Pan-cancer landscape of homologous recombination deficiency

Luan Nguyen[1], John Martens[2,3], Arne Van Hoeck[1,$], Edwin Cuppen[1,4,$, *]

[1] Center for Molecular Medicine and Oncode Institute, UMC Utrecht, The Netherlands
[2] Erasmus Medical Center, Rotterdam, The Netherlands
[3] Center for Personalized Cancer Treatment, Rotterdam, The Netherlands
[4] Hartwig Medical Foundation, Amsterdam, The Netherlands
[$] contributed equally, joint last author

[*] correspondence to ecuppen@umcutrecht.nl

Abstract

Homologous recombination deficiency (HRD) results in impaired double strand break repair and is a frequent driver of tumorigenesis. Here, we used a machine learning approach to develop a sensitive pan-cancer **C**lassifier of **HO**mologous **R**ecombination **D**eficiency (CHORD). CHORD employs genome-wide genomic footprints of somatic mutations characteristic for HRD and that discriminates BRCA1- and BRCA2-subtypes. Analysis of a metastatic pan-cancer cohort of 3,504 patients revealed HRD to occur at a frequency of 6% with highest rates for ovarian cancer (30%), comparable frequencies for breast, pancreatic and prostate cancer (12-13%) and incidental cases in other cancer types. Ovarian and breast cancer were equally driven by BRCA1- and BRCA2-type HRD, whereas for prostate, pancreatic and urinary tract cancers BRCA2-type HRD was predominant. Biallelic inactivation of BRCA1, BRCA2, RAD51C and PALB2 were found as the most common genetic causes of HRD (60% of all CHORD-HRD cases), with RAD51C and PALB2 inactivation resulting in BRCA2-type HRD. Loss of heterozygosity (LOH) was found to be the main inactivating mechanism in ovarian, breast and pancreatic cancer, whereas for prostate cancer deep deletions (primarily of BRCA2) are also a major cause. From the remaining 40% of CHORD-HRD patients, 35% had a monoallelic mutation in one of the HRD associated genes, suggesting that the second allele could be inactivated by epigenetic or regulatory mechanisms. For only 5% of CHORD-HRD patients, no mutations were identified that could explain the HRD phenotype. Taken together, our results demonstrate that broad genomics-based HRD testing is valuable for cancer diagnostics and could be used for patient stratification towards treatment with e.g. poly ADP-ribose polymerase inhibitors (PARPi).

Introduction

The homologous recombination (HR) pathway is essential for high-fidelity DNA double strand break (DSB) repair and involves numerous genes including BRCA1 and BRCA2. HR deficiency (HRD) due to inactivation of such genes leads to increased levels of genomic alterations[1]. HRD is a common characteristic of many tumors and is frequently observed in breast and ovarian cancer[2]. Accurate detection of HR deficiency (HRD) is of clinical relevance as it is indicative of sensitivity to targeted therapy with poly ADP-ribose polymerase inhibitors (PARPi)[3,4] as well as to DNA damaging reagents[1].

In the clinic, germline BRCA1/2 mutation status is currently the main genetic biomarker of HRD[5]. However, germline testing has its drawbacks: i) it is dependent on the completeness and accuracy of clinical variant annotation databases (e.g. ClinVar); ii) epigenetic silencing is overlooked; iii) partial/complete deletions of the BRCA1/2 loci are missed by current clinical genetic testing, resulting in BRCA1/2 status reporting based on the wild type allele from contaminating normal tissue; and iv) HRD can be driven purely by somatic events. Furthermore, the focus on BRCA1/2 overlooks inactivation of other HR pathway genes. Consequently, patients may receive incorrect treatment or miss out on treatment opportunities, thus necessitating the development of better biomarkers for HRD.

It was recently shown that somatic passenger mutations, which are identified efficiently by whole genome sequencing (WGS), can provide insights into the mutational processes that occurred before and during tumorigenesis, paving the way for novel opportunities for clinical tumor diagnostics[6]. HRD tumors leave behind characteristic mutational footprints in the tumor genome as a product of their dependency on alternative more error-prone pathways to repair DSBs, such as non-homologous end-joining (NHEJ) or microhomology mediated end-joining (MMEJ)[7]. Indeed, some mutational footprints were found to be associated with BRCA1/2 deficiency, namely deletions with flanking microhomology, as well as several 'mutational signatures' including two COSMIC single nucleotide variant (SNV) signatures and two structural variant (SV) signatures[8]. These features were used to develop a breast cancer-specific predictor of HRD known as HRDetect[9]. Application of this tool in tumors revealed that the prevalence of HRD extends beyond BRCA1/2 deficient breast cancer tumors. However, aside from isolated reports, e.g. in ovarian[10] or prostate cancer[11], the occurrence of HRD among other cancer types has not yet been systematically delineated.

Here, we describe the development of a random forest-based **C**lassifier of **HO**mologous **R**ecombination **D**eficiency (CHORD) for pan-cancer HRD detection. For this, we used a large cohort comprised of 3,824 whole genome sequenced tumors across 31 different cancer types[12]. With this model, we demonstrate that accurate prediction of HRD is possible across cancer types, using specific SNV, indel and SV types as input. Importantly, we identified mutational inactivation of BRCA1, BRCA2, RAD51C and PALB2 as the most frequent cause of HRD, with the latter two genes resulting in the same mutational footprints as BRCA2. In addition, we found that the underlying genetic inactivation mechanisms vary between genes and across cancer types. Finally, we highlight the potential of CHORD to improve the clinical management of patients, especially those with non-breast and ovarian cancer.

Results

Random forest classifier training

For the development of CHORD, we used WGS data of 3,824 solid tumor from 3,504 patients from the pan-cancer metastatic cohort of the Hartwig Medical Foundation (HMF)[12]. From these, we selected tumor samples with biallelic loss of BRCA1 or BRCA2, and non-mutated BRCA1/2, to obtain a high confidence set of samples belonging to 3 classes for classifier training (BRCA1-deficient, BRCA2-deficient, and BRCA1/2-proficient). To this end, we screened each sample to identify those samples carrying a pathogenic germline or somatic BRCA1/2 SNV/indel (as annotated in ClinVar, or a frameshift) in combination with LOH, as well as those having complete copy number loss (i.e. deep deletion) of BRCA1/2. This unbiased approach revealed 35 and 90 samples with BRCA1 and BRCA2 biallelic loss of function, respectively, which were labeled as HRD for the training. Conversely, 1,358 samples were labeled as HR proficient (HRP) as these samples were observed to carry at least one functional allele of BRCA1/2. In total, 1,483 out of 3,824 samples (38.8% of the HMF dataset) were used to train the classifier (**Supplementary figure 1**).

Three main somatic mutation categories were used as features for training (**Figure 1a**), which included single nucleotide variants (SNVs) subdivided by base substitution (SBS) type; indels stratified by flanking sequence context, including tandem repeats or sequence homology; and structural variants (SV), stratified by type and length. The occurence of each mutation type was counted, yielding a 28-feature mutation contribution profile for each sample. From this, relative contributions per mutation category were calculated to account for differences in mutational load across samples (**Figure 1a**). These features are henceforth collectively referred to as 'mutation contexts'.

A random forest was then trained to predict the probability of BRCA1 or BRCA2 deficiency (**Figure 1b**). Briefly, a core training procedure performed feature selection and class resampling (to alleviate the imbalance between the 3 classes). This core procedure was subjected to 10-fold cross-validation (CV) which was repeated 100 times to filter samples from the training set that were not consistently HRD or HRP. A sample was considered HRD if the sum of the BRCA1 and BRCA2 deficiency probabilities (henceforth referred to as the HRD probability) was ≥0.5. This core procedure was reapplied to the filtered training set to yield the final random forest model, CHORD, which also uses the HRD probability of 0.5 as the classification threshold (**Supplementary figure 2a,b**; **Supplementary figure 3**).

The presence of deletions with flanking microhomology (del.mh) was found to be the most important predictor of HRD. This is consistent with previous studies showing that HRD cells rely on microhomology-mediated end joining (MMEJ) as an alternative DSB repair pathway[13,14]. Additionally, 1-10kb and to a lesser extent 10-100kb duplications (DUP_1e03_1e04_bp and DUP_1e04_1e05_bp, respectively) were important for predicting BRCA1 deficiency but not BRCA2 deficiency. This is likely due to the role of BRCA1 in the suppression of ~10kb duplications during the resolution of replication fork stalling[15], and implies that this phenotype is in fact a product of two individual mutational processes. Given that deficiencies in other HR genes may produce similar mutation contexts, we have coined the terms 'BRCA1-type HRD' and 'BRCA2-type HRD' to describe these phenotypes (**Figure 1b,c**).

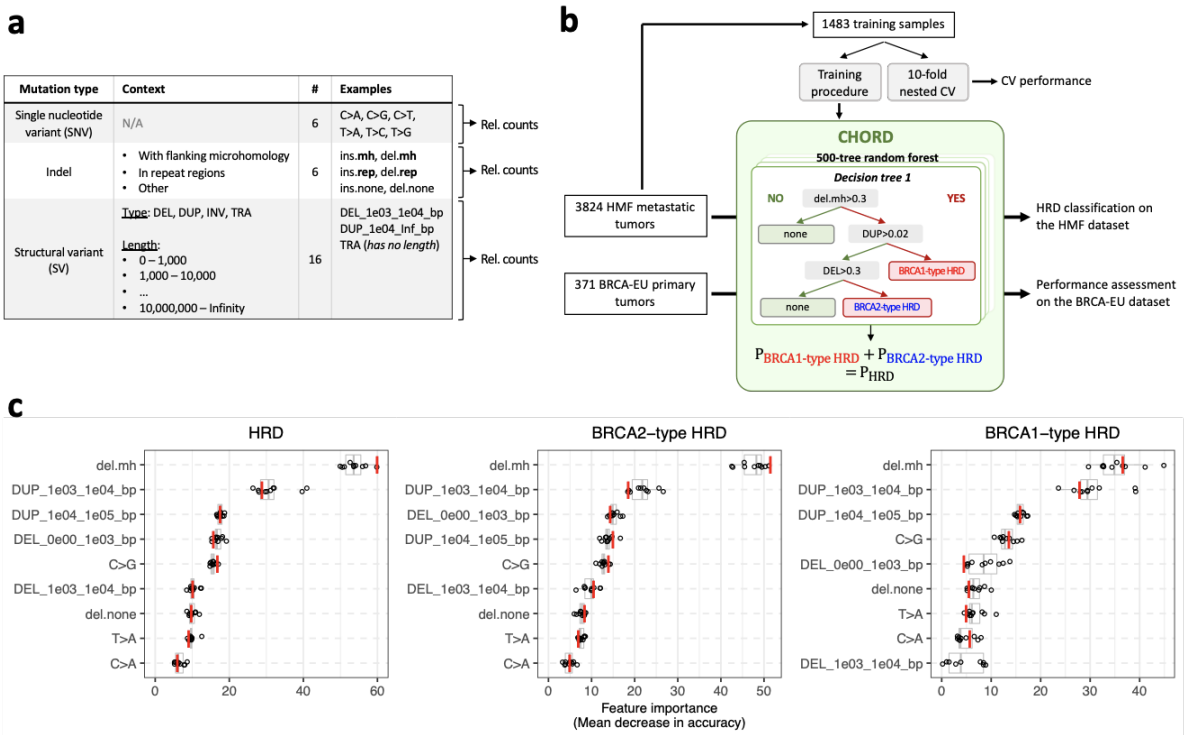


Figure 1: *CHORD is random forest **C**lassifier of **HO**mologous **R**ecombination **D**eficiency able to distinguish between BRCA1- and BRCA2-type HRD phenotypes. (**a**) The features used for training CHORD are relative counts of different mutation contexts which fall into one of three groups based on mutation type. (i) Single nucleotide variants (SNV): 6 possible base substitutions (C>A, C>G, C>T, T>A, T>C, T>G). (ii) Indels: indels with flanking microhomology, within repeat regions, or not falling into either of these 2 categories. (iii) Structural variants (SV): SVs stratified by type and length. Relative counts were calculated separately for each of the 3 mutation types. (**b**) Training and application of CHORD. From a total of 3,824 metastatic tumor samples, 1,483 samples were selected for training CHORD. The model outputs the probability of BRCA1-type HRD and BRCA2-type HRD, with the probability of HRD being the sum of these 2 probabilities. The performance of CHORD was assessed via a 10-fold nested cross-validation (CV) procedure on the training samples, as well as by applying the model to a primary breast cancer dataset (BRCA-EU) from the International Cancer Genome Consortium (ICGC) consisting of 371 tumors. Lastly, CHORD was applied to all samples in the HMF cohort in order to gain insights into the pan-cancer landscape of HRD. (**c**) The features used by CHORD to predict HRD as well as BRCA1-type HRD and BRCA2-type HRD, with their importance indicated by mean decrease in accuracy. Deletions with flanking microhomology (del.mh) was the most important feature for predicting HRD as a whole, with 1-100kb structural duplications (DUP_1e03_1e04_bp, DUP_1e04_1e05_bp) differentiating BRCA1-type HRD from BRCA2-type HRD. Boxplot and dots: Feature importance over 10-fold nested CV on the training set. Red line: Feature importance in the final CHORD model.*

Performance of CHORD

Two independent approaches were used to assess the performance of CHORD. In the first approach, 10-fold CV was performed on the training data which allows every sample to be excluded from the training set after which unbiased HRD probabilities can be determined (**Supplementary figure 2c**). The probabilities of all prediction classes (i.e. HRD, BRCA1-type HRD, BRCA2-type HRD) were highly concordant with the genetic annotations (**Figure 2a**). The concordance between predictions and annotations was quantified by calculating the area under the curve (AUC) of receiver operating characteristic (ROC) and precision-recall (PR) curves (**Figure 2b,c**). CHORD achieved excellent performance as shown by the high AUC-ROC and AUC-PR for all prediction classes (>0.97 and ≥0.88 respectively).

In the second approach, performance was evaluated on an external dataset independent from the training data. Here, we used the mutation contexts from 371 tumor samples of the 560 primary breast

cancer (BRCA-EU) dataset which were used to develop and evaluate the performance of HRDetect[9]. When applying CHORD to these samples, we obtained HRD probabilities that were in agreement with the previously reported BRCA1/2 genetic status[9] (**Figure 2d**). Furthermore, AUC-ROC and AUC-PR values were comparable to those obtained by CV on the HMF training data for all prediction classes (**Figure 2e,f**). However, we still observed some BRCA1 deficient samples predicted as HRP while HRDetect classified these as HRD, and tested whether this was due to differences in somatic calling algorithms. Indeed, using the variants obtained from the native pipeline of the HMF dataset (HMF pipeline[12]) for HRD prediction resulted in overall higher HRD probabilities in BRCA1/2 deficient samples compared to using the variants downloaded from ICGC. This was especially apparent for 2 samples (PD24186 and PD4017) which became HRD using HMF pipeline called mutation profiles, as well as 3 other samples (PD5960, PD23578, PD24191) which increased in HRD probability from slightly above the classification threshold of 0.5, to >0.75 (**Supplementary figure 6**). Except for 1 sample, the HR status of all 45 reanalyzed BRCA-EU samples were correctly predicted by CHORD. These results together demonstrate that, while HMF pipeline called mutational features provide optimal input for CHORD, the model can also accurately predict HRD from non-native input data. Furthermore, the comparable performance of CHORD on primary (BRCA-EU) and metastatic (HMF) cancer datasets indicates that HRD can be predicted independent of disease progression or staging.

