https://doi.org/10.1093/narcan/zcac013

Multi-omics data integration analysis identifies the spliceosome as a key regulator of DNA double-strand break repair

Dana Sherill-Rofe[1,†], Oded Raban[2,3,†], Steven Findlay[2,4,†], Dolev Rahat[1,5], Irene Unterman[1], Arash Samiei[2,3], Amber Yasmeen[2,3], Zafir Kaiser[6,7], Hellen Kuasne[6,7], Morag Park[6,7,8], William D. Foulkes[9,10], Idit Bloch[1], Aviad Zick[11], Walter H. Gotlieb[3,*], Yuval Tabach[1,*] and Alexandre Orthwein[2,3,8,*]

[1]Department of Developmental Biology and Cancer Research, Institute for Medical Research Israel-Canada, Hebrew University of Jerusalem-Hadassah Medical School, Jerusalem 91120, Israel, [2]Lady Davis Institute for Medical Research, Segal Cancer Centre, Jewish General Hospital, 3755 Chemin de la Côte-Sainte-Catherine, Montréal, QC H3T 1E2, Canada, [3]Division of Gynecology Oncology, Segal Cancer Center, Jewish General Hospital, McGill University, Montreal, QC H3T 1E2, Canada, [4]Division of Experimental Medicine, McGill University, Montreal, QC H4A 3J1, Canada, [5]Department of Genetics, Hadassah Medical Organization, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem 91120, Israel, [6]Department of Biochemistry, McGill University, Montreal, QC H3G 1Y6, Canada, [7]Goodman Cancer Research Centre, McGill University, Montreal, QC H3A 1A3, Canada, [8]Gerald Bronfman Department of Oncology, McGill University, Montréal, QC H4A 3T2, Canada, [9]The Research Institute of the McGill University Health Centre, Montreal, QC H4A 3J1, Canada, [10]Department of Human Genetics, McGill University, Montreal, QC H4A 3J1, Canada and [11]Department of Oncology, Hadassah Medical Center, Faculty of Medicine, Hebrew University of Jerusalem, Ein-Kerem, Jerusalem 91120, Israel

Received October 11, 2021; Revised February 25, 2022; Editorial Decision March 21, 2022; Accepted March 23, 2022

ABSTRACT

DNA repair by homologous recombination (HR) is critical for the maintenance of genome stability. Germline and somatic mutations in HR genes have been associated with an increased risk of developing breast (BC) and ovarian cancers (OvC). However, the extent of factors and pathways that are functionally linked to HR with clinical relevance for BC and OvC remains unclear. To gain a broader understanding of this pathway, we used multi-omics datasets coupled with machine learning to identify genes that are associated with HR and to predict their sub-function. Specifically, we integrated our phylogenetic-based co-evolution approach (CladePP) with 23 distinct genetic and proteomic screens that monitored, directly or indirectly, DNA repair by HR. This omics data integration analysis yielded a new database (HRbase) that contains a list of 464 predictions, including 76 gold standard HR genes. Interestingly, the spliceosome machinery emerged as one major pathway with significant cross-platform interactions with the HR pathway. We functionally validated 6 spliceosome factors, including the RNA helicase SNRNP200 and its co-factor SNW1. Importantly, their RNA expression correlated with BC/OvC patient outcome. Altogether, we identified novel clinically relevant DNA repair factors and delineated their specific sub-function by machine learning. Our results, supported by evolutionary and multi-omics analyses, suggest that the spliceosome machinery plays an important role during the repair of DNA double-strand breaks (DSBs).

INTRODUCTION

DNA repair and the DNA damage response (DDR) have emerged as essential pathways in the onset of several solid malignancies, including breast (BC) and ovarian (OvC) cancers. Indeed, germline mutations in DNA repair genes such as *BRCA1* and *BRCA2*, which are key players in the homologous recombination (HR) pathway, favor the development of hereditary breast and ovarian cancer (HBOC) and

[*]To whom correspondence should be addressed. Email: alexandre.orthwein@mcgill.ca
Correspondence may also be addressed to Yuval Tabach. Email: yuvaltab@ekmd.huji.ac.il
Correspondence may also be addressed to Walter H. Gotlieb. Email: walter.gotlieb@mcgill.ca
[†]Co-first Authors.

genetic instability is the key feature of BC and OvC (1,2). Importantly, the penetrance of HR genes can only partially explain familial cases of HBOC (3,4). Large-scale sequencing of different tumor origins, including in BC and OvC, has enabled the discovery of rare germline cancer susceptibility variants of unknown significance (VUS), due to their scarceness (5). In absence of a proper understanding of the factors involved in the regulation of HR-mediated DNA repair, proper genetic counseling and therapeutic response to these types of solid malignancies are compromised.

Computational approaches have been successful in predicting protein function and identifying novel players in different cellular pathways (reviewed in (6)). For instance, several groups have developed mathematical algorithms to cluster proteins into functionally relevant groups based on their amino acid sequence (7–9). Alternatively, studying the evolutionary relationship between two genes, also called phylogenetic profiling (PP), has been a powerful strategy to identify proteins that are functionally related (10). However, integrating the complex evolution of eukaryotic proteins has proven to be challenging, which has ultimately been bypassed by developing a normalized PP (NPP) method, where measuring sequence similarities have been adjusted to the evolutionary distance between query and reference species (11–14). In fact, we recently developed clade-based PP approaches (15–17), where different scales of co-evolution could be looked at to predict protein function, identify new factors in different pathways, and map new 'druggable' targets. For example, using our CladePP approach, we were able to provide a novel insight into HR, allowing the identification of 67 candidates that have never been previously linked to this DNA repair pathway (15).

To further refine predictive tools of protein function, we integrated our CladePP approach with omics datasets and genetic screens studying the DNA damage response into a comprehensive database (HRbase). Using machine learning, we assigned to each gene a score reflecting the strength of its association with the HR pathway. This prediction algorithm identified a total of 464 HR candidates, including 76 gold standard HR genes. Pathway enrichment analysis of our HRbase coupled with functional GFP-based DNA repair assays highlighted the spliceosome machinery as a significant regulator of DSB repair pathways. Importantly, we identified the spliceosome factors SNRNP200, SNW1, and SF3B3 as prognostic biomarkers for both BC and a subset of OvC patients, highlighting the power of our HRbase in identifying novel DNA repair factors with direct relevance for the pathobiology of BC and OvC.

METHODS

HR gold standard list

We compiled a list, based on a literature review, of 78 recognized HR genes that affect either DNA repair or directly regulate the HR pathway (18–27) (Supplementary Table S1). Included are genes from the closely related Fanconi Anemia (FA) pathway and genes such as *TP53BP1* which function in other pathways, yet are known to regulate HR (25,28). Each gene was categorized into a specific functional module within the HR pathway: DSB recognition, DNA end resection, FA pathway, regulation (DNA damage response, DDR), strand invasion & D-loop formation and synthesis, and Holiday Junctions (HJ) processing (15).

