and tumour cellularity were included as covariates in the model. We used a log fold change threshold of 1 and a Benjamini-Hochberg adjusted p-value threshold of 0.05 to indicate significant differential expression. Functional annotation of the differentially expressed genes was carried out by comparing the differentially expressed genes with a background list of all protein-coding genes and testing for enrichment of the differentially expressed genes in curated gene lists from GO: BP, CC and MF and KEGG. This was done using clusterProfiler[11] with p-value and q-value thresholds of 0.05.

Gene expression signatures for HRD

We defined a gene expression signature for HR deficiency by running principal component analysis on the variance stabilising transformed counts of the differentially expressed genes using all of the samples. The first principal component was taken as the gene expression signature for HRD with HR deficient samples having significantly lower values of the signature than HR proficient samples. We tested whether HR deficient and proficient samples had significantly different levels of the signature using a Wilcoxon Rank Sum test (Mann-Whitney U test)(Supplementary Figure 6a)-b)).

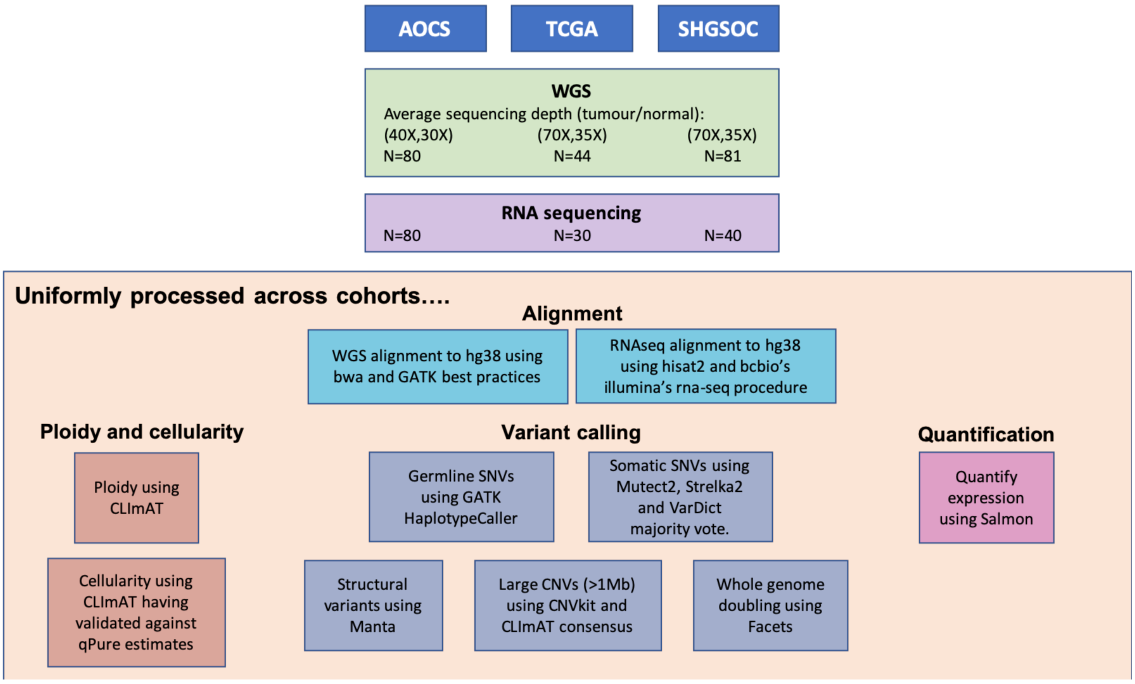
The ability of a gene expression signature for HRD to predict HRD was assessed by identifying differentially expressed genes between only 80% of true HR deficient and 80% of true HR proficient samples and examining whether the HR deficient and proficient samples in the test set lay at significantly different points along the main axis of variation (first principal component) in the expression of these genes within the test set. The difference in the levels of the signature for HR deficient and proficient samples within the test set was tested using a Wilcoxon Rank Sum test. We found that the expression of DE genes identified between HR deficient and HR proficient samples in a subset of the cohort failed to accurately discriminate HR deficient from HR proficient samples in the unexamined remainder of the cohort (Wilcox p-value=0.92)(Supplementary Figure 6c), 6d)). This is consistent with previous reports[12] and suggests that although the transcriptome is perturbed in the presence of HRD, such perturbations are not consistent, and consequently these expression changes are poor predictors of HRD.