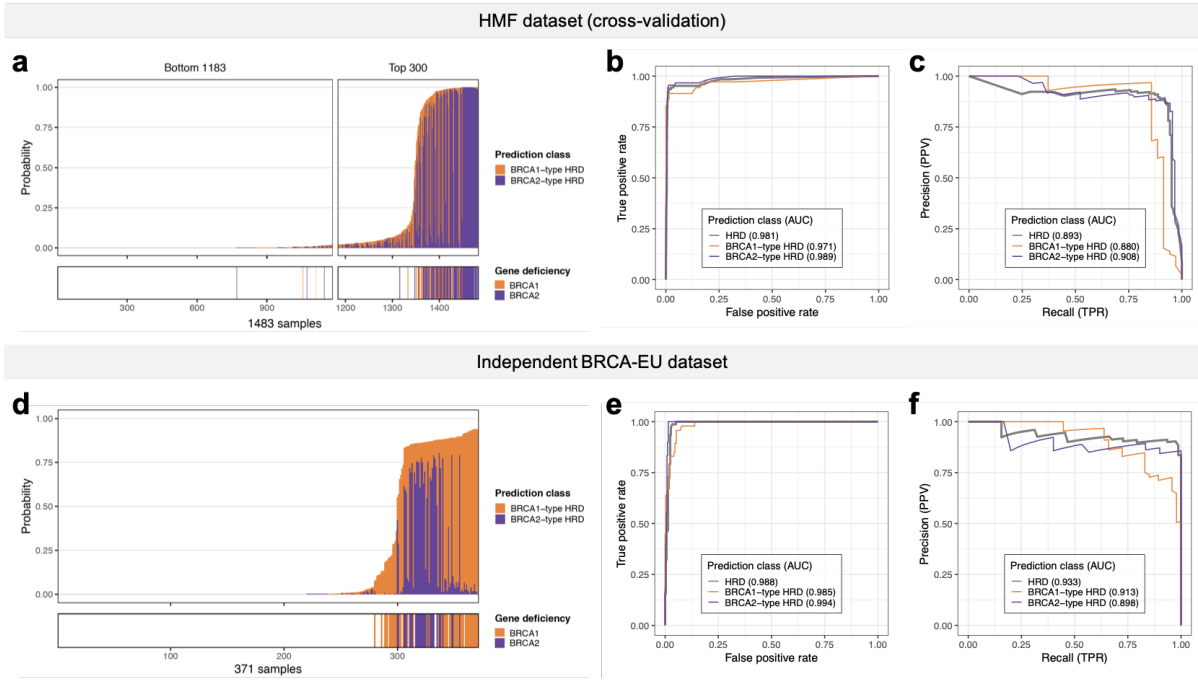


***Figure 2**: Performance of CHORD as determined by 10-fold cross-validation (CV) on the HMF training data or prediction on an independent dataset (BRCA-EU; primary breast cancer dataset). (**a, d**) The probability of HRD for each sample (total bar height) with each bar being divided into segments indicating the probability of BRCA1- (orange) and BRCA2-type HRD (purple). Stripes below the bar plot indicate biallelic loss of BRCA1 or BRCA2. In (**a**), probabilities have been aggregated from the 10 CV folds. (**b, e**) Receiver operating characteristic (ROC) and (**c, f**) precision-recall curves (PR) showing the performance of CHORD when predicting HRD as a whole (grey), BRCA1-type HRD (orange), or BRCA2-type HRD (purple).*

We note that CHORD performs similarly to HRDetect, which unlike CHORD, relies on COSMIC SNV signatures[6,7] and SV signatures[9]. To validate whether mutational signatures result in improved HRD prediction, we also trained a random forest model (CHORD-signature) that uses COSMIC and SV signatures as input instead of mutation contexts. CHORD-signature achieved similarly to CHORD (**Supplementary figure 8**), which can be explained by the reliance on similar features (**Supplementary figure 7**), namely microhomology deletions and SV signature 3 (SV3; analogous to 1-100kb duplications). COSMIC signature 3 (proposed as a sensitive marker for HRD in recent studies[16–18]) is actually a less important feature for predicting HRD than microhomology deletions in

both models, indicating that microhomology deletions serves as a better (univariate) marker of HRD compared to COSMIC signature 3, which may hamper HRD detection in tumors with already prominent endogenous signature 3 contribution. We conclude that accurate detection of HRD does not require mutational signatures, thereby simplifying HRD calling and avoiding the complications associated with the fitting step required for computing signature contributions in individual samples (for which there is currently no consensus approach)[19].

Prerequisites for accurate HRD prediction

Although CHORD detects HRD independent of cancer type, it is important to note that microsatellite instability (MSI) negatively affects CHORD's ability to accurately predict HRD. MSI is a hypermutator phenotype characterized by an exceptional number of indels in regions with (tandem) repeats. As CHORD uses relative values of mutation contexts as input, MSI results in a reduction of the relative contribution of microhomology deletions and thus an underestimated HRD probability ('false negatives'). These false negatives can be seen in **Supplementary figure 4**, where all MSI samples were predicted HRP by CHORD even though four of these samples had biallelic loss of BRCA2 (**Supplementary table 4**). This could be circumvented by incorporating a MSI trained HRD classifier within CHORD. However, the number of HRD predicted samples with MSI is currently too small for training such a classifier and therefore we have built in MSI status checking as a quality control (QC) step within CHORD.

To determine the minimum number of mutations required for accurate HRD prediction, we progressively down-sampled all mutations for each sample in the training set and measured the reduction in performance. We found that at least 50 indels were required for accurately predicting HRD, while at least 30 SVs were required for distinguishing BRCA1-type from BRCA2-type HRD (**Supplementary figure 5**). These threshold levels may be particularly relevant for samples with low tumor purity and/or read coverage, and are as such also included as a QC step within CHORD.

All subsequent analyses included only tumors fulfilling the above criteria (i.e. MSI absent, ≥50 indels, ≥30 SVs, thereby excluding 80 samples). Furthermore, a single tumor per patient was selected for those with multiple biopsies in the cohort database (based on highest tumor purity), though all patients had consistent HRD probabilities across all biopsies (**Supplementary table 1**). In total, 3,504 tumors were selected to represent each patient for our subsequent pan-cancer analyses (**Supplementary table 9**).

BRCA2, RAD51C and PALB2 are associated with BRCA2-type HRD while only BRCA1 is associated with BRCA1-type HRD

We applied CHORD to the HMF metastatic cancer cohort and found that 211 (5.9%) patients were classified as being homologous recombination deficient (CHORD-HRD) (**Figure 3a**), with 65 (31%) characterized as having BRCA1-type HRD and 146 (69%) as having BRCA2-type HRD. The remaining 3,293 patients were classified as being proficient in homologous recombination-based DNA repair (CHORD-HRP) (**Figure 3c**).

We then sought to identify the key mutated genes underlying the HRD phenotype by performing an enrichment analysis of biallelically inactivated genes in CHORD-HRD vs. CHORD-HRP patients. For this analysis, we started from a list of 781 genes that are cancer related (based on the catalog of genes from Cancer Genome Interpreter) and/or HR related (manually curated based on the KEGG HR pathway, as well as via literature search) (**Supplementary table 8**). For these genes, we considered likely pathogenic variants (according to ClinVar) as well as predicted impactful variants such as nonsense mutations to contribute to gene inactivation (see *Methods*). This revealed that, in addition

to BRCA1 and BRCA2 ($q<10^{-5}$ for both genes, Fisher's exact test), RAD51C and PALB2 ($q<0.001$ and $q<0.1$ respectively, Fisher's exact test) were also significantly enriched amongst HRD patients using a q-value threshold of 0.1 (**Figure 3b**). Of all CHORD-HRD patients, 59.7% (126/211) could be explained by biallelic inactivation of either BRCA2 (cluster 1; n=84), BRCA1 (cluster 5; n=31), RAD51C (cluster 2; n=5), or PALB2 (cluster 3; n=6) (**Figure 3c**). RAD51C and PALB2 were recently linked to HRD as incidental cases using mutational signature based approaches[16,20] and our results now confirm that biallelic inactivation of these two genes results in HRD and is actually a common cause of HRD (albeit to a lesser extent than for BRCA1/2). RAD51C and PALB2 deficient patients shared the BRCA2-type HRD phenotype (absence of duplications) with BRCA2 deficient patients (clusters 1-3; **Figure 3c**), consistent with previous studies[20,21]. On the other hand, only BRCA1 deficient patients (cluster 5) harbored the BRCA1-type HRD phenotype (1-100kb duplications).

Of note, we observed one patient (**Figure 3c**; patient #1) bearing a known pathogenic frameshift mutation in BRCA1 (**Supplementary table 4**; patient HMF001925, c.1961dupA), which based on current practices for detecting HRD in the clinic (testing for pathogenic SNVs/indels)[5] would be considered the driver mutation. However, our genetic analysis indicates that the deep deletion in BRCA2 (which would be missed by testing for SNVs/indels) was the cause of HRD, which is supported by the lack of LOH in BRCA1, as well as the BRCA2-type HRD phenotype of this patient.

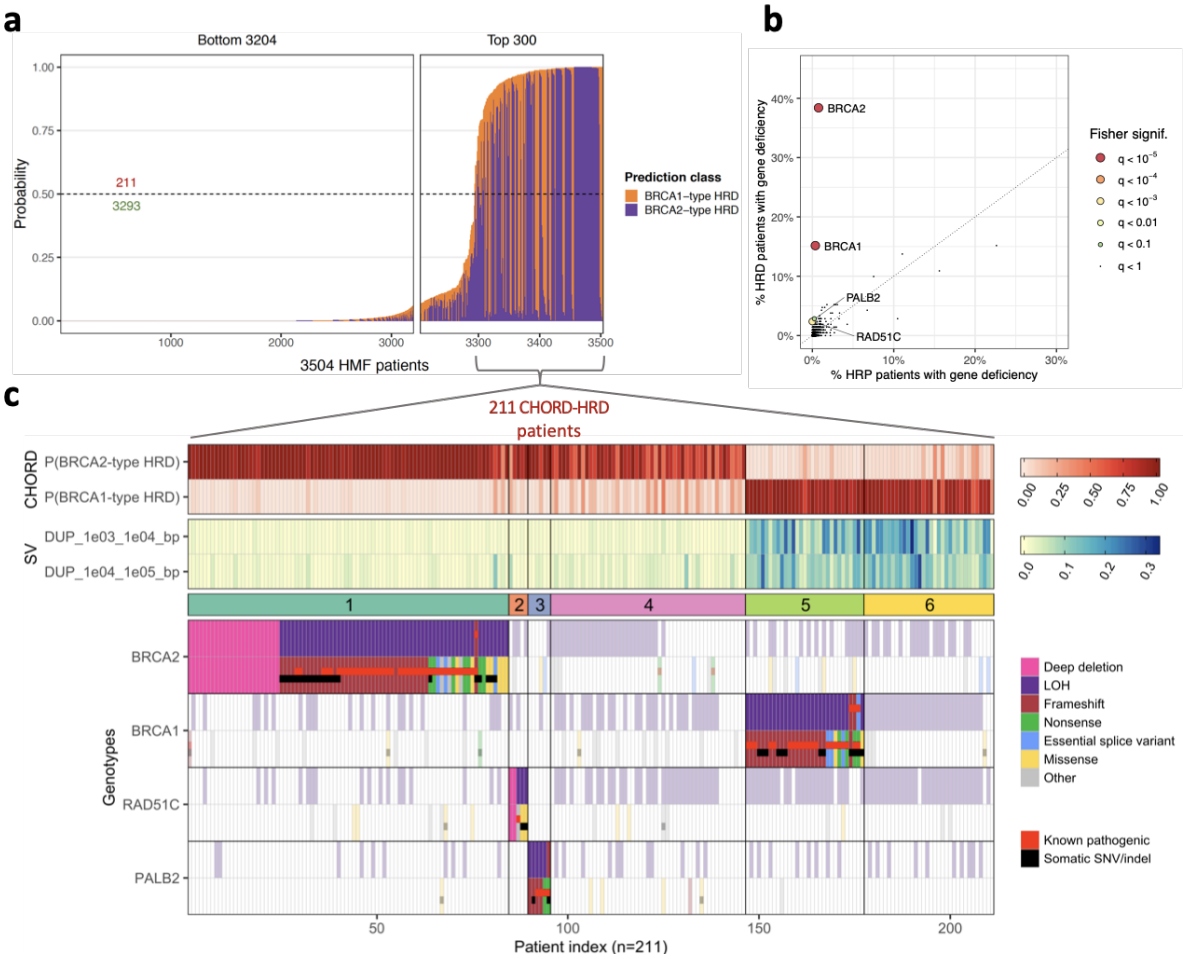


***Figure 3**: The genetic causes of HRD. (**a**) The bar plot shows the probability of HRD for each patient (total bar height) with each bar being divided into segments indicating the probability of BRCA1- (orange) and BRCA2-type HRD (purple). 211 patients were predicted HRD while 3,293 were predicted HRP by CHORD. (**b**) A Fisher's exact test identified enrichment of BRCA1, BRCA2, RAD51C and PALB2 biallelic inactivation in CHORD-HRD vs. CHORD-HRP patients (from a list of 781 cancer and HR related genes). Each point represents a gene with its size/color corresponding to the statistical significance as determined by the Fisher's exact test, with axes indicating the percentage of patients (within either the CHORD-HRD or CHORD-HRP group) in which biallelic inactivation was detected. Multiple testing correction was performed using the*

*Benjamini-Hochberg procedure. (**c**) Biallelic inactivation of BRCA2, RAD51C and PALB2 was associated with BRCA2-type HRD, whereas only BRCA1 inactivation was associated with BRCA1-type HRD. Top: BRCA1- and BRCA2-type HRD probabilities from CHORD. Middle: SV contexts used in CHORD to distinguish BRCA1- from BRCA2-type HRD. Bottom: The biallelic status of each gene. Patients were clustered according to the likely cause of HRD, with clusters 1, 2, 3, and 5 corresponding to patients with identified inactivation of BRCA2, RAD51C, PALB2 and BRCA1, while clusters 4 and 6 correspond to patients without clear biallelic inactivation of these 4 genes. Tiles marked as 'Known pathogenic' refer to variants having a 'pathogenic' or 'likely pathogenic' annotation in ClinVar.*

In 40.3% of CHORD-HRD patients (n = 85), there was no clear indication of biallelic loss of BRCA1/2, RAD51C or PALB2 (henceforth referred to as the 'HRD associated genes') (clusters 4 and 6, 85/211; **Figure 3c**) or any other DNA repair-related gene. However, 75 patients in clusters 4 and 6 (=35.5% of CHORD-HRD patients) had a deleterious event in a single allele of one of the HRD associated genes (all had LOH except for one patient where a pathogenic nonsense variant was found), with a similar gene-tissue distribution in these patients as the biallelically affected patients (with the exception of prostate cancer due to the high frequency of BRCA2 deep deletions) (**Supplementary figure 12**). We found enrichment of LOH in the HRD associated genes in CHORD-HRD versus CHORD-HRP patients (184/211 (87%) vs. 1738/3293 (53%); p=1.397E-6, Fisher's exact test) which implies the involvement of LOH in gene inactivation for the 75 patients in clusters 4 and 6. This is consistent with the finding by Jonsson et al.[17] that LOH is enriched in tumors with BRCA1/2 germline pathogenic variants or somatic loss-of-function variants. Davies et al.[9] showed that promoter methylation of BRCA1/2 was present in 35% of ovarian and 24% of breast primary cancers with HRD (**Table 1**). Aside from BRCA1/2, RAD51C promoter methylation was also reported in HRD tumors (though at a lower incidence compared to BRCA1/2) in previous studies[16,21]. Thus, BRCA1/2 and RAD51C promoter methylation (or deregulation of other regulatory elements), likely in combination with LOH, may have led to the HRD phenotype for a sizable portion of the 50% of ovarian and 49% of breast cancer patients (**Table 1**) with no clear biallelic loss of the HRD associated genes, and potentially for patients with other cancer types as well. Unfortunately, we could not systematically assess this as methylation and transcript data was not available for the HMF dataset.