Naive Bayesian classifier

To prioritize HR candidate genes, we integrated 24 omics datasets (Supplementary Figure S1A), including our previous CladePP analysis (15) (Figure 1A) and we utilized a naïve Bayesian Classifier as previously described (13). Some of the 24 datasets, including text mining, GO terms, pathway annotations and OMIM rely on previously curated knowledge. While some of these datasets are highly reliable in confirming genes with previous link to HR, such as our gold standard HR genes, dependence on these datasets may diminish our ability to detect genes whose link to HR has yet to be described in the literature (Supplementary Figure S1B; Supplementary Table S1). To reduce biases that result from the duplicated information between databases, we trained two versions of the Classifier, one using all 24 datasets, and one which omitted the datasets mentioned above, such as text mining. To generate a single and easily interpretable list we merged the ranked lists of genes from both versions of the Classifier.

For each gene g we defined $rank\ (g) = \min\{rank_{full}(g), rank_{no-cur}(g)\}$ where rank(g) is the position of g in the combined ranked list (with lower rank indicating higher confidence that the gene is HR related), $rank_{full}(g)$ the rank of g in the Classifier using all datasets and $rank_{no-cur}(g)$ the rank of g in the Classifier trained without the curation based datasets.

To utilize the Classifier's output to predict if a gene with rank s is related to the HR pathway, we calculated the false positive rate $FPR(s) = \frac{FP(s)}{FP(s)+TN(s)}$

With $FP(s) =$ # of non gold standard genes with rank s or lower

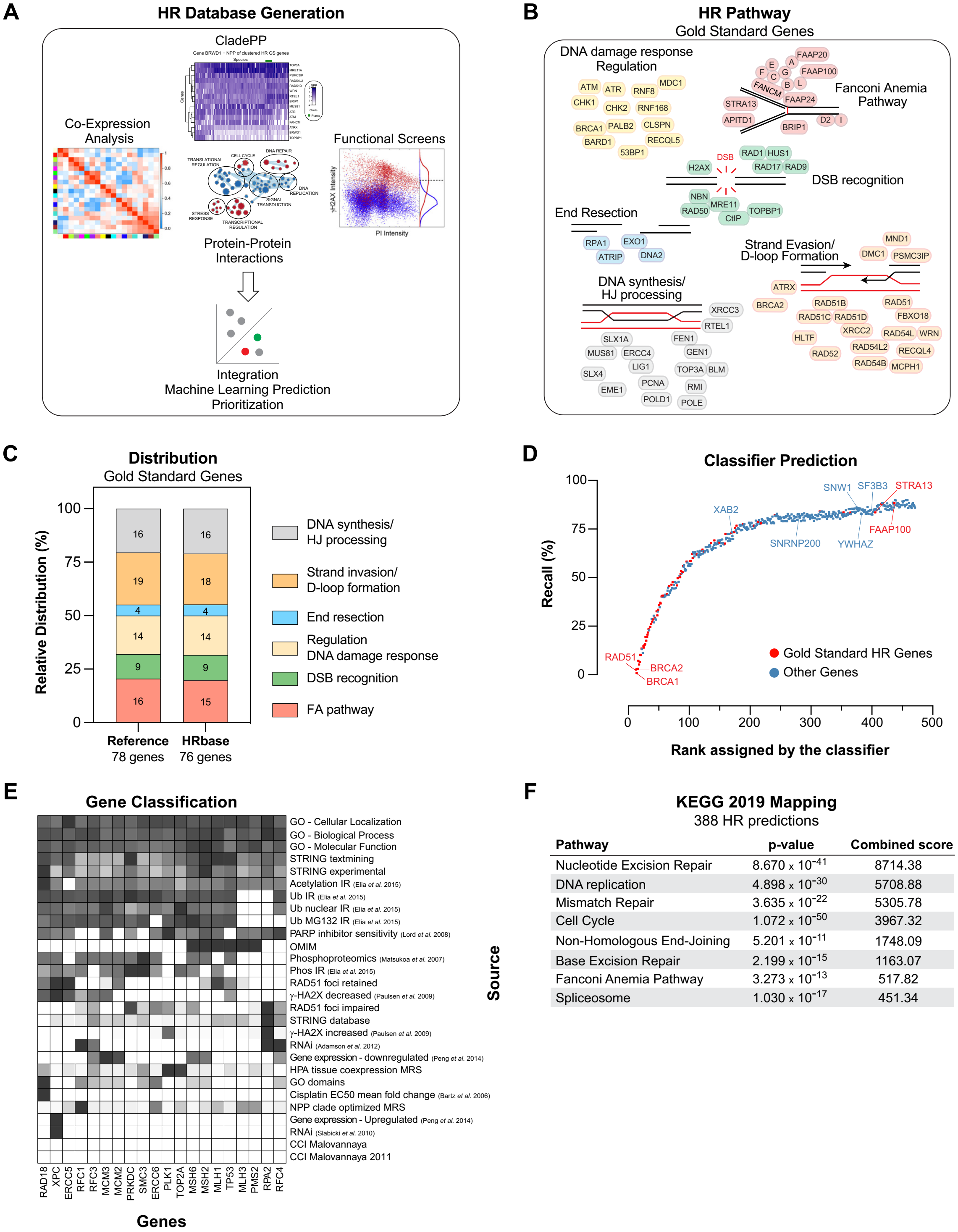
$TN(s) =$ # of non gold standard genes with rank higher than s.

We then considered a gene with rank s as HR related if $FPR(s) \leq 0.02$

We observed that this threshold is sufficiently sensitive to accurately classify most gold standard genes while only introducing a moderate fraction of false positives. Further, beyond this threshold we found a clear saturation in our capacity to detect additional gold standard genes, suggesting that inclusion of genes assigned a lower level of confidence by the Classifier is likely to introduce substantial false positives (Figure 1D).