References

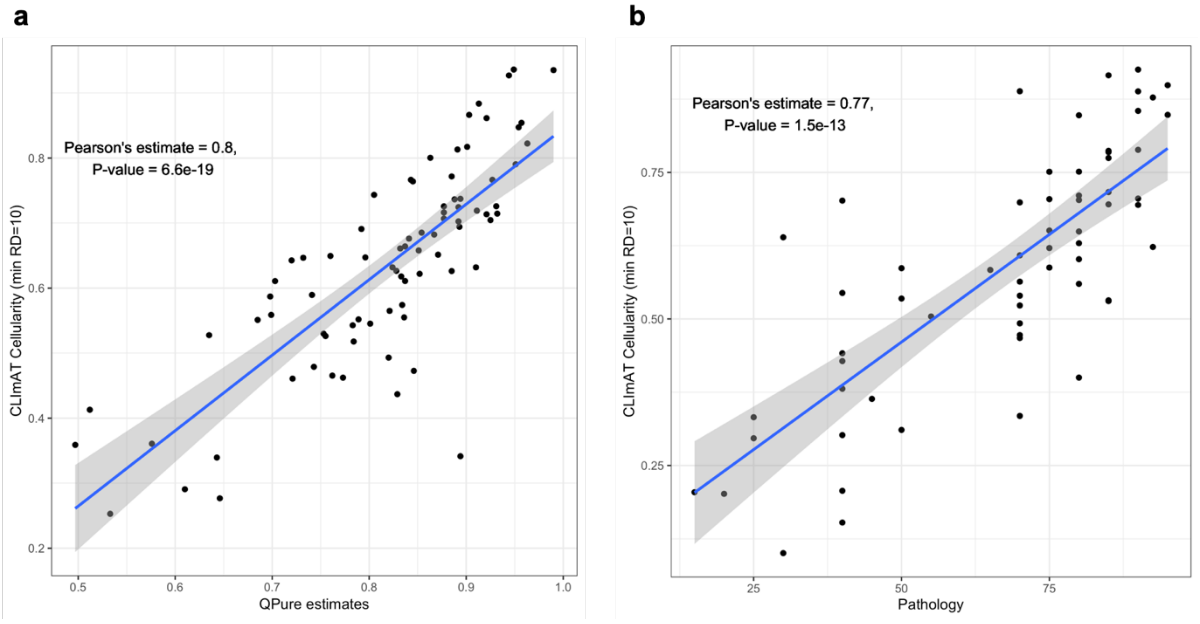
1. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* (2017). doi:10.1038/nm.4292
2. Timms, K. M. *et al.* Association of BRCA1/2 defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes. *Breast Cancer Res.* **16**, 1–9 (2014).
3. Zhao, E. Y. *et al.* Homologous Recombination Deficiency and Platinum-Based Therapy Outcomes in Advanced Breast Cancer. *Clin. Cancer Res.* **23**, 7521–7530 (2017).
4. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
5. eyzhao/SignIT: Mutation Signatures in Individual Tumours Deciphered by MCMC. Available at: https://github.com/eyzhao/SignIT. (Accessed: 21st April 2020)
6. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
7. Konstantinopoulos, P. A. *et al.* Gene Expression Profile of BRCAness That Correlates With Responsiveness to Chemotherapy and With Outcome in Patients With Epithelial Ovarian Cancer. *J Clin Oncol* **28**, 3555–3561 (2010).
8. Lu, J., Wu, D., Li, C., Zhou, M. & Hao, D. Correlation between gene expression and mutator phenotype predicts homologous recombination deficiency and outcome in ovarian cancer. *J. Mol. Med.* **92**, 1159–1168 (2014).
9. Peng, G. *et al.* Genome-wide transcriptome profiling of homologous recombination DNA repair. *Nat. Commun.* **5**, 3361 (2014).
10. McGrail, D. J. *et al.* Improved prediction of PARP inhibitor response and identification of synergizing agents through use of a novel gene expression signature generation algorithm. *npj Syst. Biol. Appl.* **3**, 8 (2017).
11. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omi. A J. Integr. Biol.* **16**, 284–287 (2012).