We also cannot rule out the possibility that deficiencies in other HR genes that did not reach significance in our enrichment analysis, underlie the HRD phenotype for a small number of patients in clusters 4 and 6. We indeed identified 14 patients with biallelic inactivation of a HR gene other than BRCA1/2, RAD51C or PALB2, and 1 patient with a likely inactivating biallelic event (patient #99; LOH in combination with a nonsense variant in CHEK1) (**Supplementary figure 13**). Notably, the 4 patients with RAD51B (patients #113, 121) and XRCC2 (patients #107, 120) deficiency were all predicted to have BRCA2-type HRD, a phenotype shared with RAD51C deficient patients. Given that these 3 genes all belong to the RAD51 paralog complex BCDX2[22], the BRCA2-type HRD suggests that RAD51B and XRCC2 deficiency could have led to HRD in these patients . Likewise, the 5 patients with deficiencies in the BRCA1-binding proteins, ABRAXAS1[23] (patient #202), BARD1[24] (patient #182), BRIP1[25] (patient #196), FANCA[26] (patient #201) and RBBP8[27] (patient #190), were all predicted as having BRCA1-type HRD. Thus, while we could not conclusively determine the cause of HRD for patients in clusters 4 and 6, we postulate that HRD in these patients may have been a result of epigenetic silencing of BRCA1/2 or RAD51C, deficiencies in other HR genes (not associated to HRD in our analysis), or possibly a result of other unknown regulatory mechanisms.

CHORD as a tool to uncover novel pathogenic variants

The HRD predictions from CHORD can provide supportive evidence for interpreting variants of unknown significance (VUS), either germline or somatic, especially for rare variants for which pathogenicity has not been determined experimentally. From 11 CHORD-HRD patients (**Figure 3c**; BRCA2: patients #77-84; BRCA1: patient #177; RAD51C: patients #88-89), we identified 13 variants (5 germline and 7 somatic) not previously described to be pathogenic (**Supplementary table 6**; 9 missense, 2 nonsense, 2 splice variants) but which in combination with LOH could explain biallelic loss

of a HR gene. Furthermore, biallelic loss of the respective gene corresponded to the associated HRD subtype, providing additional support that these variants are indeed most likely pathogenic.

Of the 13 variants, 9 (4 in BRCA2; 1 in BRCA1; 2 in RAD51C) were the sole variant found in the respective gene and respective patient, and may thus be pathogenic driver variants. Two of these (both in BRCA2: missense c.8045C>T, splice acceptor c.8351G>A) had a 'VUS' annotation in ClinVar but a low population frequency according to gnomAD[28] as well as being predicted as 'deleterious' by SIFT[29] and 'probably damaging' by PolyPhen[30], supporting their potential pathogenicity. For the remaining 4 variants (all in BRCA2) of the 13, two missense variants (c.9230T>C, c.9254C>T) were found in patient HMF000429. Here, further validation is required to determine which was the driver mutation (or possibly the combination). The other two variants were found in patient HMF002600: a missense variant (c.1792A>G) with a high population frequency according to gnomAD and a 'tolerated' annotation by SIFT, and splice acceptor variant (n.457-3dupT), suggesting that the latter was the driver mutation. In total, we identified 10 variants which may be pathogenic but warrant further investigation.

The incidence and genetic cause of HRD varies across tissue types

We next investigated the differences in the incidence and genetic causes of HRD in CHORD-HRD patients based on primary tumor location (**Figure 4a, Supplementary table 7**). HRD was found to occur commonly in ovarian (30%), breast (12%), pancreatic (13%), and prostate (13%) cancer; to a lesser extent in biliary (6.8%) and urinary tract (6.4%) cancer; and sporadically in other cancer types. Overall, non-breast and -ovarian cancer patients accounted for 43% (92/211) of all CHORD-HRD patients. This finding, together with the observation that HRD occurs equally frequently in pancreatic and prostate cancer as with breast cancer is striking given that HRD is typically associated with breast (and ovarian) cancer[5].

Across different cancer types, we observed differences in gene deficiencies that contributed to HRD (**Figure 4b, Supplementary table 7**). BRCA2-type HRD related gene deficiencies (including BRCA2, RAD51C, PALB2 deficiencies) were disproportionately frequent in pancreatic (91%), prostate (95%), and urinary tract (89%) cancer. On the other hand, BRCA1-type HRD related gene deficiencies were more often found in ovarian (50%), breast (39%), and biliary (50%) cancer. Whether these differences can be linked to a biological cause or have prognostic value remains to be determined.

In 95% of all CHORD-HRD patients, we found mono- or biallelic inactivation of at least one of the four main HRD associated genes (BRCA1, BRCA2, PALB2, RAD15C; **Figure 4c, Supplementary table 7**). In the case of biallelic inactivation, we observed LOH to be the dominant secondary event, occurring in combination with a germline SNV/indel in 33% of patients and with a somatic SNV/indel in 16% of patients. LOH of BRCA1/2, RAD51C or PALB2 was also found as a monoallelic event, mainly in ovarian (43%) and breast (40%) cancer patients. As indicated above, the other allele may be inactivated by epigenetic mechanisms in these patients. Alternatively, since the majority of these patients belong to clusters 4 and 6 (**Figure 3c)**, it is possible that deficiencies in other HR genes were responsible for some of these HRD cases and that the observed LOH in BRCA1/2, RAD15C or PALB2 are passenger events (**Supplementary figure 13**).

Interestingly, we find that deep somatic deletions are frequent contributors to biallelic loss of BRCA2 or RAD51C, occurring in 12% of CHORD-HRD patients across different cancer types (**Figure 4c, Supplementary table 7**). However, deep deletions (primarily of BRCA2; **Supplementary figure 14**) occurred much more frequently in prostate cancer (37%) compared to other cancer types, consistent with previous literature[31]. When excluding prostate cancer, the overall frequency of deep deletions dropped to ~5%. Nevertheless, deep deletions of HRD genes were found in every cancer type with a

high frequency of CHORD-HRD indicating that complete somatic gene loss is a common and highly underestimated cause of HRD.

We find that biallelic gene loss is often caused exclusively by somatic events, also when involving SNVs/indels and occurs in 77% of prostate, 46% of ovarian, 22% of pancreatic, 30% of breast and 33% of biliary CHORD-HRD cancer patients with biallelic gene loss (**Figure 4d**, **Supplementary table 7**). The high somatic HRD frequency in prostate cancer was driven by the high frequency of deep deletions (15 out of 26 patients), as well as LOH in combination with a somatic SNV/indel (11 out of 26 patients) (**Figure 4c, Supplementary table 7**). Although these frequencies may not be fully representative for each cancer type due to the proportion of patients with unknown mutation status in at least one allele (indicated as 'Unknown' in **Figure 4d**), these observations do emphasize that somatic-only events should not be overlooked as a mechanism of HR gene inactivation.

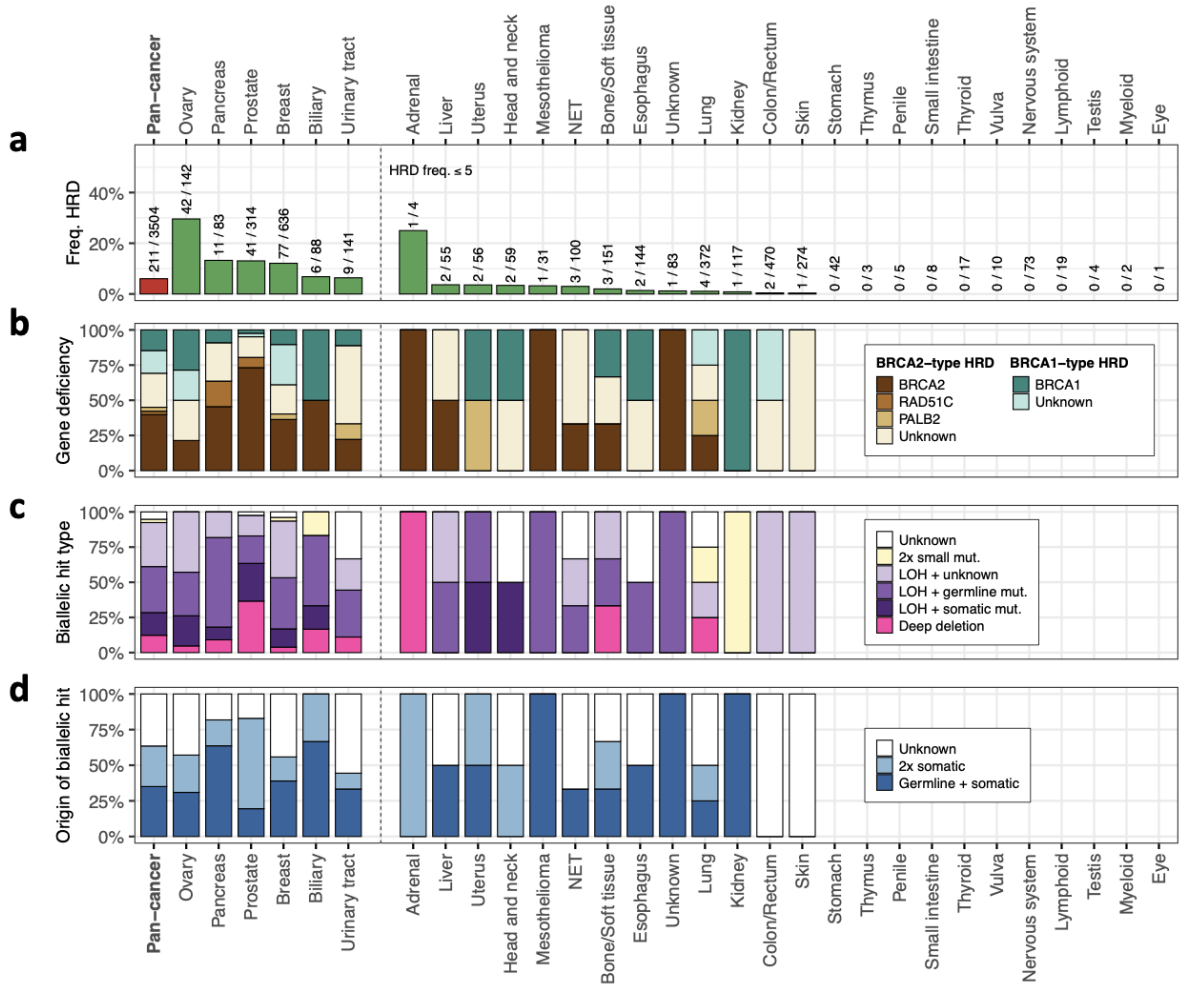
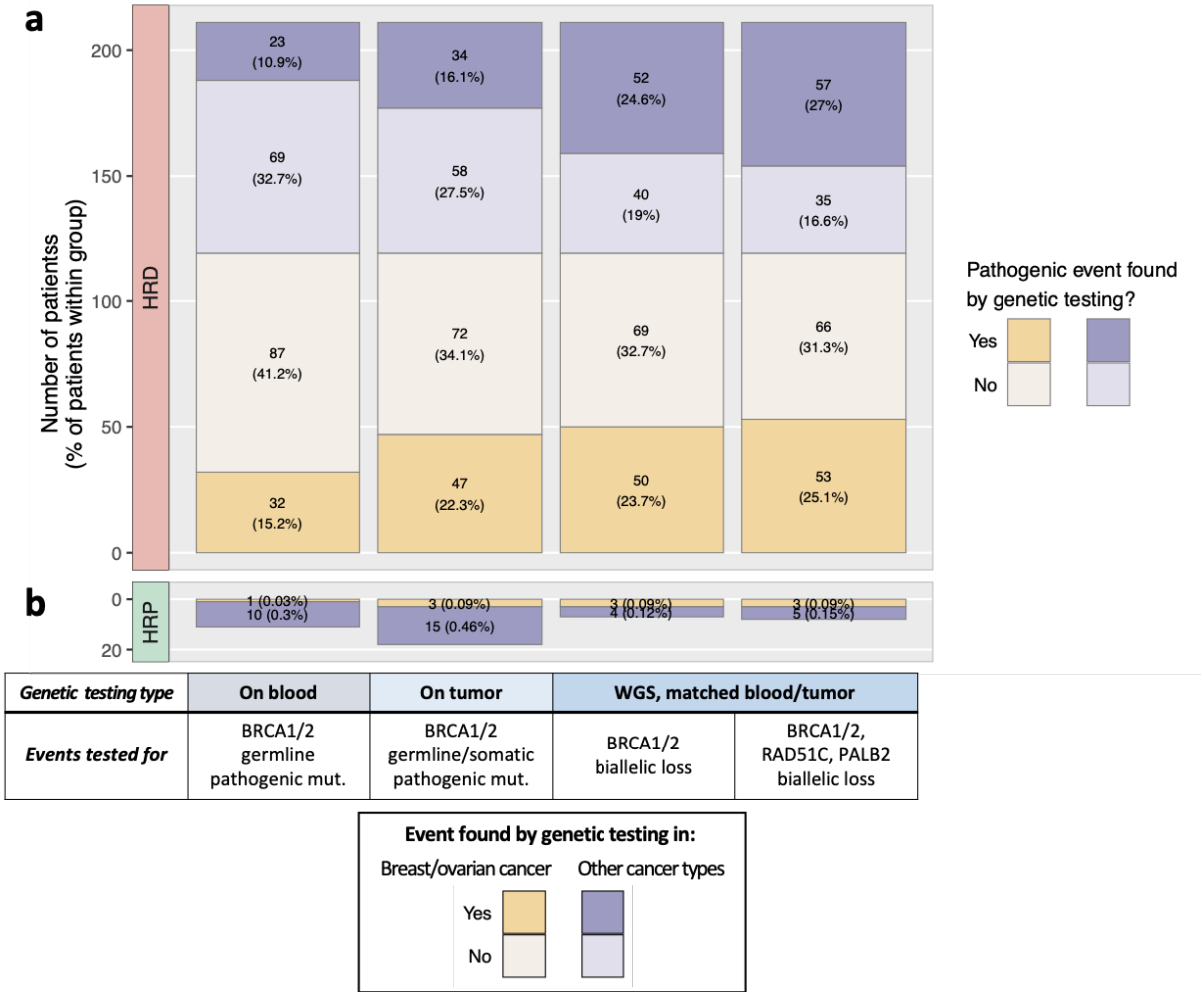


***Figure 4**: Percentage breakdown of the incidence and genetic causes of HRD in CHORD-HRD patients pan-cancer and by cancer type. (**a**) Frequency of HRD. Cancer types with low frequency (≤ 5 cases) of HRD lie on the right side of the dashed line. (**b**) The gene deficiency associated with HRD. Bar segments are grouped into BRCA2-type HRD genes (BRCA2, RAD51C, PALB2) and BRCA1-type HRD genes (BRCA1 only). (**c**) The likely combination of biallelic events in BRCA1/2, RAD51C or PALB2 causing HRD. (**d**) Whether the genetic cause of HRD was purely due to somatic events, due to germline predisposition, or unknown. In (**c** and **d**), 'Unknown' and/or 'LOH + unknown' bar segments refer to patients where no clear biallelic loss of the aforementioned BRCA1/2, RAD51C, or PALB2 was identified (i.e. clusters 4 and 6, **Figure 3c**).*

CHORD can detect HRD in a substantial number of cases that would be missed by genetic testing

To assess the potential value of CHORD in a clinical setting, we compared CHORD's predictions to the hypothetical outcomes of common genetic testing approaches (**Figure 5**).