Characterizing the top Classifier hits

To assign each top ranked gene from the Naïve Bayesian Classifier to a functional module, we constructed a second Classifier, using the Extreme Gradient Boosting algorithm (XGBoost) (29). XGBoost is an ensemble algorithm of decision trees. Weak tree learners trained on parts of the data are combined to generate the output. The weights of the weak learners are adjusted iteratively to reduce the misclassification rate. We treated the 78 gold standard HR genes as labeled examples of the six functional modules. We implemented a one-vs-all multiclass classification approach using



KEGG 2019 Mapping
388 HR predictions

Pathway	p-value	Combined score
Nucleotide Excision Repair	8.670×10^{-41}	8714.38
DNA replication	4.898×10^{-30}	5708.88
Mismatch Repair	3.635×10^{-22}	5305.78
Cell Cycle	1.072×10^{-50}	3967.32
Non-Homologous End-Joining	5.201×10^{-11}	1748.09
Base Excision Repair	2.199×10^{-15}	1163.07
Fanconi Anemia Pathway	3.273×10^{-13}	517.82
Spliceosome	1.030×10^{-17}	451.34

Figure 1. Multi-omics data integration analysis that generated the HRbase. (**A**) Schematic representing how the different datasets were integrated to create the HRbase. (**B**) Schematic representing the different functional modules that participate in the HR pathway and the gold standard HR genes. (**C**) Representation of the distribution of the gold standard HR genes (78 genes) and their presence as part of the HRbase. (**D**) Distribution of the 464 HR predictions based on their score defined by the Classifier algorithm. (**E**) Representation of the top HR predictions and the relative contribution of each dataset in defining their respective score. (**F**) Pathway enriched analysis of the 388 HR predictions using the KEGG 2019 mapping.

the scikit-learn (30) OneVsRestClassifier combining results from 6 different estimators, to help discriminate between the modules. We used the same datasets described above, with datasets based on previously curated knowledge omitted, aggregated by functional module.

Leave-one-out cross-validation (performed with scikit-learn KFold cross-validation) was used to evaluate model performance (Figure 2A). The model is trained on all gold-standard genes except one and classifies the remaining example. This process is repeated n times, n being the number of training examples. The final model was trained on all available data.

We used Local Interpretable Model-agnostic Explanations (LIME) to explain the most confident predictions for each functional module (Supplementary Figure S2B) (31). LIME analysis explains the contribution of individual features to the overall prediction of a single example. It approximates the model locally and enables extracting the contributions of each feature to a specific prediction. The SHAP (SHapley Additive exPlanations) method was used to estimate overall feature importance (32). SHAP scores the contribution of each individual feature to the model. It combines LIME with Shapley values. Overall importance is calculated as the average absolute Shapley value per feature.

Visualization

Heatmaps were generated in R using the ComplexHeatmap package (33). Network representations (Figure 2D) were generated in Cytospace (34). All other visualizations were generated in Python using the Seaborn and Matplotlib packages.

Cell lines and transfection

HeLa cells were cultured in Dulbecco's Modified Eagle medium (DMEM; Wisent) supplemented with 10% fetal bovine serum (FBS, Sigma) and 1% Penicillin-Streptomycin (P/S, Wisent). U2OS cells were cultured in McCoy's 5A Modified medium (Wisent) supplemented with 10% FBS and 1% P/S. All cell lines were regularly tested for mycoplasma contamination and STR DNA authenticated. The DNA-repair reporter cell lines DR-GFP and SA-GFP were a gift of Dr. Jeremy Stark (City of Hope National Medical Center). Plasmid encoding I-SceI was kindly provided by Dr. Daniel Durocher (Lunenfeld-Tanenbaum Research Institute). Plasmid transfections were carried out using Lipofectamine 2000 Transfection Reagent (Invitrogen) following the manufacturer's protocol.

RNA interference

All siRNAs employed in this study were siGENOME Human siRNAs purchased from Dharmacon (Horizon Discovery). RNAi transfections were performed using Lipofectamine RNAiMax (Invitrogen) using forward transfections. Except when stated otherwise, siRNAs were transfected 48 h prior to experimental procedures. The individual siRNA duplexes used are: siCTRL, D-001810-03; CtIP, M-011376-00; CDCA5, MQ-015256–01; DHX9, MQ-009960–01; SNW1, MQ-012446–00; CDC7, MQ-00324–02; SF3B3, MQ-020085–01; STAG2, MQ-021351–01; XAB2, MQ-004914–01; BRDT, MQ-004938–02; IK, MQ-012190–01; MYO3B, MQ-004863–01; ESCO2, MQ-025788–01; PDS5A, MQ-014071–02; SNRNP200, MQ-014161–00, SNRNP200-1, D-014161–01, SNRNP200-2, D-014161–02.

Cell cycle profiling

U2OS cells were sub-cultured to 60% confluency. Cells were transfected with the indicated siRNA and harvested 48hrs post-transfection. Cells were counterstained with DAPI and at least 10 000 events were acquired on a BD FACSCanto II (Becton Dickinson).

GFP-based DNA repair assays

For DR- and SA-GFP reporter assays, U2OS or HeLa cells carrying the respective GFP expression cassette were transfected with the indicated siRNAs. Twenty-four hours after transfection, cells were transfected with empty vector (EV, pDEST-FRT-FLAG) or I-SceI plasmids. After 48 hours, cells were trypsinized, harvested, washed and resuspended in PBS. The percentage of GFP-positive cells were determined by flow cytometry. The data was analyzed using the FlowJo software and presented as previously described (35).

Patient cohort RNA expression analysis

Patients considered for the study were diagnosed and treated for high grade serous ovarian cancer (HGSOvC) at the Jewish General Hospital between 2003 and 2017. Tissue and blood samples were collected at the time of surgery and stored in the gynecologic oncology tumor bank (protocol #03–041). This study was approved by the Jewish General Hospital Research Ethics Board s(protocol #15–070). All patients participating in biobanking, and research activities gave informed written consent.

Patient cohort outcome analysis

Analysis of breast and ovarian patient outcome was performed using KM-plotter (36). Relapse free survival analysis was performed on the whole cohort of breast cancer patient ($n = 4929$ patients). Progression free survival and overall survival analyses were performed on patient diagnosed with serous ovarian cancer ($n = 1104$ and $n = 1207$ patients, respectively). Treatment outcome analysis was performed on breast cancer patients treated with chemotherapy ($n = 1372$ patients) or serous ovarian cancer patients treated with a platinum-based regimen ($n = 979$ patients)

Quantitative real time PCR

RNA was extracted using the RNeasy Mini kit (Qiagen). One μg of RNA was used to prepare cDNA using the LunaScript RT SuperMix (New England Biolabs). cDNA was then diluted 10-fold and 1 μL was used per qRT-PCR reaction. Reactions were performed in triplicate with the Luna Universal qPCR Master mix (New England Biolabs) in a total volume of 10 μL. Primers for reactions are outlined in Supplementary Table S2.