12. Staaf, J. *et al.* Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.* doi:10.1038/s41591-019-0582-4

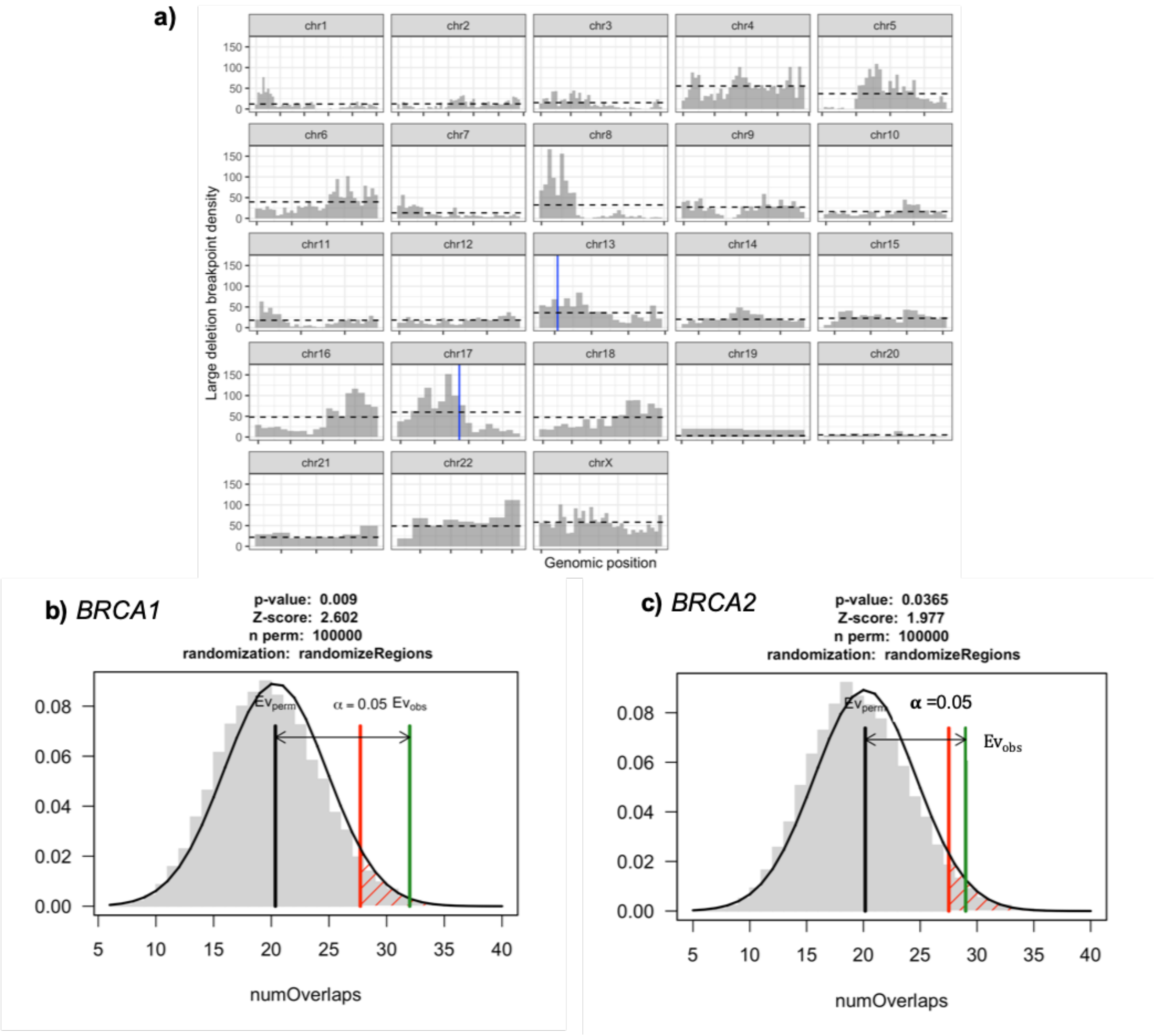
Supplementary Figures


AOCS
TCGA
SHGSOC
WGS
Average sequencing depth (tumour/normal):
(40X,30X) (70X,35X) (70X,35X)
N=80 N=44 N=81
RNA sequencing
N=80 N=30 N=40
Uniformly processed across cohorts....
Alignment
WGS alignment to hg38 using bwa and GATK best practices
RNAseq alignment to hg38 using hisat2 and bcbio's illumina's rna-seq procedure
Ploidy and cellularity
Ploidy using CLImAT
Cellularity using CLImAT having validated against qPure estimates
Variant calling
Germline SNVs using GATK HaplotypeCaller
Somatic SNVs using Mutect2, Strelka2 and VarDict majority vote.
Structural variants using Manta
Large CNVs (>1Mb) using CNVkit and CLImAT consensus
Whole genome doubling using Facets
Quantification
Quantify expression using Salmon