In the clinic, HRD detection is currently often done by screening for pathogenic BRCA1/2 SNVs/indels based on annotations from curated databases (e.g. ClinVar). This is performed either on blood biopsies (blood genetic testing), analogous to screening for germline SNVs/indels in sequencing data of the HMF cohort (which was analysed by tumor-normal pair whole genome sequencing); or on tumor biopsies (tumor genetic testing), analogous to screening both germline and somatic SNVs/indels[5]. Our genetic analyses (**Figure 5a**) indicate that blood genetic testing would identify a pathogenic BRCA1/2 SNV/indel (according to ClinVar; or an out-of-frame frameshift) in 15% of CHORD-HRD breast/ovarian cancer patients (cancer types which genetic testing is often restricted to), while tumor genetic testing would increase this proportion to 22%. If patients with other cancer types would be included, blood and tumor genetic testing would identify 26% and 38% of CHORD-HRD tumors (respectively) with a pathogenic BRCA1/2 SNV/indel.



Genetic testing type	On blood	On tumor	WGS, matched blood/tumor	
Events tested for	BRCA1/2 germline pathogenic mut.	BRCA1/2 germline/somatic pathogenic mut.	BRCA1/2 biallelic loss	BRCA1/2, RAD51C, PALB2 biallelic loss

***Figure 5**: CHORD identifies a large proportion of HRD patients that would be missed by genetic testing. (**a**) and (**b**) show the percentage of CHORD-HRD or CHORD-HRP patients (respectively) from the HMF dataset in which a pathogenic event was found based on four genetic testing setups. In the first two setups ('On blood', 'On tumor'), a pathogenic event was identified if a pathogenic SNV/indel was found on one allele, which was defined as a frameshift, or a likely pathogenic or pathogenic variant according to ClinVar. In the 'WGS based testing' setups, the pathogenic SNV/indel must also have occurred in combination with loss of heterozygosity (LOH); or, a deep deletion was identified.*

While not currently routinely performed in the clinic, WGS based genetic testing with matched blood/tumor biopsies (WGS genetic testing) would allow the detection of any event (including structural events such as LOH or deep deletions) that contributes to HR gene inactivation, and enables the determination of biallelic gene status. Our analyses show that WGS genetic testing would increase the number of patients that are considered HRD compared to SNV/indel-based blood/tumor genetic testing, with biallelic loss of BRCA1/2 being identified in 24% of CHORD-HRD breast/ovarian cancer patients and in 48% patients pan-cancer (**Figure 5a**). Additionally, WGS genetic testing would consider 7 CHORD-HRP patients as HRD ('genetic testing false positives'; although we cannot exclude that some of these are CHORD false negatives), a marked decrease compared to tumor genetic testing which would identify 18 false positives (**Figure 5b**). By including the two other main HRD associated genes (RAD51C and PALB2) in WGS genetic testing, biallelic gene loss would be identified in 52% of patients (**Figure 5a**), while only increasing the number of genetic testing false positives from 7 to 8 patients (**Figure 5b**). Our findings show that while WGS genetic testing (for biallelic loss) offers improved detection of HRD patients compared to testing for pathogenic SNVs/indels, it still misses ~50% of HRD patients as classified by CHORD. Most of these patients have monoallelic loss (almost all due to LOH and not by SNV/indels) of one of the HRD associated genes (**Figure 3c**). Importantly, patients who have no clear biallelic loss of BRCA1/2, RAD51C or PALB2 but with a CHORD-HRD phenotype could still benefit from PARPi, as was shown by Staaf et al.[21]. While these cases could be identified by screening for monoallelic variants in HRD-associated genes, it should be noted that this identifies many variants that would be classified as VUS, requiring further evidence to be informative in a clinical setting.

In our analysis of genetic testing outcomes, the HR status of each patient as predicted by CHORD was assumed to be the ground truth. However, we do acknowledge that a major limitation of using mutational scars is that they represent genomic history and not necessarily current on-going mutational processes. Consequently, a patient might exhibit a CHORD-HRD phenotype but in reality be HRP – for example by restoring the reading frame by a secondary frameshift[32]. To identify such cases, we examined the subclonal fraction of microhomology deletions but observed no drop in subclonal microhomology deletions in any CHORD-HRD patients indicating that the HRD mutational process, unlike APOBEC[33], has a stable mutation rate. Secondary frameshift mutations have been described as a mechanism reverting HRD status[34,35], but we identified such potential frameshifts in only four CHORD-HRD patients. However, it remains unclear whether HR function is restored in these patients as we did not observe a decrease in the proportion of microhomology deletions among subclonal mutations (**Supplementary figure 15**). It should be noted that the number of revertants by secondary frameshifts may be much more frequent as a resistance mechanisms, but this could not be determined with the cohort used here as only 11 out 3,504 patients had undergone a PARPi treatment before the tumor sample was collected.

We also identified 8 patients with biallelic loss of BRCA1/2 or PALB2, but which were classified as HRP by CHORD. This discrepancy could be due to incorrect local ploidy determination, which becomes less reliable at lower tumor purities of the tested sample, and would thus represent false positives from WGS genetic screening. These results could also be explained by recent acquirement of the HRD phenotype and thus represent false-negative CHORD cases. Indeed, one of these 8 cases had a non-zero HRD probability of 0.118 suggesting that this patient has to some extent acquired HRD associated mutational scars (**Supplementary table 5**).

Thus, while CHORD can detect HRD independent of the underlying cause, genetic testing of HRD genes is complementary and can provide supporting information for making a final verdict on a patient's HR status, especially for cases where HRD only recently occurred. The unique advantage of using WGS, although not routine in clinical diagnostics yet, but likely in the near future[36], is that both genetic testing and HRD detection with CHORD can be performed simultaneously with the same assay.

Discussion

Here we describe a classifier (CHORD) that can detect HRD across cancer types as well as the HRD subtypes based on mutation profiles. By using this tool in a systematic pan-cancer analysis, we reveal novel insights into the mechanisms and incidence of HRD across cancer types with potentially important clinical relevance.

HRD targeted therapy with PARPi is mostly restricted to breast and ovarian cancer[5], although its use for treating pancreatic cancer was recently approved by the FDA (US Food and Drug Administration)[37]. Our results indicate, however, that HRD is present in a non-negligible proportion of other cancer types as well, suggesting that a large number of such patients who would potentially benefit from PARPi therapy currently remain unnoticed. Given that the underlying cause of HRD is ultimately unrelated to cancer type, CHORD would serve as a valuable tool for cancer type agnostic patient stratification for future PARPi trials. This is especially important for cancer types currently lacking good markers for patient stratification for such treatments such as prostate[38] and biliary[39] cancer.

We also demonstrated that genetic based detection of HRD commonly used in the clinic by germline mutation scanning of BRCA1/2 can miss a substantial number of HRD patients. HR gene inactivation often occurs purely by somatic events across all cancer types. Furthermore, germline testing is particularly unsuitable for prostate cancer where gene inactivation is frequently caused by somatic deep deletions, which prevent the identification of any SNVs/indels at the affected locus when using panel- or PCR-based sequencing methods (exon scanning). This problem also exists for other cancer types where deep deletions also make up a non-negligible fraction of HR gene inactivation cases. While somatic mutation testing improves diagnostic yield and is indeed increasingly performed in the clinic[5], WGS based genetic testing is ultimately necessary to capture the full spectrum of genetic alterations and to accurately determine the mutational status of HR genes. However, we show that even such broad genetic testing with focus on biallelic gene inactivation still potentially misses roughly half of all HRD patients. Although exon scanning for mono-allelic mutations could in principle be incorporated in routine procedures, this likely results in many passenger mutations being identified, increasing the VU interpretation challenge. Moreover, most mono-allelically mutated CHORD positive patients were driven by LOH events, which are typically missed by exon scanning approaches. Finally, when just relying on genetic information, one would preferably obtain solid evidence on the inactivation of the non-mutated allele to be conclusive in these cases. While the incorporation of additional procedures would enable this (e.g. bisulfite sequencing to detect methylation in the other allele), this may become impractical and expensive. As we show here, this challenge can also be overcome by detecting the consequences of HRD by tools like CHORD rather than the underlying genetic causes.

Current FDA approved mutational scar based HRD detection methods rely on the identification of large genomic alterations using SNP array data, including telomeric allelic imbalance (TAI), large-scale transitions (LST), and/or genome-wide LOH levels[5,40,41]. While CHORD also uses large genomic alterations including structural duplications and deletions for detecting HRD, small microhomology deletions were found to be most predictive of HRD, but these cannot be detected by SNP arrays. We thus expect improved HRD classification by CHORD and hence better prediction of patient PARPi response over the current approved methods, although this will need to be evaluated in direct comparisons and prospective clinical trials.

We envision that the findings from our analyses incentivizes improvements to current clinical practices for detecting HRD, and that the application of genomics-based approaches, like CHORD, in the clinic will support these endeavors and provide additional treatment options for patients. CHORD is freely available as an R package at https://github.com/UMCUGenetics/CHORD.

Acknowledgements

This publication and the underlying study have been made possible partly on the basis of the data that Hartwig Medical Foundation and the Center of Personalised Cancer Treatment (CPCT, The Netherlands) have made available to the study. We thank Neeltje Steeghs (Netherlands Cancer Institute), Martijn Lolkema (Erasmus Medical Center Rotterdam), Geert Cirkel (Meander Medical Center), Els Witteveen (UMC Utrecht), Mariette Labots (Amsterdam UMC, location VUmc) and Laurens Beerepoot (Elisabeth-TweeSteden Ziekenhuis, Tilburg) for study inclusion of a significant part of the patients that were used in this study and Peter Bouwman (Netherlands Cancer Institute, Amsterdam) for critically reading the manuscript. This work was financially supported by the gravitation program CancerGenomiCs.nl from the Netherlands Organisation for Scientific Research (NWO) and Oncode Institute to E.C.

Methods

Datasets

Germline and somatic VCF files of the 3,824 metastatic tumor samples from 3,504 patients were obtained from Hartwig Medical Foundation (HMF) https://www.hartwigmedicalfoundation.nl/en/appyling-for-data/. For patients with multiple biopsies that were taken at different timepoints, patient IDs were suffixed by 'A' for the first biopsy, 'B' for the second biopsy, etc (e.g. HMF001423A, HMF001423B). Somatic variant TSV files of the 560 breast cancer (BRCA-EU) dataset was downloaded from the International Cancer Genome Consortium (ICGC; https://dcc.icgc.org/).

Variant calling

Variant calling in the HMF dataset was performed previously by HMF (https://github.com/hartwigmedical/pipeline)[12]. Briefly, reads were mapped to GRCh37 using BWA-MEM v0.7.5a with duplicates being marked for filtering. Indels were realigned using GATK v3.4.46 IndelRealigner. GATK Haplotype Caller v3.4.46 was used for calling germline variants in the reference sample. For somatic SNV and indel variant calling, GATK BQSR3 was first used to recalibrate base qualities, followed by Strelka v1.0.14 for the variant calling itself. Somatic SV calling was performed using GRIDSS v1.8.0. Copy-number calling was performed using PURity & PLoidy Estimator (PURPLE), that combines B-allele frequency (BAF), read depth and structural variants to estimate the purity and copy number profile of a tumor sample[42].

Determining gene biallelic status

For tumors in the HMF cohort, biallelic status was determined for 781 genes (**Supplementary table 8**) which included genes associated with cancer, according to Cancer Genome Interpreter (https://www.cancergenomeinterpreter.org/genes), as well as a manually curated set of genes involved in HR (based on the KEGG HR pathway (https://www.genome.jp/), as well as via a literature search). This was performed using an in-house pipeline that interprets copy-number, and germline and somatic SNV/indel data from the HMF variant calling pipeline to determine biallelic gene status (https://github.com/luannnguyen/hmfGeneAnnotation).

First, the copy number status in the gene region was determined. If the minimum copy number was <0.3, the gene was considered to have a deep deletion (and by default biallelically inactivated). Else, the gene was screened for 2 mutation events, which included following combinations: (i) loss of heterozygosity (LOH) with a germline or somatic SNV/indel; (ii) a germline and somatic SNV/indel; or (iii) 2 somatic SNV/indels.

For SNV/indels, variant pathogenicity was assessed based on pathogenicity annotations from ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/; GRCh37, database date 2018-12-07), and for variants without an entry in ClinVar, based on variant type as determined by SnpEff (http://snpeff.sourceforge.net/; v4.1h). Briefly, variants can be given one of the following annotations from ClinVar: pathogenic, likely pathogenic, variant of unknown significance (VUS), likely benign, and benign. A pathogenicity score (P-score) of 1-5 was also assigned to each annotation, with 1=benign and 5=pathogenic. Additionally, variant types as determined by SnpEff were assigned similar annotations and scores: out-of-frame frameshifts were considered pathogenic (P-score=5); nonsense and splice variants were considered likely pathogenic (P-score=4); missense variants, essential splice variants, and inframe frameshifts were considered VUS's (P-score=3); the remaining variant types were considered likely benign or benign (P-score ≤2). The final P-score of a variant was the ClinVar P-score if a ClinVar annotation exists for that variant, and if not, the SnpEff P-score was used. See **Supplementary table 10** for details on pathogenicity scoring.

LOH was considered pathogenic and was automatically given a P-score of 5. LOH occurred if the 'minimum minor allele ploidy' (determined by PURPLE[42]) was < 0.2. Briefly, this value represents the minimum tumor purity adjusted ploidy of the minor allele (i.e. the non-reference allele of heterozygous SNVs observed in the control sample) within the gene region.

P-scores from pairs of mutation events (i.e. SNV, indel, or LOH) were summed to yield a biallelic pathogenicity score (BP-score), giving a maximum possible score of 10. Deep deletions were automatically given a score of 10. Per gene, the biallelic event with the highest score was taken the biallelic status of the gene. If multiple events had the same score, a biallelic event was greedily selected.

Extracting mutation contexts

The absolute frequency of 3 types of mutation contexts (SNV, indel, and structural variant (SV) contexts) was determined from the somatic variant data from the HMF and BRCA-EU cohorts. This was performed using an R package that we have developed, *mutSigExtractor* (https://github.com/UMCUGenetics/mutSigExtractor).

The SNV contexts comprised of 96 trinucleotide contexts, which are composed of one of six classes of base substitutions (C>A, C>G, C>T, T>A, T>C, T>G) in combination with the immediate 5' and 3' flanking nucleotides.

The 6 indel contexts were extracted based on the presence of: short tandem repeats; short stretches of identical sequence at the breakpoints, also known as microhomology; or the absence of either. Indels in repeat regions were defined as the presence of ≥1 copy of the indel sequence downstream (i.e. in the 3' direction) from the breakpoint, where sequence length must be <50bp. Indels with flanking microhomology were defined as the presence of the following sequence features up or downstream from the breakpoint: (i) ≥1 copy of the indel sequence where the indel sequence length is ≥50bp; (ii) ≥2bp sequence identity to the indel sequence; (iii) ≥1bp sequence identity if the indel sequence length is ≥3bp. For (ii) and (iii) the number of up or downstream bases searched was equal to the length of the indel.

The 16 SV contexts were composed of the SV type (deletion, duplication, inversion, translocation) and the SV length (1-10kb, 10-100kb, 100kb-1Mb, 1Mb-10Mb, >10Mb). Note that SV length is not applicable for translocations.