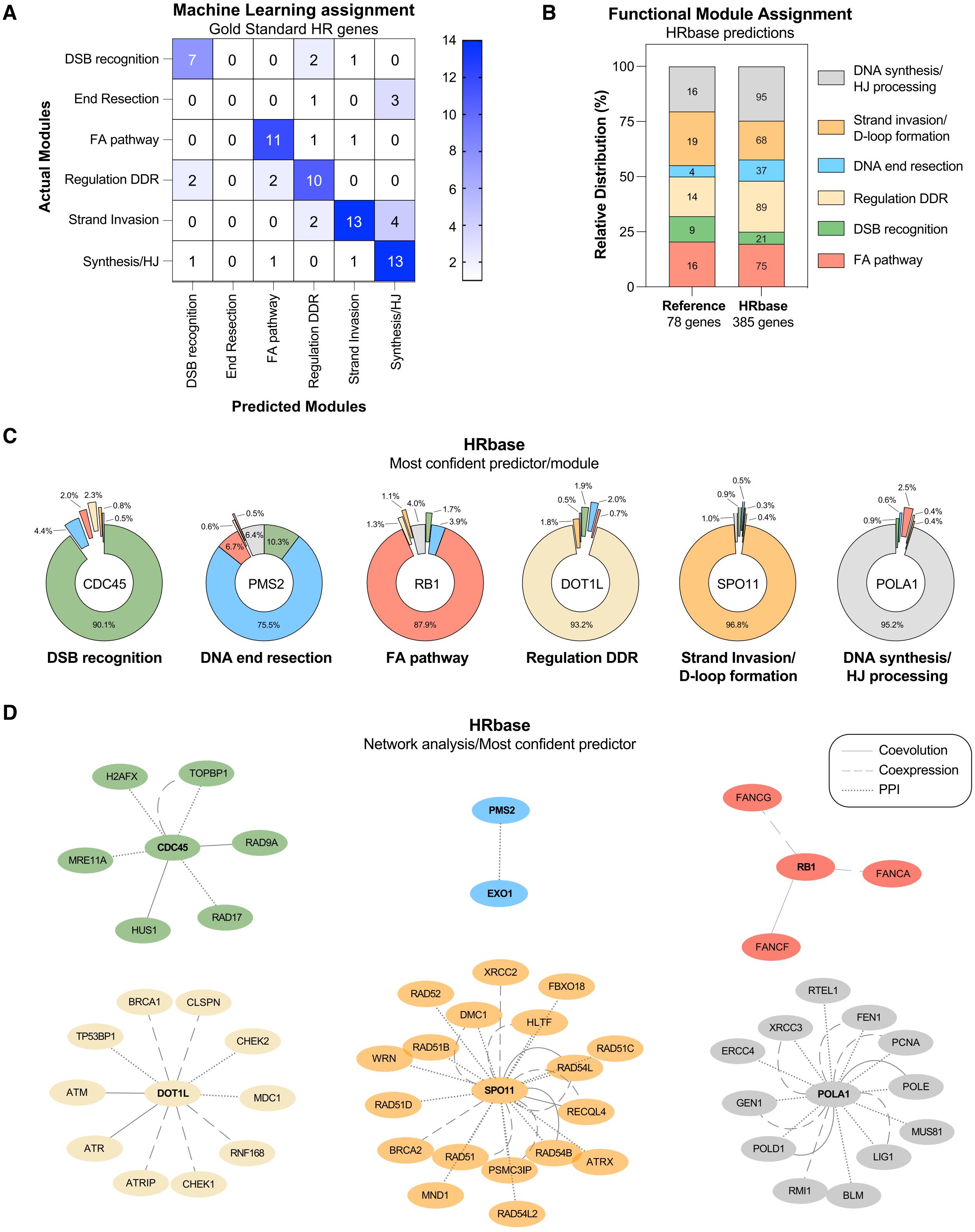


Figure 2. In-depth analysis of the HRbase and assignment of the 388 HR predictions to the different functional HR modules. (**A**) Representation of the actual and prediction distributions of the 78 gold standard HR genes per functional HR modules. (**B**) Distribution of the HRbase predictions per functional HR modules. Only 385 out of the 388 HR predictions were assigned by machine learning to the different function HR modules. (**C**) Examples of HR prediction and their predicted contribution to each functional HR module. (**D**) The network analysis for each HR prediction analysed in Figure 2C. The genes from the relevant HR module which were either co-expressed, co-evolved or with protein interaction evidence with the predicted protein are represented.

RNA-sequencing analysis

HeLa cells were transfected with either siCTRL or two distinct siRNAs targeting SNRNP200 as described above using Lipofectamine RNAiMax (Invitrogen). After 48 hours, cells were either treated with cisplatin (10μM; Tocris) for 6 or 24hrs or solvent (H_2O) before being resuspended in TRIzol® (Invitrogen). RNA was extracted following the manufacturer's protocol and processed for llumina next-generation sequencing to the IRIC Genomics Platform. Adaptor sequences and low-quality bases in the resulting FASTQ files were trimmed using Trimmomatic version 0.35 (37) and genome alignments were conducted using STAR version 2.7.1a (38). The sequences were aligned to the human genome version GRCh38, with gene annotations from Gencode version 37 based on Ensembl 103. As part of quality control, the sequences were aligned to several different genomes to verify that there was no sample contamination. Gene expressions were obtained both as raw readcount directly from STAR as well as computed using RSEM (39), to obtain gene and transcript level expression in reads per transcripts per million (TPM) for these stranded RNA libraries. DESeq2 version 1.30.1 (40) was then used to normalize gene readcounts and compute differential expression between the various experimental conditions. Sample clustering based on normalized log readcounts produces a hierarchy of samples. A principal component analysis is also used to validate that samples correlate as expected.

Statistical analyses

All quantitative experiments are graphed with mean +/- SEM with data from the independent number of independent experiments in the figure legend. All data sets were tested for normal distribution by Shapiro-Wilk Test. Statistical significance was determined using the test indicated in the legend. All statistical analyses were performed in Prism v9 (Graphpad Software).

RESULTS

Functional mapping of the HR pathway by integrating genetic screens and omics datasets

To delineate the extent of genes involved in DNA repair by HR, we selected 24 distinct omics datasets containing a wide range of information related to the relationship of a given human gene to this pathway (Figure 1A) (15,41–45). These datasets can be divided into three main categories (see methods section; Supplementary Table S3): in the first group, we compiled evidence linking a gene to known HR factors (e.g. CladePP, protein-protein interactions) (15,45); the second category encompassed data correlating a given gene to a HR dysregulation phenotype (e.g. RAD51 focus formation, DR-GFP assay) (43,44); finally, we organized any proof of gene alteration upon activation of the DSB response (e.g. ATM/ATR substrates, IR/UV-induced acetylation/ubiquitination sites) (41,42). We applied a naïve Bayesian Classifier approach that we previously used for studying the RNAi machinery (46), to integrate the datasets into a score that predicts association at different levels of a given human gene to the HR pathway (see methods section). For each gene, we established a score where each dataset was independently weighted (Supplementary Table S3). At the moment of designing our algorithm, only RNAi-based functional screens were published (43,44), and incorporated to our multi-omics data integration analysis

While we observed a significant variance in the ability of each dataset to validate well-established HR genes, integrating them into one algorithm yielded a comprehensive list of 464 genes, named HRbase, which includes 76 gold standard HR genes (Figure 1B-C, Supplementary Table S1), and 388 candidate genes identified by either the Classifier or both CladePP and Classifier analyses (Figure 1C-E, Supplementary Table S3).