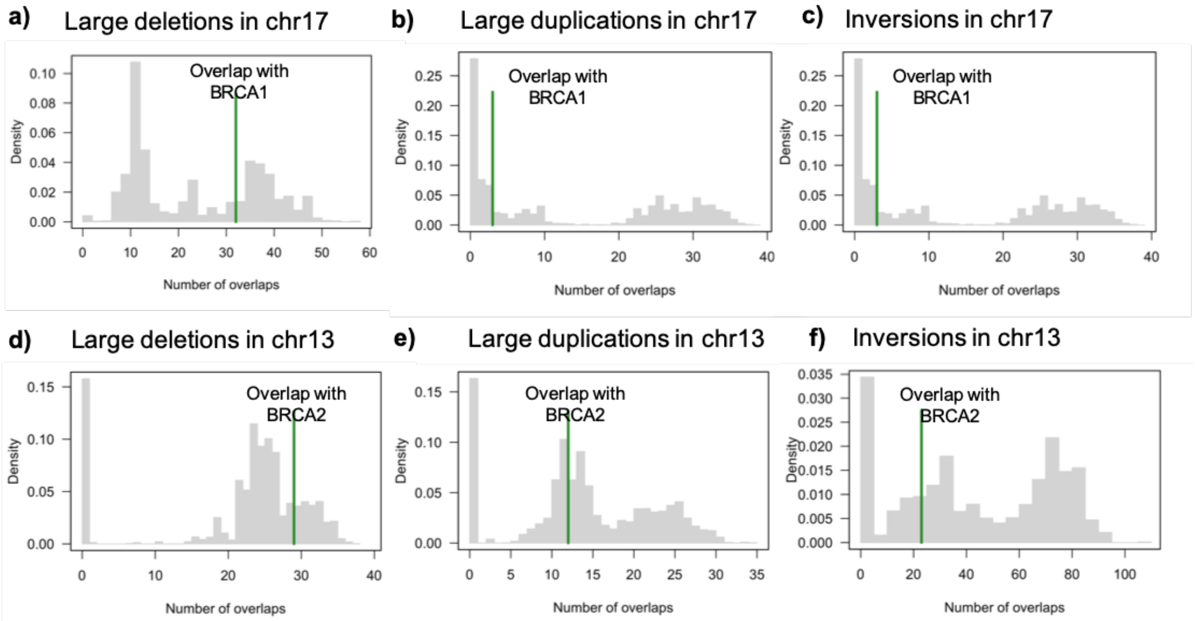
Supplementary Figure 1: Uniform primary processing of three large HGSOC cohorts. WGS and RNA-seq fastqs were downloaded for AOCS and TCGA and the SHGSOC cohort was sequenced for the first time. Sequencing reads were aligned uniformly for all cohorts to hg38 and variant detection was carried out to detect a range of types of variant using existing published tools. Ploidy and cellularity were estimated using the allele-specific copy number caller CLImAT and gene-level expression counts were quantified using Salmon.



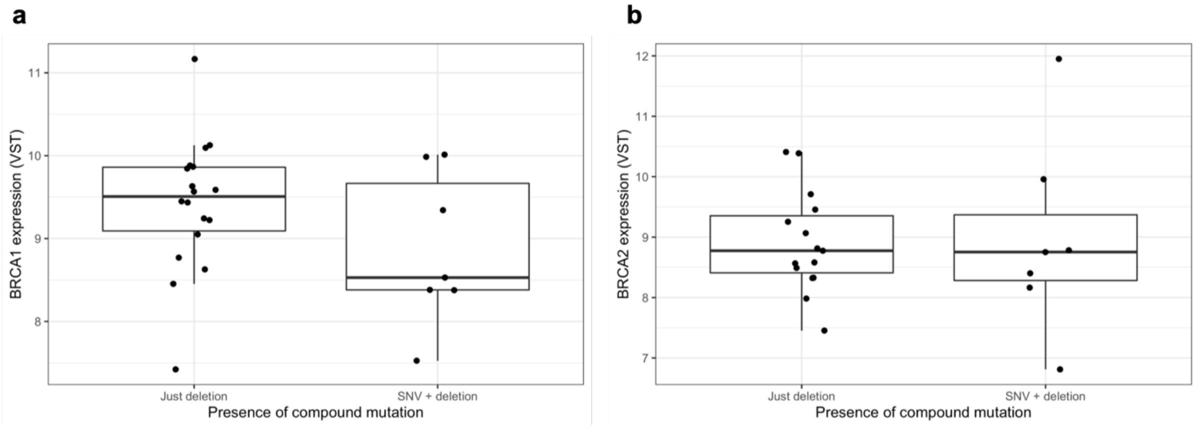
Supplementary Figure 2 - Comparison of methods to estimate the tumour cellularity in two cohorts. a) Estimates of tumour cellularity from allele-specific copy number tool CLImAT in comparison to estimates using qPure for the AOCS cohort. b) Estimates of tumour cellularity from allele-specific copy number tool CLImAT in comparison to scores from manual examination of the histopathology for the SHGSOC cohort.