Random forest training

Features

To construct the features for training the **C**lassifier of **HO**mologous **R**ecombination **D**eficiency (CHORD), the 96 trinucleotide contexts were first simplified to 6 base substitution contexts by discarding the 5' and 3' flanking nucleotide information. Then, relative contribution was calculated for each feature per mutation context type (i.e. SNV, indel, SV contexts separately). For training the mutational signature model, the 96 trinucleotide contexts were fitted to the 30 COSMIC SNV signatures[43,44] using the non-negative least squares algorithm (incorporated in *mutSigExtractor*). Similarly, the SV contexts were fitted to the 6 SV signatures as reported previously[8]. The relative frequency of the SNV signatures, SV signatures, and indel contexts was then calculated per mutation type.

Training set

The training set consisted of samples which we could confidently consider BRCA1/2 deficient or proficient based on the P-scores/BP-scores as described in *Determining gene biallelic status* and **Supplementary table 10**. BRCA1/2 deficiency was defined as having: (i) a deep deletion or (ii) LOH in combination with a pathogenic SNV/indel or an out-of-frame frameshift (BP-score = 10). Within the BRCA1/2 deficient group, samples where the absolute frequency of indels within repeat regions was >14000 were considered to have microsatellite instability (MSI) and were removed. This threshold was determined by correlating the frequency of indels within repeat regions for a selection of samples to a 5-gene PCR panel for detecting MSI (BAT25, BAT26, NR21, NR24 and MONO27 markers; data not shown). This filtering step was done as the relative contribution of indels in repeat regions are grossly overrepresented in samples with MSI, thereby masking the contribution of microhomology deletions. This sample group ultimately consisted of 35 BRCA1 ('BRCA1' class) and 90 BRCA2 ('BRCA2' class) deficient samples which were both considered HRD during the training. Conversely, BRCA1/2 proficiency required the following criteria: (i) Absence of deep deletions or LOH; (ii) all SNV/indels had a P-score ≤ 3 (VUS or lower in impact); (iii) for the highest impact pair of SNV/indels (i.e. highest BP-score), the germline variant had a P-score ≤ 3 (VUS or lower in impact) and the somatic variant a P-score ≤ 2 (likely benign or lower in impact). This BRCA proficient group ('none' class) consisted of 1358 samples which were considered HRP during the training (**Supplementary figure 1**).

Training procedure

The training procedure for both CHORD and the mutational signature model was identical and is illustrated in **Supplementary figure 2**. A core training procedure, which performs feature selection and class resampling, forms the basis for the full training procedure (**Supplementary figure 2a**). Feature selection was done to retain features which had a higher mean relative frequency in BRCA1/2 deficient samples compared to BRCA1/2 proficient samples, where the difference between the means must be significant ($p < 0.01$) as determined by a t-test. Class resampling serves to reduce the difference in the number of samples between each class (i.e. class imbalances). Here, a grid search was performed to determine the optimal pair of the following parameters: (i) down-sampling of the 'none' class: no down-sampling, 2x or 4x; (ii) up-sampling of the 'BRCA1' class: no up-sampling, 1.5x or 2x. For each iteration of the grid search, a random forest-based 10x repeated 10-fold cross-validation (CV) was performed, after which the area under the precision-recall curve (AUC-PR) was calculated. The parameter pair with the highest AUC-PR was chosen. With the selected features and resampling parameters, a random forest was then trained that predicts the probability of a new sample being in one of the aforementioned 3 classes (i.e. 'BRCA1', 'BRCA2' or 'none'). We defined the HRD probability as the sum of the probability of belonging to the 'BRCA1' and 'BRCA2' classes, where a sample was considered HRD if the HRD probability was ≥ 0.5. Random forests were trained using the *randomForest* R package.

The full training procedure was split into 2 stages (**Supplementary figure 2b**). The first stage serves to filter 'BRCA1' or 'BRCA2' samples from the which are likely not HRD (e.g. due to reversal of biallelic inactivation via a second frameshift bringing the gene in frame), or 'none' samples which are likely not HRP (e.g. due to deficiencies in other HR genes). Here, the core training procedure is encapsulated by a 10-fold CV loop to allow every sample to be excluded from the training set to subsequently calculate an unbiased HRD probability. This was repeated 100 times and the number of times each sample was HRD or HRP was calculated. 'BRCA1' or 'BRCA2' samples that were predicted HRD < 60 times were blacklisted while 'none' samples that were predicted HRD > 40 times were blacklisted. In the second training stage, the core training procedure was performed on a training set without the blacklisted samples. This yielded the final random forest model (either CHORD or the mutational signature model).

The performance of the final random forest model was assessed using 2 approaches: (i) 10-fold CV of the training set by further encapsulating the full training procedure in a 10-fold CV loop; (ii) applying the final random forest model to an external dataset (BRCA-EU dataset). An AUC-PR was then calculated for both approaches. In the case of the BRCA-EU dataset, BRCA1/2 deficiency annotations were obtained from Davies *et al.* 2017[9].

Code availability

CHORD is available as an R package at https://github.com/UMCUGenetics/CHORD.

Determining the genetic cause of HRD

To determine the genetic cause of HRD, tumors were first selected from the HMF cohort based on the absence of MSI and having ≥50 indels and ≥30 SVs (**Supplementary table 9**). Furthermore, a single tumor per patient was selected (based on highest tumor purity) for those with multiple biopsies . In total, 3504 tumors were selected (from the 3824 in total) to represent each patient. The following procedure was then employed for identifying biallelic loss in each of the 781 genes. First, high frequency germline SNV/indels (present in >1.6% of patients; **Supplementary figure 9**) were marked as benign (P-score = 0). Then, each gene was screened for the following events: (i) a deep deletion; (ii) LOH in combination with a germline SNV/indel with a P-score ≥ 4 (likely pathogenic or higher in impact); (iii) LOH in combination with a somatic SNV/indel with a P-score ≥ 3 (VUS or higher in impact); or (iv) two SNVs/indels (germline + somatic, or 2x somatic) both with a P-score = 5 (pathogenic). See **Supplementary table 10** for details of the P-score thresholds used.

After applying CHORD to the HMF cohort, we then determined whether each of the 781 genes was significantly more frequently deficient in CHORD-HRD vs. CHORD-HRP patients using the Fisher's exact test, with multiple testing correction performed using the Benjamini-Hochberg procedure. This was ultimately to determine the genes most likely to cause HRD when inactivated. Six genes were found with a q-value < 0.1: BRCA1, BRCA2, RAD51C, PALB2, NF1, and STARD13 (**Supplementary figure 10**). NF1 and STARD13 have not been reported with be involved in HR, and thus further analyses were performed to validate the enrichment for these 2 genes.

Since BRCA1 and NF1 are both located on Chr17, we reasoned that copy number alterations (CNA; in this case referring to deep deletions or LOH) that affect BRCA1 also affect NF1. This leads to frequent biallelic loss of NF1 even though the gene is likely not associated with HRD. A similar situation was suspected for BRCA2 and STARD13 which are both located on Chr13. Thus, Fisher exact tests were performed to determine whether CNAs in each of the 781 genes significantly co-occurred with a CNA in BRCA1 or BRCA2. Multiple testing correction was performed using the Benjamini-Hochberg procedure. Indeed, enrichment in the co-occurrence of BRCA1 and NF1 CNAs was found, and was similarly the case for BRCA2 and STARD13 (**Supplementary figure 11**). We thus concluded that biallelic loss of NF1 and STARD13 are likely not associated with HRD and were therefore excluded from **Figure 3a**.

For each of the 211 CHORD-HRD patients, the gene with biallelic event with the highest BP-score was considered the gene that caused HRD, where the pair of P-scores constituting the BP-score had to be 5 and ≥3. Patients with no biallelic events fulfilling these criteria were considered to have an unknown cause of HRD.

References

1. Lord, C. J. & Ashworth, A. BRCAness revisited. *Nat. Rev. Cancer* **16**, 110–120 (2016).
2. Heeke, A., Lynce, F., Baker, T., Pishvaian, M. & Isaacs, C. Prevalence of Homologous Recombination Deficiency (HRD) Among All Tumor Types. (2017).
3. Audeh, M. W. *et al.* Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial. *The Lancet* **376**, 245–251 (2010).
4. Mateo, J. *et al.* DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. *http://dx.doi.org/10.1056/NEJMoa1506859* (2015) doi:10.1056/NEJMoa1506859.
5. Hoppe, M. M., Sundar, R., Tan, D. S. P. & Jeyasekharan, A. D. Biomarkers for Homologous Recombination Deficiency in Cancer. *JNCI J. Natl. Cancer Inst.* **110**, 704–713 (2018).
6. Van Hoeck, A., Tjoonk, N. H., van Boxtel, R. & Cuppen, E. Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer* **19**, 457 (2019).
7. Sfeir, A. & Symington, L. S. Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? *Trends Biochem. Sci.* **40**, 701–714 (2015).
8. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
9. Davies, H. *et al.* HRDetect is a predictor of *BRCA1* and *BRCA2* deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).
10. Patel, J. N. *et al.* Characterisation of homologous recombination deficiency in paired primary and recurrent high-grade serous ovarian cancer. *Br. J. Cancer* **119**, 1060–1066 (2018).
11. Sztupinszki, Z. *et al.* Genomic aberration based molecular signatures efficiently characterize homologous recombination deficiency in prostate cancer. *bioRxiv* 681114 (2019) doi:10.1101/681114.
12. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* 1–7 (2019) doi:10.1038/s41586-019-1689-y.
13. Ceccaldi, R. *et al.* Homologous-recombination-deficient tumours are dependent on Polθ-mediated repair. *Nature* **518**, 258–262 (2015).
14. Truong, L. N. *et al.* Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. *Proc. Natl. Acad. Sci.* **110**, 7720–7725 (2013).
15. Willis, N. A. *et al.* Mechanism of tandem duplication formation in *BRCA1*-mutant cells. *Nature* **551**, 590–595 (2017).
16. Polak, P. *et al.* A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **49**, 1476–1486 (2017).
17. Jonsson, P. *et al.* Tumour lineage shapes BRCA-mediated phenotypes. *Nature* **571**, 576–579 (2019).
18. Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M., Cortés-Ciriano, I. & Park, P. J. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat. Genet.* **51**, 912–919 (2019).
19. Maura, F. *et al.* A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* **10**, 1–12 (2019).
20. Nones, K. *et al.* Whole-genome sequencing reveals clinically relevant insights into the aetiology of familial breast cancers. *Ann. Oncol.* **30**, 1071–1079 (2019).
21. Staaf, J. *et al.* Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.* **25**, 1526–1533 (2019).
22. Chun, J., Buechelmaier, E. S. & Powell, S. N. Rad51 Paralog Complexes BCDX2 and CX3 Act at Different Stages in the BRCA1-BRCA2-Dependent Homologous Recombination Pathway. *Mol. Cell. Biol.* **33**, 387–395 (2013).
23. Castillo, A. *et al.* The BRCA1-Interacting Protein Abraxas Is Required for Genomic Stability and

Tumor Suppression. *Cell Rep.* **8**, 807–817 (2014).
24. Zhao, W. *et al.* BRCA1–BARD1 promotes RAD51-mediated homologous DNA pairing. *Nature* **550**, 360–365 (2017).
25. Cantor, S. B. & Guillemette, S. Hereditary breast cancer and the BRCA1-associated FANCJ/BACH1/BRIP1. *Future Oncol. Lond. Engl.* **7**, 253–261 (2011).
26. Folias, A. *et al.* BRCA1 interacts directly with the Fanconi anemia protein FANCA. *Hum. Mol. Genet.* **11**, 2591–2597 (2002).
27. Yu, X. & Chen, J. DNA damage-induced cell cycle checkpoint control requires CtIP, a phosphorylation-dependent binding partner of BRCA1 C-terminal domains. *Mol. Cell. Biol.* **24**, 9478–9486 (2004).
28. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019) doi:10.1101/531210.
29. Sim, N.-L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457 (2012).
30. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
31. Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell* **174**, 758-769.e9 (2018).
32. Kondrashova, O. *et al.* Secondary Somatic Mutations Restoring RAD51C and RAD51D Associated with Acquired Resistance to the PARP Inhibitor Rucaparib in High-Grade Ovarian Carcinoma. *Cancer Discov.* **7**, 984–998 (2017).
33. Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282-1294.e20 (2019).
34. Sakai, W. *et al.* Secondary mutations as a mechanism of cisplatin resistance in BRCA2 -mutated cancers. *Nature* **451**, 1116–1120 (2008).
35. Edwards, S. L. *et al.* Resistance to therapy caused by intragenic deletion in BRCA2. *Nature* **451**, 1111–1115 (2008).
36. Nangalia, J. & Campbell, P. J. Genome Sequencing during a Patient's Journey through Cancer. *N. Engl. J. Med.* **381**, 2145–2156 (2019).
37. Golan, T. *et al.* Maintenance Olaparib for Germline BRCA-Mutated Metastatic Pancreatic Cancer. *N. Engl. J. Med.* **381**, 317–327 (2019).
38. Arora, K. & Barbieri, C. E. Molecular Subtypes of Prostate Cancer. *Curr. Oncol. Rep.* **20**, 58 (2018).
39. Tariq, N.-A., McNamara, M. & Valle, J. Biliary tract cancers: current knowledge, clinical candidates and future challenges. *Cancer Manag. Res.* **11**, 2623–2642 (2019).
40. Myriad myChoice CDx - P190014. *FDA* http://www.fda.gov/medical-devices/recently-approved-devices/myriad-mychoice-cdx-p190014 (2019).
41. FoundationFocus CDxBRCA LOH - P160018/S001. *FDA* http://www.fda.gov/medical-devices/recently-approved-devices/foundationfocus-cdxbrca-loh-p160018s001 (2018).
42. Cameron, D. L. *et al.* GRIDSS, PURPLE, LINX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number. *bioRxiv* 781013 (2019) doi:10.1101/781013.
43. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep.* **3**, 246–259 (2013).
44. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149**, 979–993 (2012).

Tables

Cancer type	(A) HRD patients with BRCA1/2 promoter methylation (Davies et al.)		(B) CHORD-HRD patients without biallelic loss of HRD associated genes	
	%	# / total	%	# / total
Ovarian	35%	16/46	50%	21/42
Breast	24%	21/86	49%	38/77

Table 1: **(A)** The number of ovarian and breast cancer HRD patients with BRCA1/2 promoter methylation as reported by Davies et al.[9] (from supplementary tables 1, 3, and 8) in comparison to **(B)** the number of CHORD-HRD ovarian and breast cancer patients without biallelic loss of the HRD associated genes (BRCA1/2, RAD51C, PALB2). In **(A)**, only breast cancer samples which had no missing data (marked as 'is used for evaluation' in supplementary table 1) were considered.