Using the VarElect software (47), which ranks genes based on their relationship to a given phenotype, we confirmed that 348 HR predictions were associated with the HR pathway (Supplementary Table S4). Importantly, 76 out of the 78 gold standard HR genes were scoring very high using our approach (Figure 1D, Supplementary Table S3). Independent validation using a dataset where each gene was systematically tested by CRISPR for its impact on PARPi response in three distinct cell lines (48), identified 286 out of 320 (89%) predictions to significantly impact PARPi sensitivity in at least one cell line tested (NormZ score ←1 or >1; Supplementary Figure S1C, Supplementary Table S5). Importantly, only 320 out of 388 predictions were scored in these screens due to technical limitations (48).

As expected, KEGG-based pathway enrichment analysis identified several DNA repair pathways as being significantly enriched in our database (Figure 1F), including the FA pathway (p-value = $1.1*10^{-61}$), nucleotide excision repair (NER; p-value = $5.3*10^{-58}$), the HR pathway (p-value = $6.7*10^{-49}$), mismatch repair (MMR; p-value = $1.3*10^{-36}$), base excision repair (BER; p-value = $1.1*10^{-30}$), and non-homologous end-joining (NHEJ; p-value = $7.5*10^{-14}$). Aside from cell cycle (p-value = $4.3*10^{-66}$) and DNA replication (p-value = $5.0*10^{-46}$), we noted that spliceosome-related genes (p-value = $1.7*10^{-16}$) were significantly enriched in our HRbase, suggesting a potential contribution of this pathway to HR. KEGG-based pathway enrichment analysis of our HRbase highlighted the complexity of the HR machinery (Figure 1F). For instance, the E3 ubiquitin ligase RAD18, which plays a critical role in translesion DNA synthesis (49), scored very highly in our Classifier (position 27, Supplementary Table S3), reflecting the known relationship between DNA synthesis and the response to DSBs. We also noted that genes involved in processes (e.g. MMR, NER and BER) closely related to HR, such as *MSH6* and *MSH2, s*cored very well (position 11 and 12 respectively, Supplementary Table S3), confirming previous reports that have documented a role for MMR in regulating the HR pathway (50–55). Altogether, our data suggest that omics data integration analysis can predict novel factors involved, directly or indirectly, in the HR pathway.

Assigning our predictions to the different functional HR modules using machine learning

The HR pathway can be sub-divided into 6 basic functional modules: DNA double strand break (DSB) recognition, DNA end resection, strand invasion/D-loop formation, DNA synthesis/HJ processing, regulation of the DNA damage response (DDR), and the FA pathway (Supplementary Figure S2A). To delineate the functional contribution of our HR predictions, we assigned them to one of these functional modules using the machine learning algorithm XGBoost (29). To assess model performance, we first performed leave-one-out cross validation on the gold standard HR genes that were part of our HRbase. Machine learning assigned 54 out of 74 accurately (73%) (Figure 2A, Supplementary Table S6), with an optimal performance in the FA pathway (11 out of 14 genes), likely due to the strong protein-protein interactions and co-evolution profiles observed between the different factors of this module. The inaccurate assignment observed by machine learning likely reflects the multiple roles of several gold standard HR genes during DNA repair (Supplementary Table S6), exemplified by *MCPH1* (56) and *BLM* (57). Next, we applied our multiclass classification algorithm to HRbase, which assigned 385 out of 388 predictions to the different HR modules (Figure 2B, Supplementary Table S7): 21 candidates were predicted to be involved in DSB recognition, 37 in DNA end resection, 68 in strand invasion/D-loop formation, 95 in DNA synthesis/HJ processing, 89 in regulation of the DDR, and 75 in the FA pathway (Figure 2B, Supplementary Table S7).

In-depth analysis of the 95 predictions assigned to DNA synthesis/HJ processing confirmed a significant enrichment of DNA replication-related genes (p-value = $5.0*10^{-46}$, Figure 1F), including several DNA polymerases (e.g. POLA1, POLD2, and POLI) (Supplementary Table S7). Furthermore, a series of MMR genes (e.g. MSH2, MSH3 and MSH6) were part of this functional module, reflecting the central contribution of MMR during HR (58). Finally, we noted the presence of several proteasomal-related genes as part of this process, including *PSMA-1*, *-4*, *-5* and *-14* (p-value = $1.0*10^{-12}$), likely reflective of a previously unknown role of proteolysis during the resolution of HJ (59).

We focused our attention on the top assignments of each functional HR module. For instance, machine learning predicted CDC45 to participate in DSB recognition (Figure 2C-D), due to its co-evolution with established DSB recognition genes (e.g. RAD9A and HUS1) and PPIs with H2AFX and MRE11A (Figure 2D). Of note, the replicative CMG helicase, composed of CDC45, MCM2-7 and GINS, has been previously linked to single strand break and intrastrand crosslinks (60), and studies in yeast has shown that the CMG complex is required for break induced replication (BIR) (61). Machine learning assigned the MMR factor PMS2 as a regulator of DNA end resection (Figure 2C-D), suggesting a more complex contribution of this gene to HR than previously documented (62). Finally, the histone methyltransferase DOT1L was assigned to the regulation of the DDR, supporting its role in regulating DNA repair by promoting H3K79 methylation and 53BP1 recruitments to DSBs (63).

To understand how machine learning assigned our predictions to the different functional HR modules, we performed a model interpretability analysis using the SHAP method (32). We noted that strong PPIs with other factors of the same module greatly influenced machine learning assignment of our HRbase predictions (Supplementary Figure S2B-C). Globally, CladePP and PPIs were the primary features impacting the attribution of our new HR genes to a functional module (Supplementary Figure S2B-C).