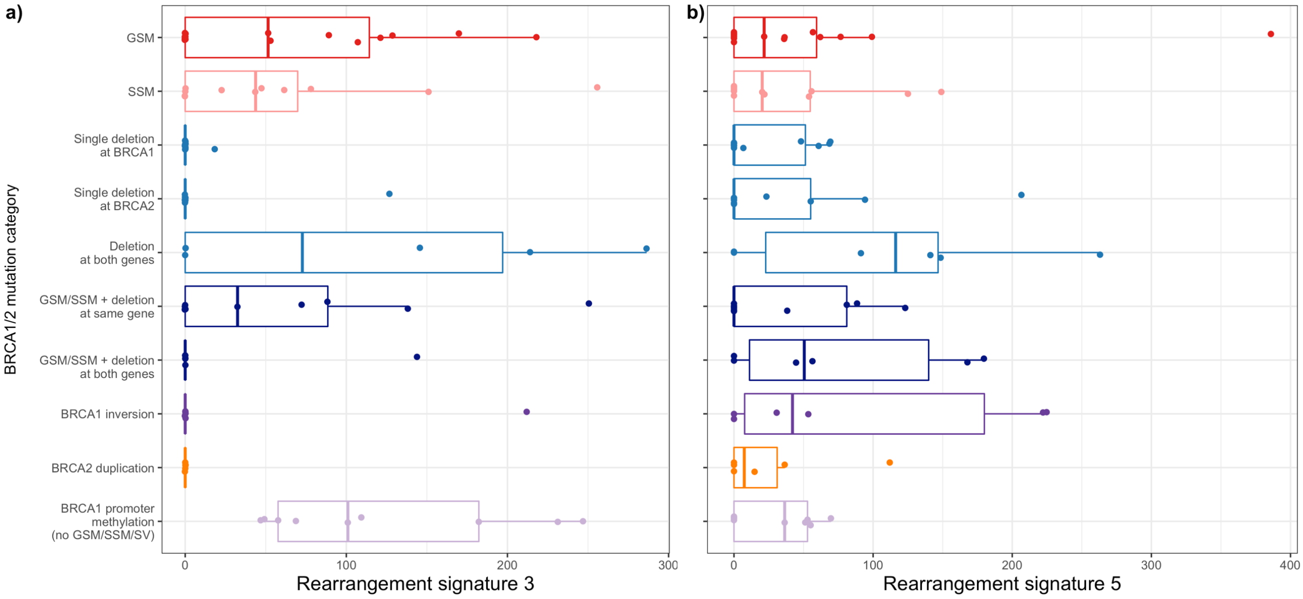
Supplementary Figure 3 – Enrichment of large deletions (>1Mb) at *BRCA1/2* in comparison to the rest of the high grade serous ovarian cancer genome. a) The distribution of large deletion breakpoints throughout the genome by chromosome. The observed number of breakpoints from large deletions in 5Mb bins is shown in grey. The mean number of breakpoints per bin per chromosome is shown by a black dashed line. The locations of *BRCA1* and *BRCA2* are shown in blue on chromosomes 13 and 17. b) The null distribution (in grey) of the number of overlaps between observed large deletions throughout the HGSOC genomes and 100,000 random regions the same size as *BRCA1* sampled from throughout the genome which is masked to exclude unmappable, repetitive regions. In green the observed number of overlaps between the observed large deletions and *BRCA1*. c) The null distribution (in grey) of the number of overlaps between observed large deletions throughout the HGSOC genomes and 100,000 random regions the same size as *BRCA2* sampled from throughout the masked genome. In green, the observed number of overlaps between the observed large deletions and *BRCA2*.