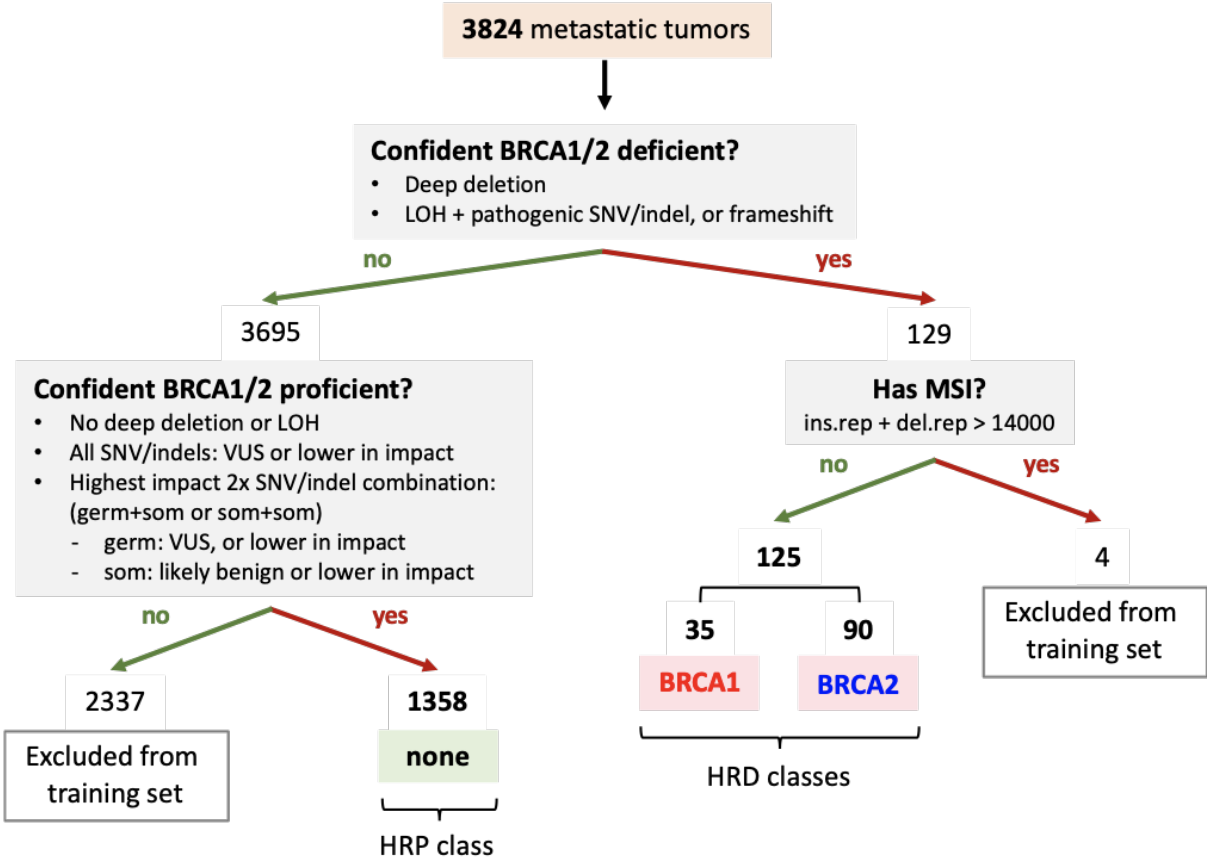
Supplementary tables

Supplementary table 2: Predictions and cross-validation output from all random forest models
Supplementary table 3: Features from HMF and BRCA-EU datasets

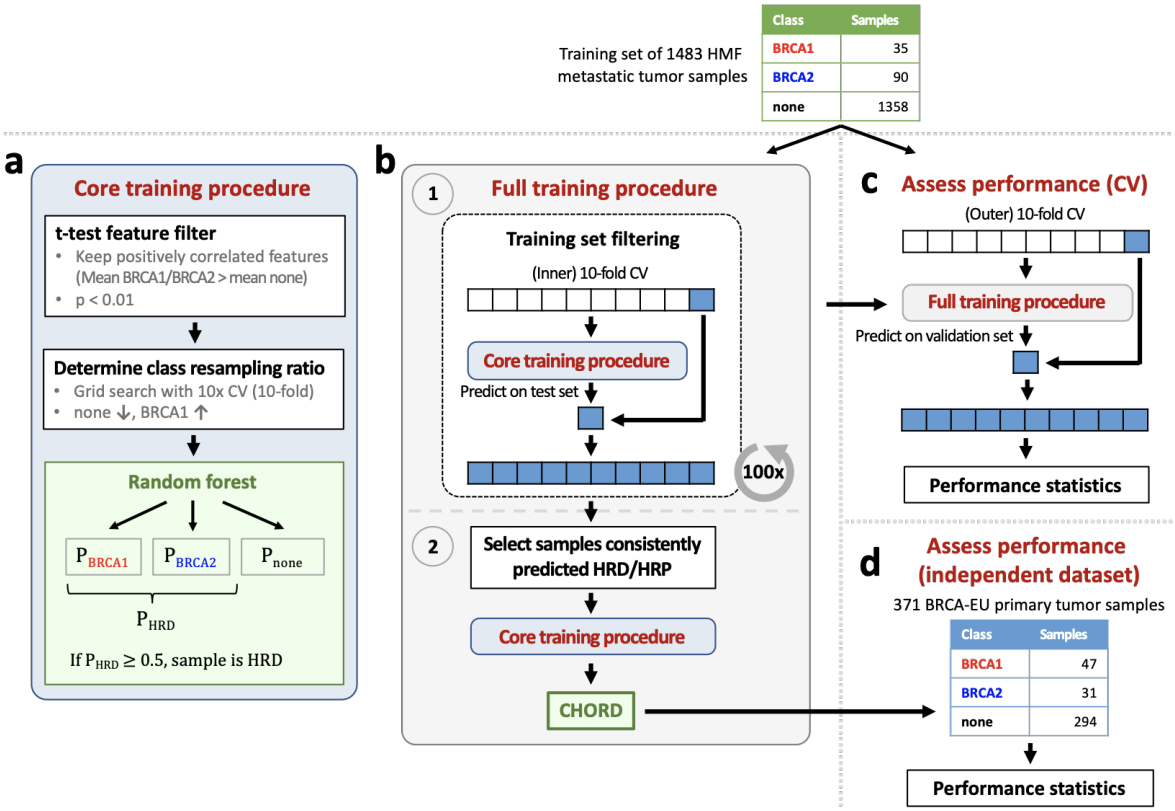
Supplementary table 4: BRCA1/2, RAD51C and PALB2 genotypes for HRD tumors
Supplementary table 5: HRP tumors with biallelic loss of BRCA1/2, RAD51C and PALB2
Supplementary table 6: Likely pathogenic VUS's
Supplementary table 7: Incidence and genetic cause of HRD by cancer type

Supplementary table 8: List of 781 cancer and HR related genes
Supplementary table 9: Sample selection for pan-cancer analysis of HRD
Supplementary table 10: Pathogenicity scoring

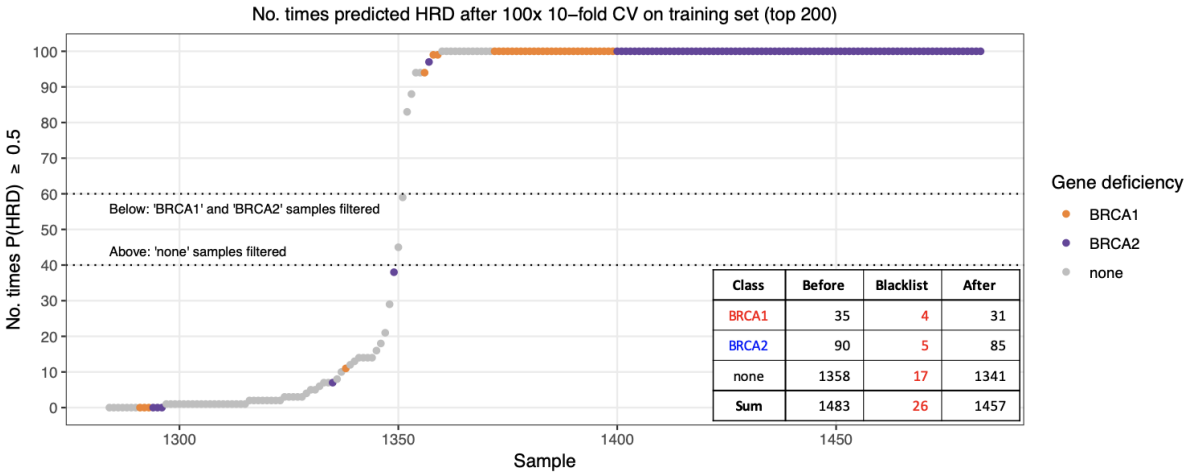
Supplementary figures



Supplementary figure 1: Sample selection for training CHORD. Only samples where biallelic inactivation of BRCA1/2 could confidently be identified (and did not have MSI), or where disruption of either allele of BRCA1/2 was not apparent were recruited into the training set. A total of 1,483 samples were eventually selected, consisting of 35 BRCA1 and 90 BRCA2 deficient which were both considered HRD during training, and 1,358 BRCA1/2 proficient ('none') which were considered HRP.

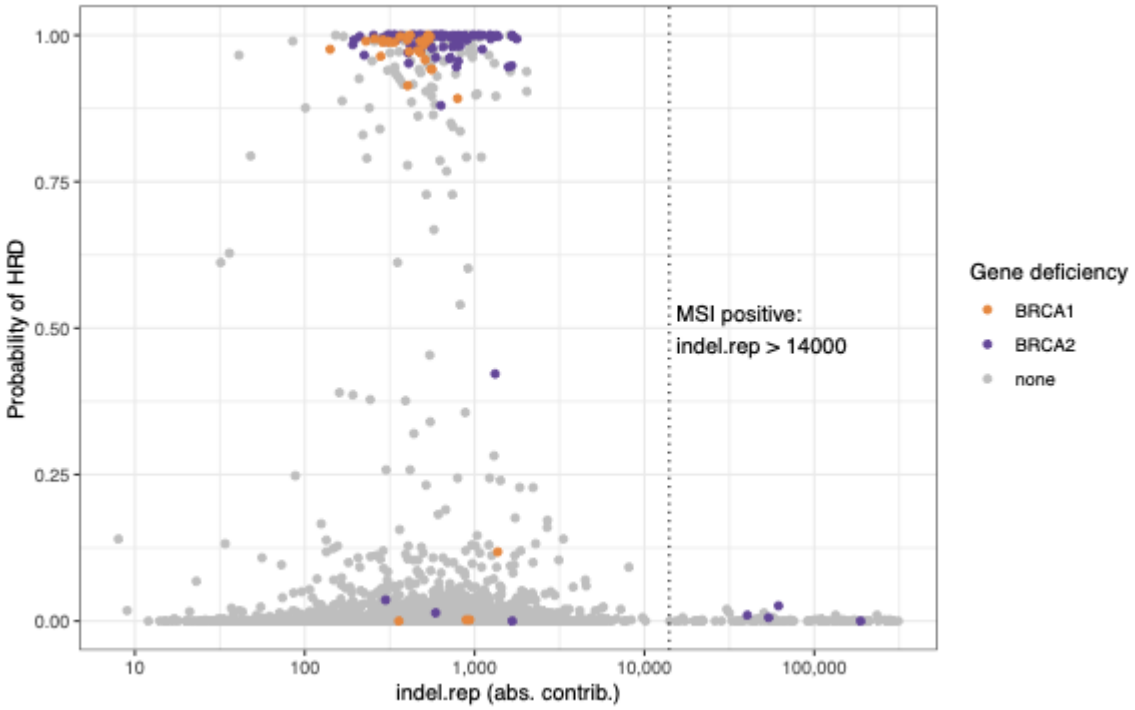


Supplementary figure 2: Workflow for training CHORD. 'BRCA1', 'BRCA2' and 'none' classes refer to BRCA1 deficient, BRCA2 deficient and BRCA1/2 proficient sample groups, respectively. (**a**) The core training procedure on which the full training procedure is based on performs feature selection and class resampling. This returns a random forest that outputs the probability of a new sample being in one of the aforementioned 3 classes. The probability of HRD (P_{HRD}) is the sum of the probability of belonging to the 'BRCA1' and 'BRCA2' classes, where a sample is considered HRD if P_{HRD} is ≥ 0.5. (**b**) The full training procedure is split into 2 stages. The first stage serves to blacklist 'BRCA1' or 'BRCA2' class samples which are likely not HRD (e.g. due to reversal of biallelic inactivation via a secondary frameshift), or 'none' samples which are likely not HRP (e.g. due to deficiencies in other HR genes). Here, the core training procedure is encapsulated by a 10-fold cross-validation (CV) loop to allow every sample to be excluded from the training set to subsequently calculate an unbiased P_{HRD}. This was repeated 100 times and the number of times each sample was HRD or HRP was calculated. 'BRCA1' or 'BRCA2' samples that were predicted HRD < 60 times were blacklisted. 'none' samples that were predicted HRD > 40 times were blacklisted. In the second training stage, the core training procedure was performed on a training set without the blacklisted samples. This produced the final random forest model, CHORD. (**c**) The full training procedure was further encapsulated by a 10-fold CV to assess the performance of CHORD. (**d**) Performance was also assessed by applying CHORD on an independent dataset containing 371 primary breast tumors (BRCA-EU dataset).

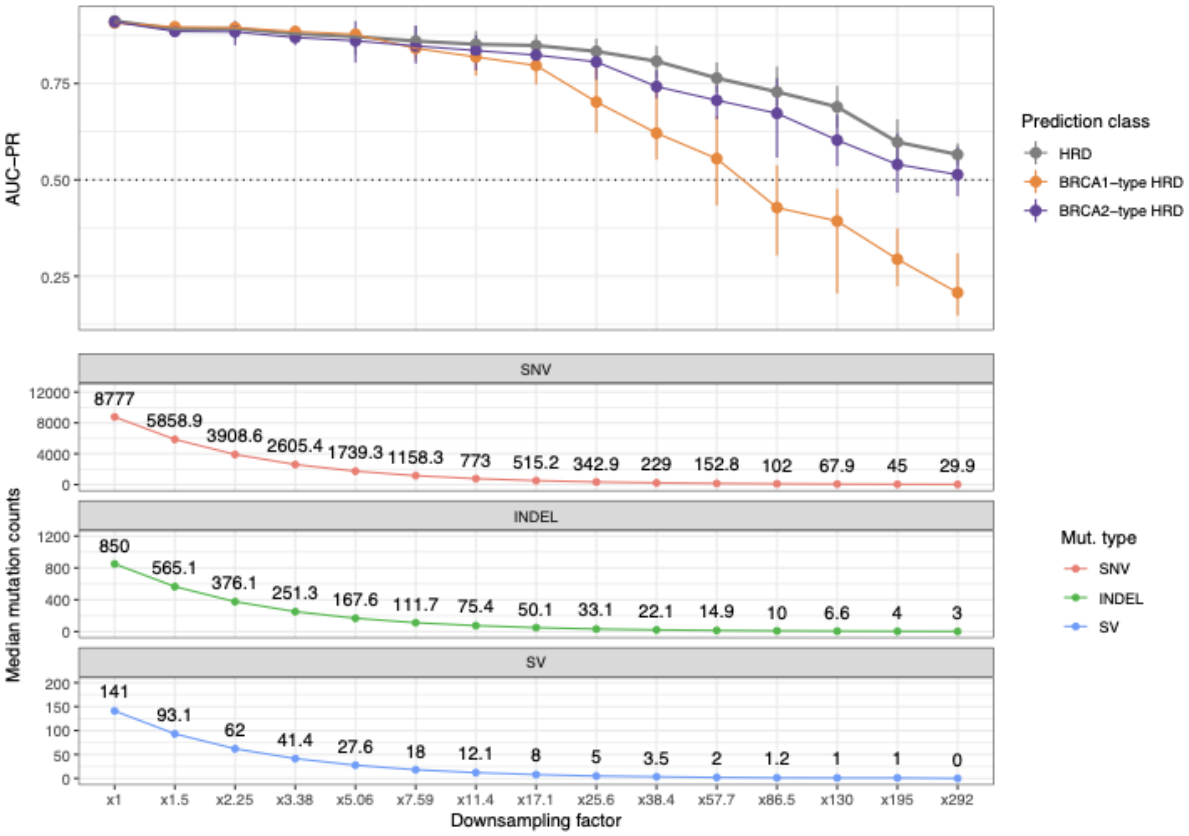


Class	Before	Blacklist	After
BRCA1	35	4	31
BRCA2	90	5	85
none	1358	17	1341
Sum	1483	26	1457