Functional validation identified several spliceosome factors as novel DNA repair factors

To functionally validate our HRbase, we endeavoured to test the relevance of 5 genes that our Classifier (*ESCO2*, *PDS5A*, *SNRNP200*) or the CladePP (*BRDT*, *MYO3B*) predicted to be involved in HR but have yet to be linked to DNA repair (Supplementary Table S8). To test their relevance for HR pathway directly, we employed a well-established GFP-based reporter system that monitors DNA repair by HR, the direct repeat GFP (DR-GFP) assay (Figure 3A-B) (64). As positive control, we targeted the gold standard HR factor CtIP by small interfering RNA (siRNA) as well as 7 HR predictions that have been previously involved in the HR pathway (*CDCA5, DHX9, SNW1, CDC7, SF3B3, STAG2, XAB2*; Supplementary Table S8). As expected, depletion of these latter 7 HR predictions resulted in a significant reduction in HR in both U2OS and HeLa DR-GFP cell lines (indicated in blue in Figure 3C, Supplementary Figure S3A). Interestingly, all 5 genes that have yet to be linked to DNA repair also significantly impaired HR in both cell lines (indicated in orange and red in Figure 3C, Supplementary Figure S3A). We focused our attention on targets that, upon depletion by siRNA, correlated with, at least, 50% reduction in HR in both cell lines, leaving us with 2 putative HR candidates (*SNRNP200* and *ESCO2*) and 4 genes that were previously linked to HR (Figure 3C). We noted a significant reduction of viability in both cell types upon depletion of these 6 candidates followed by DSB induction (Supplementary Figure S3B). However, SNRNP200 knock-down did not impact drastically cell viability in absence of DNA damage (Supplementary Figure S3C), likely reflecting a key role of this spliceosome factor during the DDR. Importantly, we confirmed effective knock-down of each of these targets in U2OS cells by quantitative RT-PCR (qPCR; Supplementary Figure S3D).

Next, we evaluated their relevance in another homology-based DNA repair pathway, single-strand annealing (SSA; Figure 3A), using the SA-GFP assay (Figure 3B) (65). As previously mentioned, we used CtIP as positive control. Interestingly, we noted that depletion of all 6 candidate genes led to a significant reduction in the GFP signal (Figure 3D), indicative of impaired SSA-mediated DNA repair. As homology-driven DNA repair potential is intimately linked to cell cycle positioning, we performed a flow cytometry-based cell cycle analysis of our different siRNA conditions using propidium iodide (PI) DNA staining. As expected, silencing of *CDCA5* in U2OS cells, which have been previously shown to regulate mitotic entry and progression, significantly impaired the progression to the G1 phase of the cell cycle, compared to control conditions (scramble siRNA, Ctrl; Figure 3E). Importantly, four (SNRNP200, SNW1, SF3B3 and XAB2) out of the five HR predictors that validated in both DR- and SA-GFP DNA reporter assays without drastically impacting cell cycle distribution, are linked to the spliceosome machinery. Of note, target-