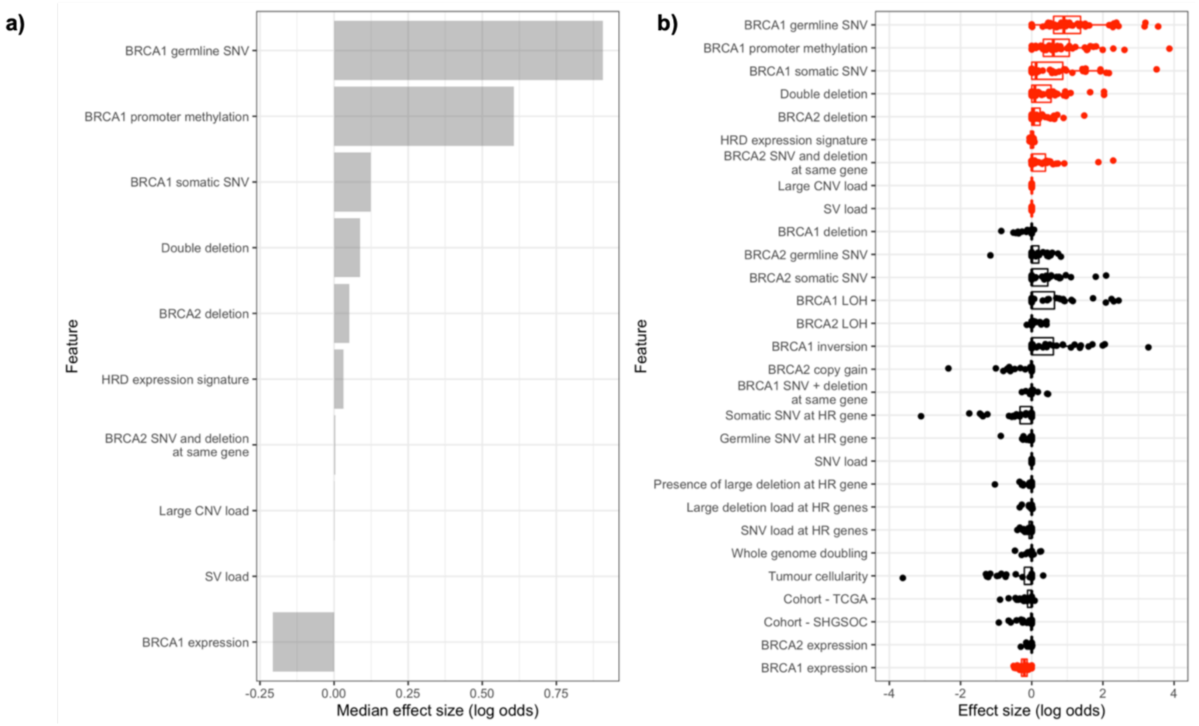
Supplementary Figure 4 – Enrichment of deletions, duplications and inversions at *BRCA1/2* within chromosomes 13 and 17. In all panels, the grey histogram represents the circularly permuted null distribution of overlaps between the structural variants and regions that are the same size as *BRCA1/2* throughout their respective chromosomes. The green line represents the observed number of overlaps between the structural variant type in question and *BRCA1/2*. In all cases the events occurring at *BRCA1* or *BRCA2* are well within the range expected given the permuted null distribution and are therefore not significantly enriched**.** a), b) and c) consider the enrichment of large deletions, large duplications and inversions at *BRCA1* within chromosome 17 and d), e) and f) consider the enrichment of large deletions, large duplications and inversions at *BRCA2* within chromosome 13.