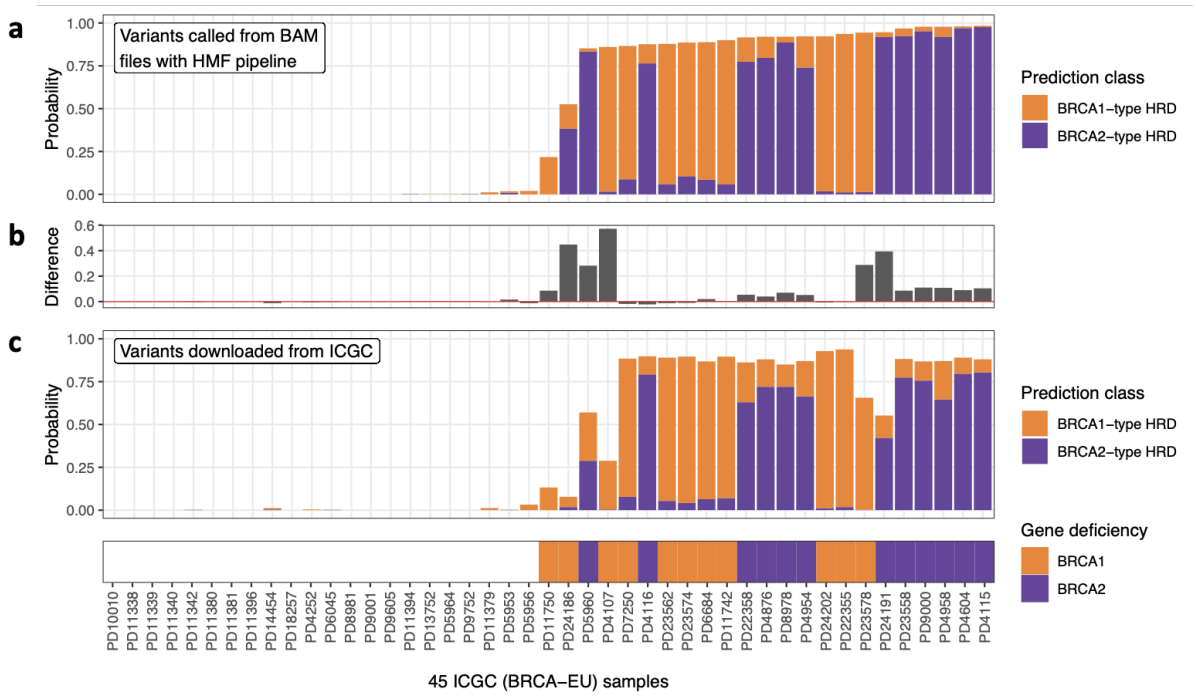
Supplementary figure 3: Results of the 100x repeated 10-fold CV procedure performed when training the CHORD (see **Supplementary figure 2b** for details) in order to blacklist 'BRCA1' or 'BRCA2' class samples that were consistently predicted HRP (no. times HRD < 60), or 'none' class samples that were consistently predicted HRD (no. times HRD > 40). Samples with a probability of HRD ≥ 0.5 were considered HRD. The overlaid table summarizes the number of samples per class before and after removing the blacklisted samples from the training set. Note that only the top 200 samples are shown here as the remaining samples belonged to the 'none' class and were always predicted HRP.



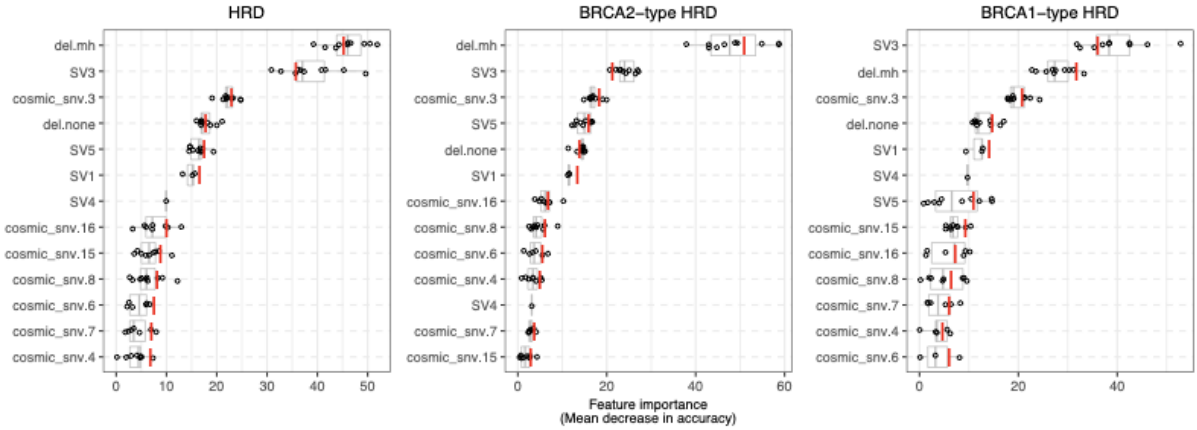
Supplementary figure 4: Microsatellite instability (MSI) impacts HRD prediction by CHORD. After applying CHORD to all 3824 tumors of the HMF dataset, all MSI positive tumors (those with more than 14,000 indels within repeat regions) had a low HRD probability, including 4 with BRCA2 biallelic loss.



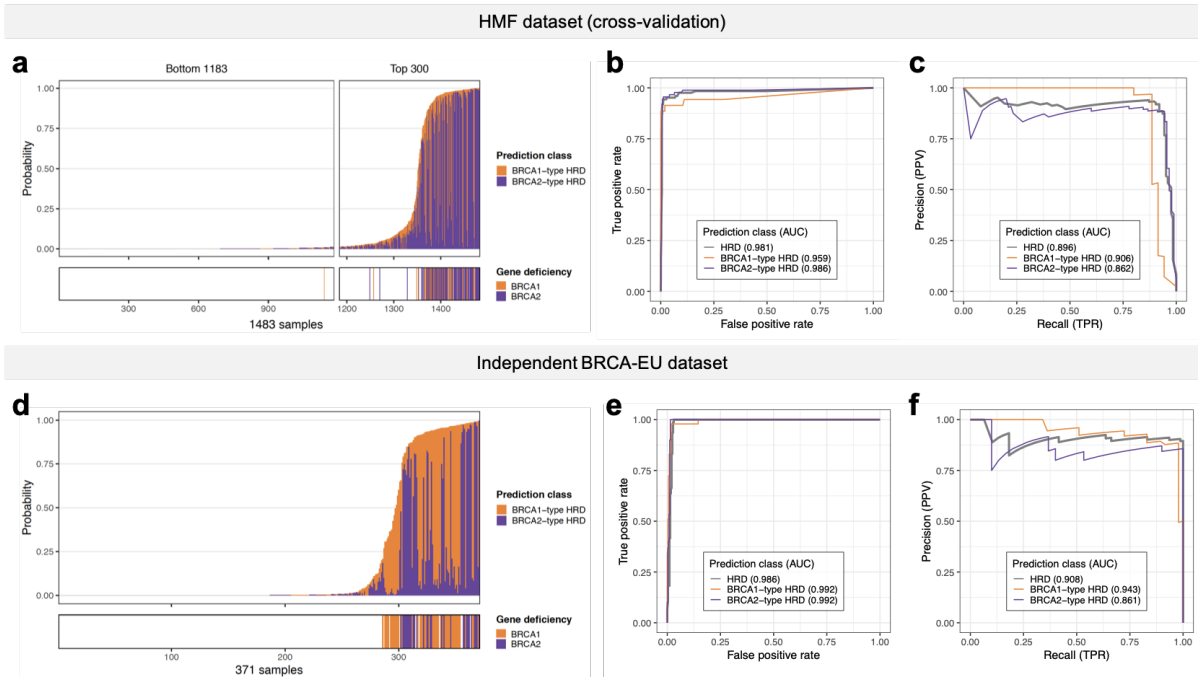
Supplementary figure 5: Performance of predicting HRD declines below ~50 indels. Similarly, performance for distinguishing BRCA1-type from BRCA2-type HRD declines below ~30 SVs. The set of mutations in each sample in the training set (used for training CHORD) were progressively downsampled. At each stage, CHORD was applied to the downsampled set of mutations, and thereafter, the area under the precision-recall curve was calculated.



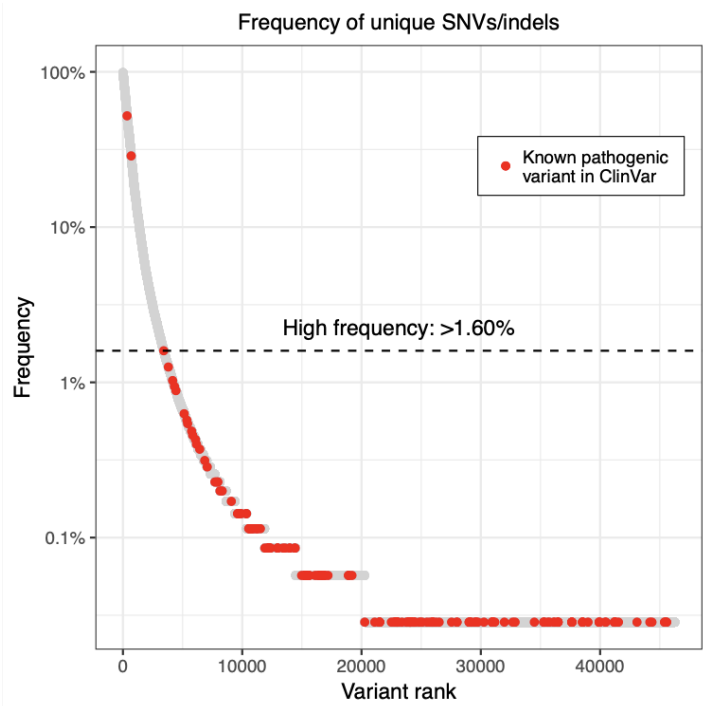
Supplementary figure 6: Variant calling pipeline differences affects CHORD performance. (**a**) Using variants called with the native pipeline of the HMF dataset (HMF pipeline) for HRD prediction with CHORD resulted in overall higher HRD probabilities in BRCA1/2 deficient tumors when compared to (**c**) using variants downloaded from ICGC. The differences in HRD probabilities are quantified in (**b**).



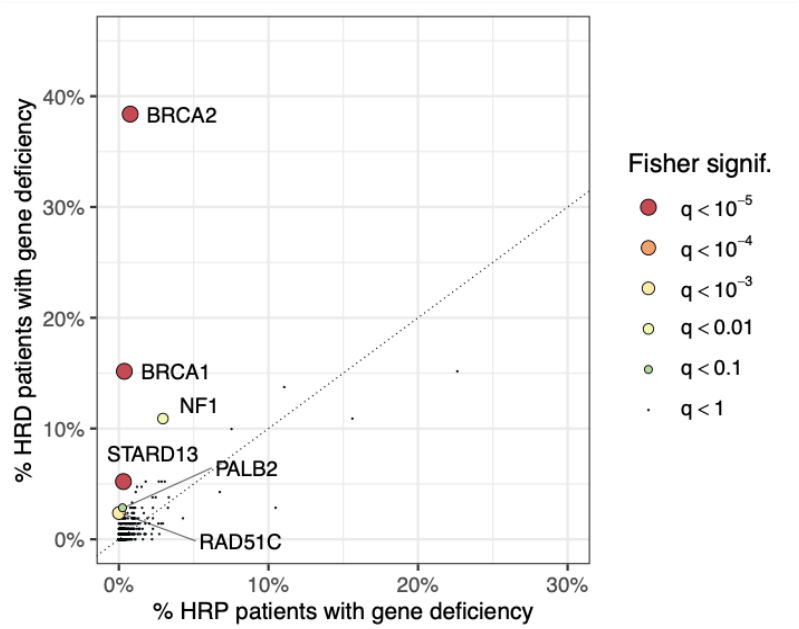
Supplementary figure 7: The features used by CHORD-signature to predict HRD and BRCA1/2-type HRD, with their importance indicated by mean decrease in accuracy. Boxplot and dots: Feature importance over 10-fold nested CV on the training set. Red line: Feature importance in the final model.



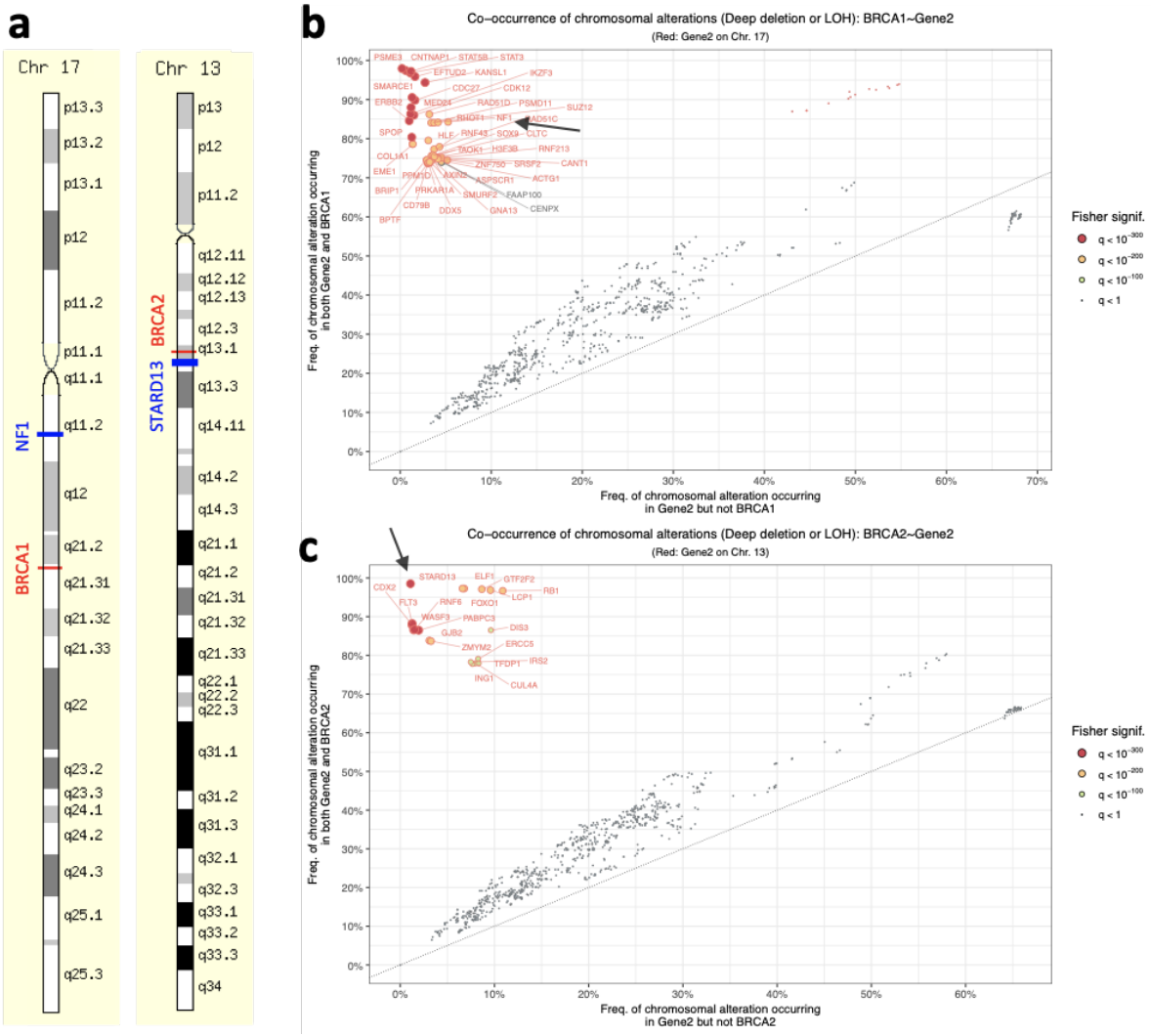
Supplementary figure 8: Performance of CHORD-signature by 10-fold cross-validation (CV) on the HMF training data or prediction on an independent dataset (BRCA-EU; primary breast cancer dataset). (**a, d**) The probability of HRD for each sample (total bar height) with each bar being divided into orange and purple bar segments indicating the probability of BRCA1- and BRCA2-type HRD, respectively. Below the bar plot, lines indicate detected biallelic loss of BRCA1 or BRCA2. In the case of (**a**), probabilities have been aggregated from the 10 CV folds. (**b, e**) Receiver operating characteristic (ROC) and (**c, f**) precision-recall curves (PR) showing the performance of CHORD when predicting HRD as a whole (grey), BRCA1-type HRD (orange), or BRCA2-type HRD (purple).