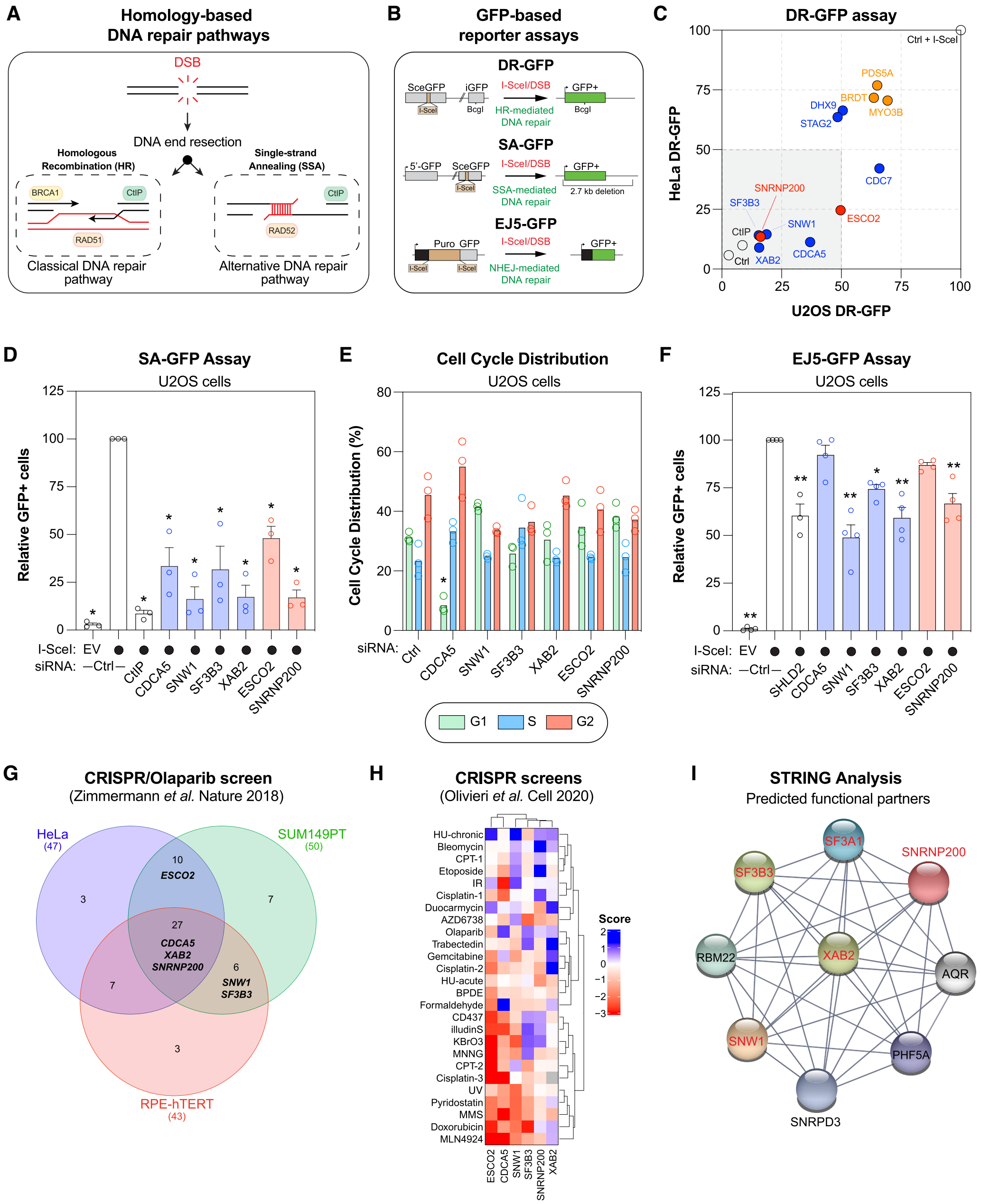


Figure 3. Functional validation of a subset of predictions identified SNRNP200 and its co-factors as regulators of several DNA repair pathways. (**A**) Schematic representing the main homology-based DNA repair pathways, HR and SSA. (**B**) Schematic representing the three GFP-based DNA repair assays used to validate our predictions, the DR-GFP assay for HR-mediated DNA repair (top panel), the SA-GFP assay for SSA-mediated DNA repair (middle panel), and the EJ5-GFP assay for NHEJ-mediated DNA repair. (**C**) U2OS and HeLa cells containing the DR-GFP reporter construct were transfected with the indicated siRNA. Twenty-four hours post-transfection, cells were transfected with the I-SceI expression plasmid or an empty vector (EV), and the GFP+ population was analyzed 48h post-plasmid transfection. The percentage of GFP+ cells was determined for each individual condition and subsequently normalized to the non-targeting condition provided with I-SceI (Ctrl + I-SceI). Data are represented as the mean (x axis for the U2OS cells, y axis for the HeLa cells; $n \geq 3$ biological replicates). (**D**) U2OS cells containing the SA-GFP reporter plasmid were processed and analyzed as in (C). Data are represented as the mean ± SEM, each replicate being representing as a round symbol ($n = 3$ biological replicates). Significance was determined by one-way ANOVA followed by a Dunnett's test. *P≤0.0001. (**E**) Cell cycle distribution was monitored in U2OS cells transfected with the indicated siRNA. Forty-eight hours post-transfection, cells were harvested and stained with propidium iodide. Data are the percentage of cells in G1, S and G2 phases of the cell cycle for each indicated condition and are represented as a bar graph showing the relative mean, each replicate being representing as a round symbol (n = 3 biological replicates). Significance was determined by two-way ANOVA followed by a Sidak's test. *$P<0.05$. (**F**) U2OS cells containing the EJ5-GFP reporter plasmid were processed and analyzed as in (C). Data are represented as the mean ± SEM, each replicate being representing as a round symbol ($n = 3$ biological replicates). Significance was determined by one-way ANOVA followed by a Dunnett's test. *$P\leq 0.0001$, **$P<0.05$. (**G**) Venn diagram representing the overlap of the CRISPR screens published by (48) where sensitivity to the PARPi olaparib was tested in three different cell lines. Only the gold standard HR genes whose inactivation by CRISPR provided a significant effect in these respective screens tested (NormZ score<1) are plotted. The predictions that we functionally tested are indicated in this Venn diagram. (**H**) Heatmap clustering representing the NormZ of our predictions in the series of CRISPR screens published by (68). (**I**) STRING analysis of the different spliceosome factors that were functionally validated as part of our HR predictions (indicated in red).

ing of our HR predictions drastically impaired the capacity of U2OS cells to form RAD51 foci (Supplementary Figure S3E) (44). Finally, we employed the EJ5-GFP reporter assay to monitor end-joining between distal DSB ends of two tandem I-SceI sites (66), which is reflective of total NHEJ events, in the same experimental conditions as previously described (Figure 3B). Here, we used the recently identified NHEJ factor SHLD2 as positive control (35,67). Notably, targeting SNRNP200, SNW1, SF3B3 and XAB2 significantly impaired total NHEJ events in the U2OS EJ5-GFP cells (Figure 3F), suggesting that the spliceosome machinery participates in multiple DSB repair pathways.

To independently validate our 6 HR candidates, we took advantage of new omics datasets where CRISPR-based genome-wide dropout screens were completed in either neoplastic (HeLa and SUM149PT) or non-transformed (RPE1-hTERT) cell lines using the PARP inhibitor olaparib as a selective agent (48). As control, we used our 78 gold standard HR genes and were able to identify the respective NormZ score for 68 of them in this dataset (Supplementary Table S9). CRISPR-mediated inactivation of 63 gold standard HR genes, including *BRCA1*, *PALB2* and *BRCA2* sensitized to the PARP inhibitor (PARPi) olaparib in at least one cell line (NormZ score←1; Figure 3G, Supplementary Table S9) (48). Interestingly, inactivation of our six predicted HR genes (*CDCA5*, *SNW1*, *SF3B3*, *XAB2*, *ESCO2* and *SNRNP200*) correlated with an increased sensitivity to olaparib in at least two cell lines (Figure 3G, Supplementary Table S9) (48). We extended our independent validation to a series of CRISPR-based screens completed in RPE1-hTERT cells against different 31 distinct genotoxic agents by focusing our attention on the ones where HR, FA/ICL, NER and DNA replication fork quality control (QC) genes were highly enriched (total of 26 CRISPR screens) (68). As control, we monitored the NormZ scores of 19 well-established gold standard HR genes (Supplementary Table S9). As expected, CRISPR targeting of this short list of genes sensitized RPE1-hTERT cells to several genotoxic agents that rely on the HR pathway for their processing and repair, including cisplatin, camptothecin (CPT), and the alkylating agent methylnitrosoguanidine (MNNG; Supplementary Figure S3F, Supplementary Table S9) (REF). Importantly, inactivation of *CDC5A* and *ESCO2* sensitized RPE1-hTERT cells to 13 distinct genotoxic agents, including cisplatin, doxorubicin, and illudinS (Figure 3H, Supplementary Table S9) (44). *SNW1* depletion provided an increased sensitivity to 9 carcinogenic agents, including doxorubicin, gemcitabine and MLN4924. Knocking out *SF3B3* hypersensitized RPE1-hTERT cells to 6 drugs tested, while targeting *SNRNP200* modulated the response of 3 chemotherapeutic agents (cisplatin, HU acute, AZD6738). Finally, *XAB2* deletion provided sensitivity to the alkylating agent methyl methanesulfonate (MMS; Figure 3H, Supplementary Table S9). These data suggest that our HRbase contains novel DNA repair factors that have the potential to modulate the response to chemotherapeutic agents, such as PARPi. Strikingly, STRING analysis identified SNRNP200, SNW1, XAB2 and SF3B3 along another HRbase prediction, SF3A1 (indicated in red in Figure 3I) to be functional partners, suggesting that they promote DNA repair as part of a multi-protein complex.

Functional characterization of SNRNP200 during the response to cisplatin

To gain further insight into the contribution of our six candidates during DNA repair, we used machine learning to assign them to the different functional HR modules (29). Interestingly, CDCA5, SF3B3 and XAB2 were predicted to regulate the DDR, while ESCO2 was allocated to the DSB recognition module (Supplementary Figure S4A-C). Machine learning assigned SNRNP200 and its co-factor SNW1 to strand invasion/D-loop formation, reflecting their association to RAD51 and its paralogs (Supplementary Figure S4C).

To functionally characterize the contribution of the spliceosome to DNA repair, we focused our attention on SNRNP200, the core component of the U5 small nuclear RNA proteins (snRNPs) complex (Figure 4A) (69). We performed systematic transcriptomic analysis in HeLa cells transfected with either two distinct siRNAs targeting SNRNP200 (SNRNP200-1 and -2) or a scramble siRNA (siCtrl), followed by cisplatin (CIS; 10μM for 6hrs and 24hrs) or vehicle (H_2O) treatment (Figure 4B). Principal component analysis of our RNA-seq data showed limited variation between the two distinct siRNAs targeting SNRNP200 (Figure 4C), which allowed us to combine both experimental conditions and compare them to control conditions (Supplementary Tables S10-13). Importantly, none of the gold standard HR genes that we able to detect by RNA-seq were significantly affected by SNRNP200 knockdown in our different experimental conditions (Log_2FC←1.5 or >1.5, P <0.05; Figure 4D, Supplementary Table S10).