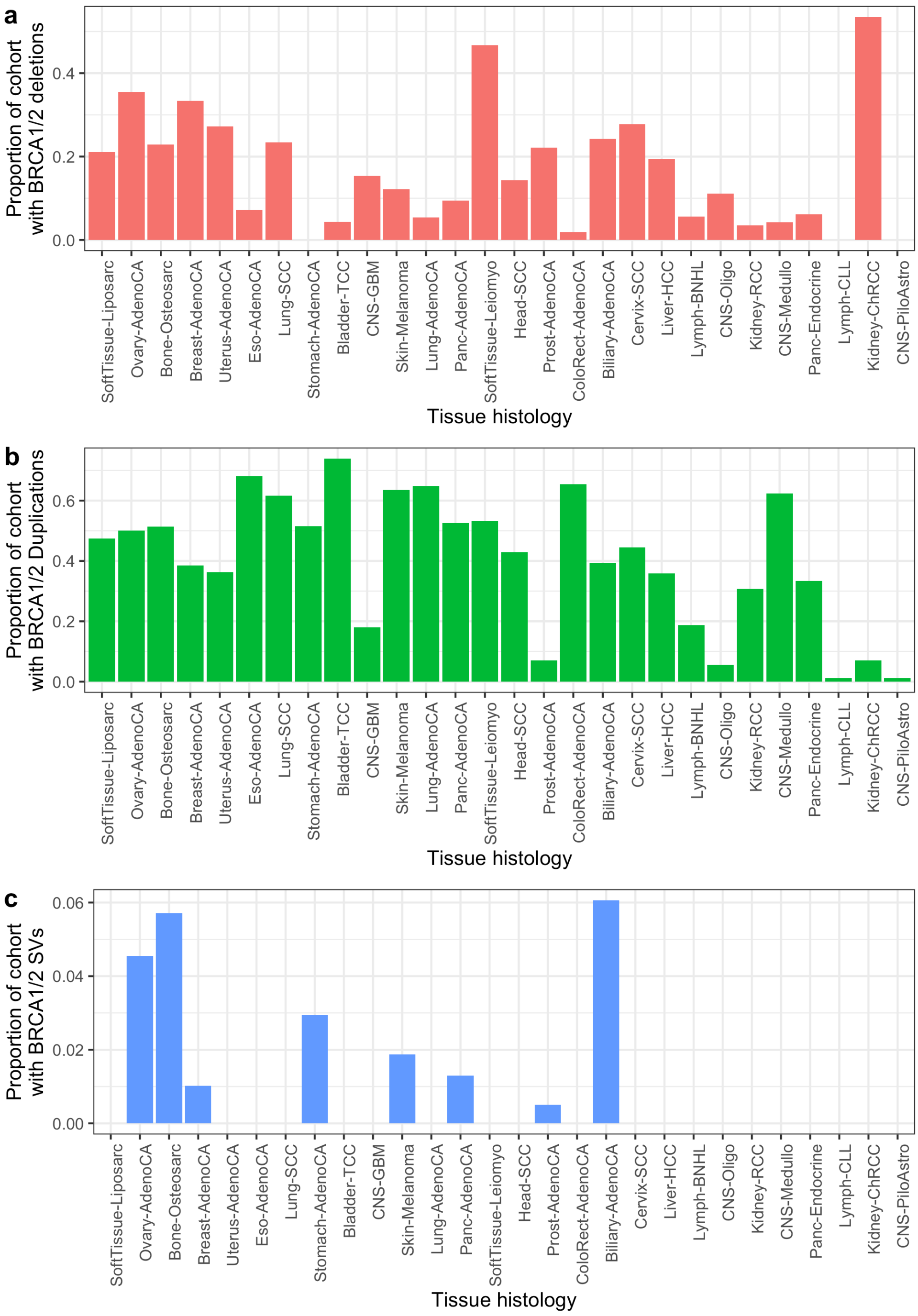
Supplementary Figure 5 - *BRCA1/2* expression in samples with deletions, with and without SNVs in the same *BRCA1/2* gene. a) Boxplot with points overlaid showing that *BRCA1* expression (variance stabilising transformed) is higher in samples with only deletions than in samples with an SNV and a deletion at *BRCA1* (DESeq2 fold change for just deletions vs SNV + deletion = 1.6, p-value=0.02). b) Boxplot with points overlaid showing no evidence of a significant difference in *BRCA2* expression between samples with only deletions and samples with an SNV and a deletion at the same gene (DESeq2 fold change for just deletions vs SNV + deletion= 1.02, p-value=0.95).