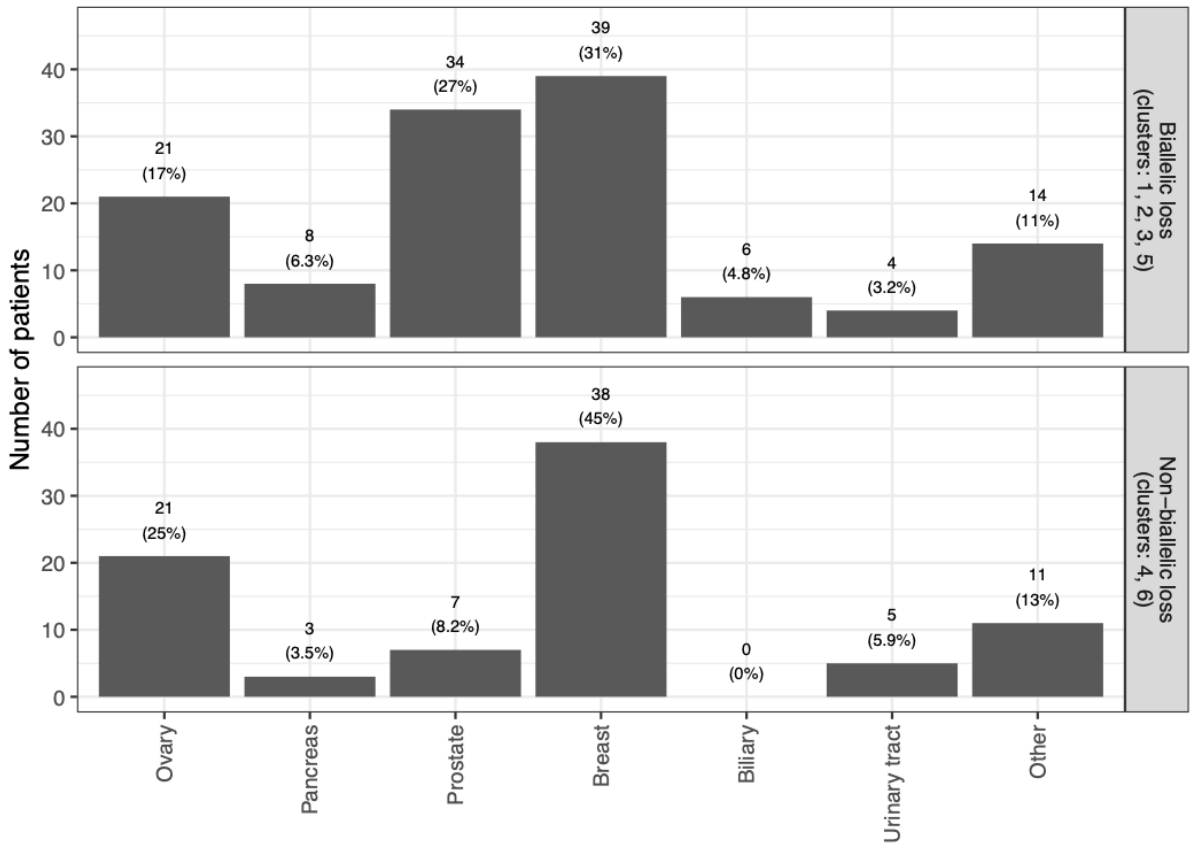
Supplementary figure 9: Frequency of unique germline SNVs/indels of 781 cancer related genes in patients of the HMF cohort. Germline variants with a frequency >1.6% were marked as benign prior to performing the pan-cancer analysis of HRD. Two known pathogenic variants with a frequency >1.56% in FGFR4 (c.1162G>A; 52.8%) and PTPRJ (c.827A>C; 28.7%) were also considered benign.



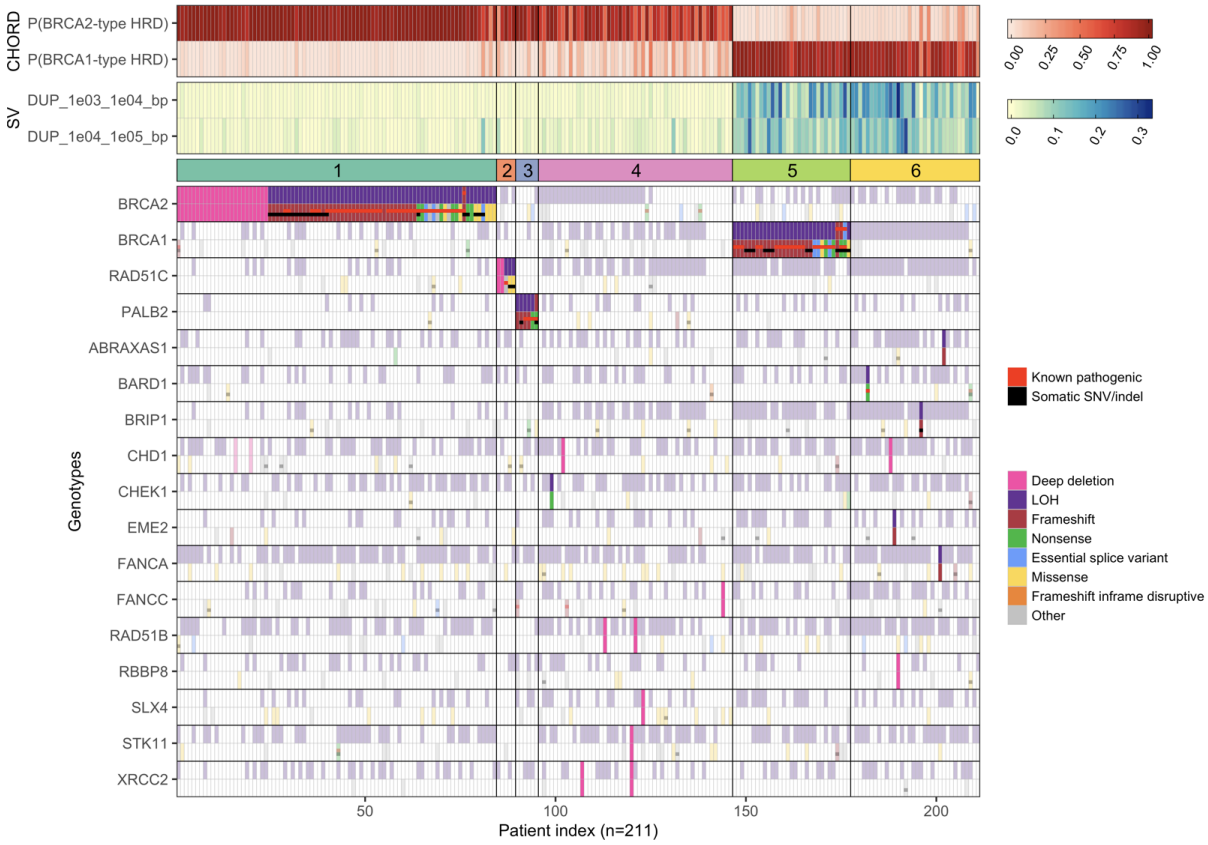
Supplementary figure 10: Same as **Figure 3b** except with NF1 and STARD13 included. A Fisher's exact test determined BRCA1, BRCA2, RAD51C, PALB2, NF1 and STARD13 to be significantly enriched (from a list of 781 cancer related genes) in CHORD-HRD vs. CHORD-HRP patient groups. Each point represents a gene with its size/color corresponding to the statistical significance as determined by the Fisher's exact test, with axes indicating the percentage of patients (within either the CHORD-HRD or CHORD-HRP group) in which biallelic inactivation was detected. Multiple testing correction was performed using the Benjamini-Hochberg procedure.



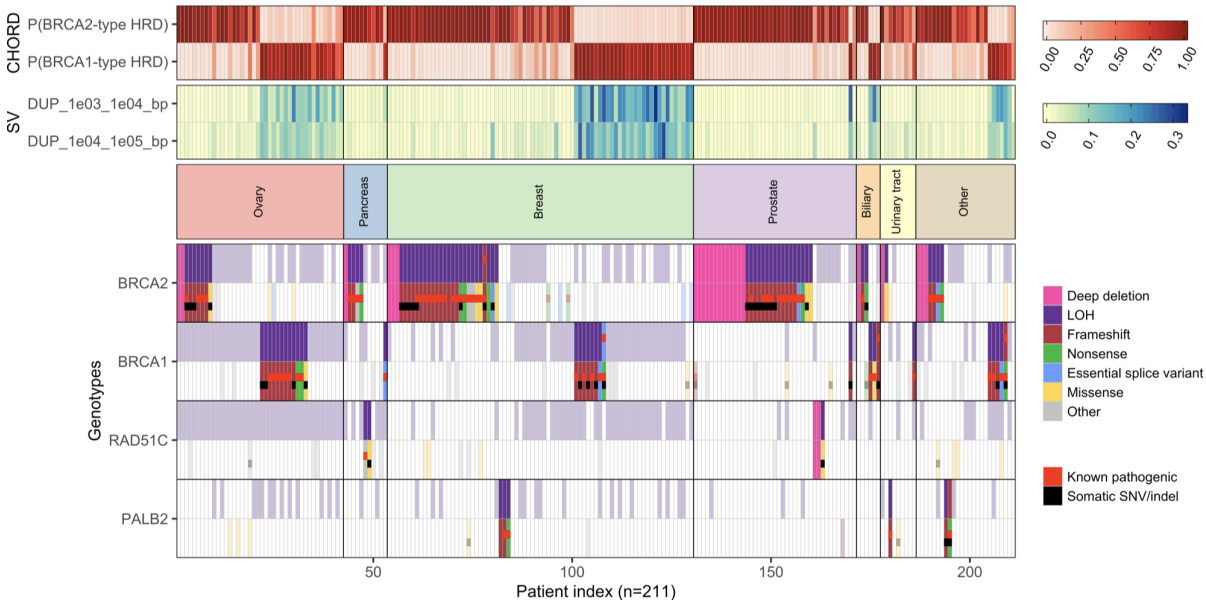
Supplementary figure 11: Chromosomal alterations (deep deletions or loss of heterozygosity (LOH)) affecting BRCA1 and BRCA2 also affects nearby genes including NF1 and STARD13 respectively. (**a**) BRCA1 and NF1 are both located on Chr17q while BRCA2 and STARD13 are both located on Chr13q. Source: www.genecards.org. (**b**) Enrichment of NF1 biallelic loss as shown in **Supplementary figure 10** is likely due to the gene being within proximity of BRCA1 and not because the gene is associated with HRD, since NF1 is not considered to be involved in HR in literature. The size/color of each point on the plot represents the significance of enrichment of CNAs occurring both in BRCA1 and in each of the 781 genes (one vs. all comparison). This was determined by the Fisher's exact test, where multiple testing correction was performed using the Benjamini-Hochberg procedure. Genes residing on the same chromosome as BRCA1 were marked with a red outline and text. (**c**) Similarly, as with (**a**), a chromosomal alteration affecting BRCA2 also affects the nearby gene STARD13. Enrichment of STARD13 biallelic loss as shown in **Supplementary figure 10** is likely due to the gene being within proximity of BRCA2 and not because the gene is associated with HRD.



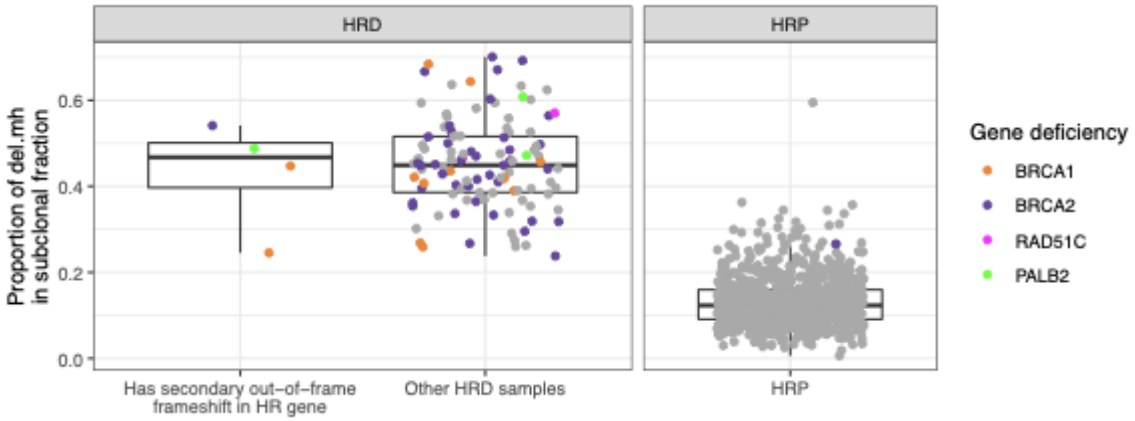
Supplementary figure 12: The number of CHORD-HRD patients of each cancer type which did (top) and did not (bottom) have biallelic loss of BRCA1, BRCA2, RAD51C, or PALB2. The top panel corresponds to patients in cluster 1,2,3 and 5 of **Figure 3c** while the bottom panel corresponds to patients in cluster 4 and 6. The percentages shown are within-group proportions.



Supplementary figure 13: Extended **Figure 3c** where biallelic status of BRCA2, BRCA1, RAD51C, PALB2, as well as other HR genes are shown for HMF cohort patients classified by CHORD as HRD. Top: BRCA1- and BRCA2-type HRD probabilities from CHORD. Middle: SV contexts used in CHORD to distinguish BRCA1- from BRCA2-type HRD. Bottom: The biallelic status of each gene. Tiles marked as 'Known pathogenic' refer to variants having a 'pathogenic' or 'likely pathogenic' annotation in ClinVar. Only HR genes with one of the following events in at least one patient from cluster 4 or 6 was shown here: deep deletion; or LOH in combination with a pathogenic/likely pathogenic variant or a frameshift/nonsense variant.



Supplementary figure 14: Biallelic status of BRCA2, BRCA1, RAD51C and PALB2 in HMF cohort patients classified by CHORD as HRD. This figure is the same as **Figure 3c** except that samples have been clustered both by cancer type and by HRD type. Top: BRCA1- and BRCA2-type HRD probabilities from CHORD. Middle: SV contexts used in CHORD to distinguish BRCA1- from BRCA2-type HRD. Bottom: The biallelic status of each gene. Tiles marked as 'Known pathogenic' refer to variants having a 'pathogenic' or 'likely pathogenic' annotation in ClinVar.



Supplementary figure 15: The subclonal proportion of microhomology deletions for 4 CHORD-HRD patients with a secondary frameshift (that reverts the out-of-frame frameshift) was the same as for other CHORD-HRD patients. This suggests that these samples are still HRD despite the secondary frameshift. Only patients with sufficient number of indels (≥100) were included in this analysis.