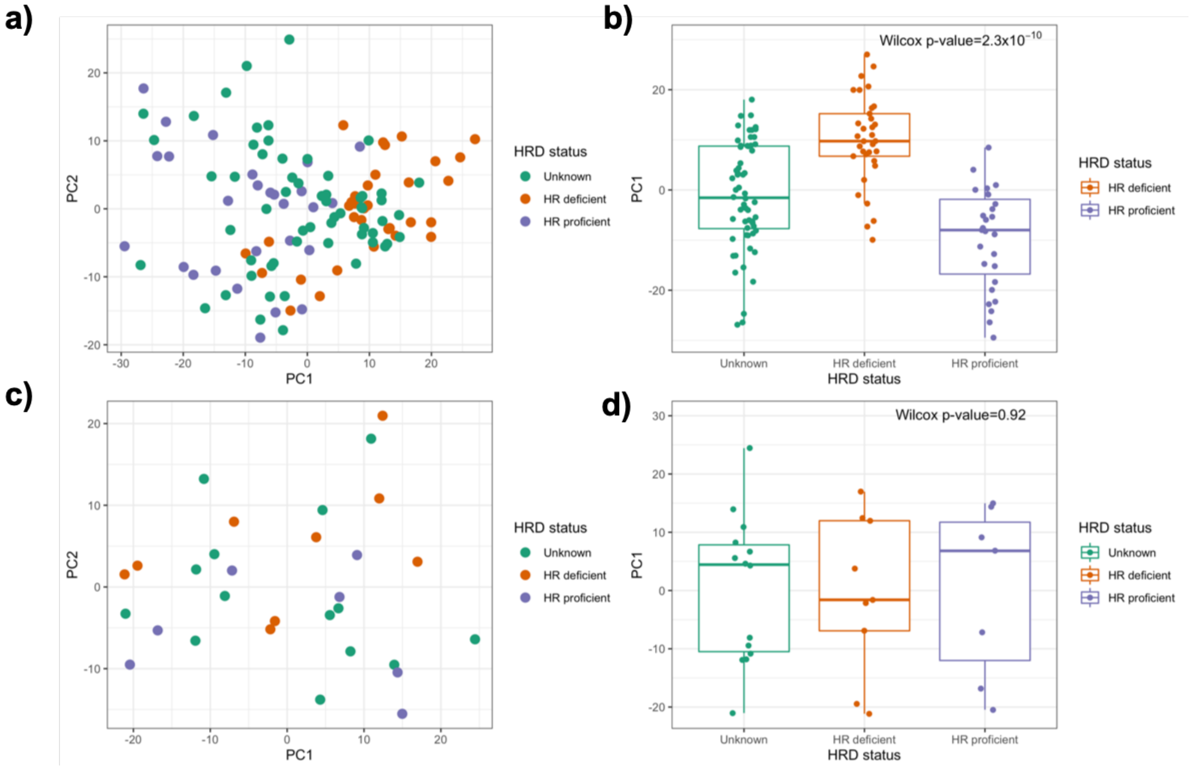
Supplementary Figure 6 – *BRCA1/2* mutation classes and rearrangement signatures in three HGSOC cohorts. a) Boxplots of the level of rearrangement signature 3 by *BRCA1/2* mutation category. Rearrangement signature 3 is characterised by tandem duplications and has been associated with *BRCA1* deficiency specifically. Rearrangement signature 3 is a term in the HRDetect predictive model. b) Boxplots of the level of rearrangement signature 5 by *BRCA1/2* mutation category. Rearrangement signature 5 is characterised by deletions <100kb in length and has been associated with both *BRCA1* and *BRCA2* deficiency. It is also included as a term in the HRDetect model.



Supplementary Figure 7 - Integrative modelling of repair deficiency in HGSOC in full dataset. a) Median effect sizes of features selected to predict HRD, using elastic net regularised regression on 50 training/test set splits. Binary mutational status variables (e.g. presence/absence of *BRCA1* somatic SNV) were included as factors and continuous variables were standardised to allow comparisons between variables. b) Distributions of effect size for each variable on HRD (log odds) in each training/test set split. Variables in red are selected for inclusion by the model in more than half of the training sets. It should be noted that, due to the lower number of samples with expression information and the increased number of features this model is likely to be underpowered to accurately identify significant features.



Supplementary Figure 8 – Deletions, duplications and other SVs at *BRCA1/2* in other cancer types in PCAWG. The proportion of samples of various histologies in PCAWG with a deletion, duplication or other SVs overlapping at least one exon at at least one of *BRCA1/2*. Histologies are ordered along the x-axis by overall SV burden as defined by PCAWG with the subtypes with the highest rates of structural variation on the left. Deletions and duplications considered separately here are those identified by the PCAWG consensus CNV pipeline. The other SVs, including translocations with a breakpoint within a *BRCA1/2* exon, inversions, and deletions and duplications picked up by paired and split read technologies, are determined by the PCAWG consensus SV pipeline. For consistency, we focus on those deletions called by the consensus CNV pipeline in our subsequent analyses.



Supplementary Figure 9 – Performance of expression signature for HRD at predicting HRD. a) The first two principal components of differentially expressed (DE) genes between HRD and HRP samples. DE genes identified in the training set and PCA fitted to the training set. b) The level of the first principal component in samples in the training set. The first principal component, is significantly different between HR deficient and HR proficient samples in the combined cohort (Wilcox p-values =2.3x10^{-10}) c) The first two principal components from PCA applied to the test set using the DE genes identified in the training set. d) The level of the first principal component in samples in the test set. PC1 discriminates poorly between HRD and HRP samples which suggests that HRD expression signatures is not a generalisable predictor. Notably these genes do not include known HR genes and given their diverse functions their dysregulation is likely to be a consequence rather than a cause of HRD.