Global gene expression profiling identified a limited subset of targets that were significantly down-regulated in all three experimental conditions (Log_2FC←1.5, P <0.05; Figure 4E-F, Supplementary Tables S11-13). We noticed that *CCDC33*, which encodes for a coiled-coil domain containing factor, was down-regulated in all three experimental conditions (Figure 4F, Supplementary Tables S11-13). Interestingly, previous yeast two-hybrid studies have linked it to several known players in the response to DNA double-strand breaks (70–72), including SMAD3 and SMARCB1 (summarized in our STRING analysis in Figure 4G). Altogether, these data suggest that the splicing factor SNRNP200 participates in the DDR, at least in part, by modulating the expression of a limited subsets of genes linked to DNA repair.

Splicing factors identified by our HRbase act as prognostic factors in both HGSOC and BC

To determine whether our Classifier predictions may have clinical relevance, we focused our attention on the four spliceosome factors that we functionally validated (SNRNP200, SNW1, XAB2 and SF3B3) and interrogated their RNA expression in an in-house cohort of high grade serous ovarian cancer (HGSOC patients; n = 26). RNA expression levels from patient samples were compared to normal fallopian tissue, which is thought to represent the cell-of-origin of HGSOC (73). We found that SNRNP200, SNW1, XAB2 and SF3B3 RNA expressions were significantly altered compared to normal fallopian tube (Fig-

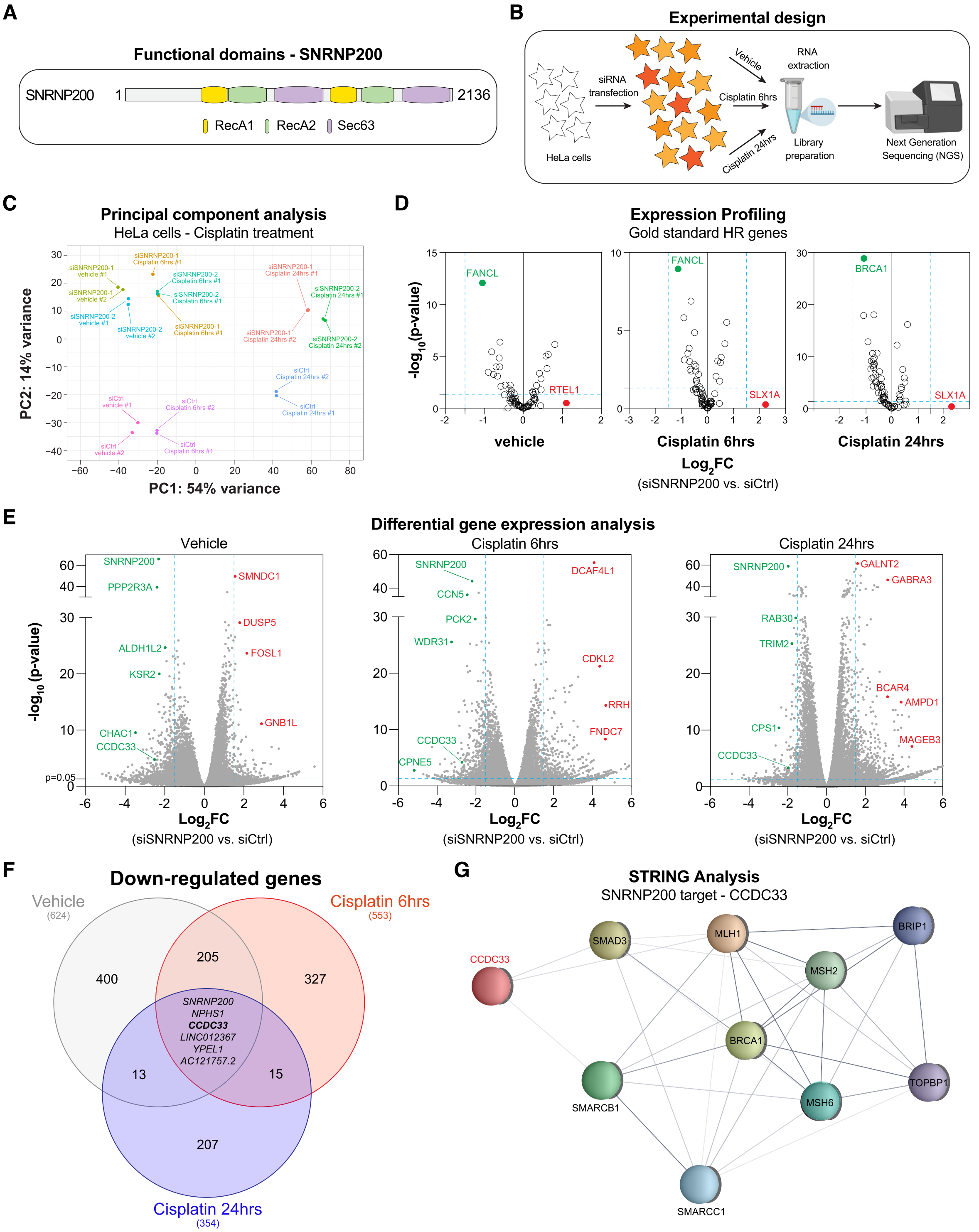


Figure 4. Differential gene expression analysis identified a limited subset of targets impacted by SNRNP200 depletion during the DNA damage response. (**A**) Schematic representing the spliceosome factor SNRNP200 and its structural domains. (**B**) Schematic representing the experimental design used to perform gene expression profiling by RNA-seq analysis in HeLa cells depleted by SNRNP200 and treated with the DNA inter-strand crosslinking agent cisplatin. (**C**) Principal component analysis performed on the first two most significant components of each sample processed for RNA-seq. (**D**) Expression profiling of 73 gold standard HR genes in the experimental conditions monitored by RNA-seq (vehicle, cisplatin 6hrs and 24 hrs). Data are represented as the mean $\log_2$FoldChange ($\log_2$FC) between compiled siSNRP200 (#1 and #2) and siCtrl conditions ($n = 2$ biological replicates) on the x-axis and the -$\log_{10}$ of adjusted p-values (padj) on the y-axis. Significant differentially expressed genes are considered for those with $\log_2\text{FC} < -1.5$ or >1.5 and padj <0.05. (**E**) Global differential gene expression profiling as described in (D). (**F**) Venn diagram representing the overlap between the different experimental conditions described in (B), where genes are significantly down-regulated ($\log_2\text{FC} < -1.5$ with a padj<0.05. We indicated the targets that were significantly downregulated in all three experimental conditions (vehicle, cisplatin 6hrs and 24 hrs). (**G**) STRING analysis of CCDC33 that was downregulated upon SNRNP200 depletion in all three experimental conditions (vehicle, cisplatin 6hrs and 24 hrs) monitored by RNA-seq.

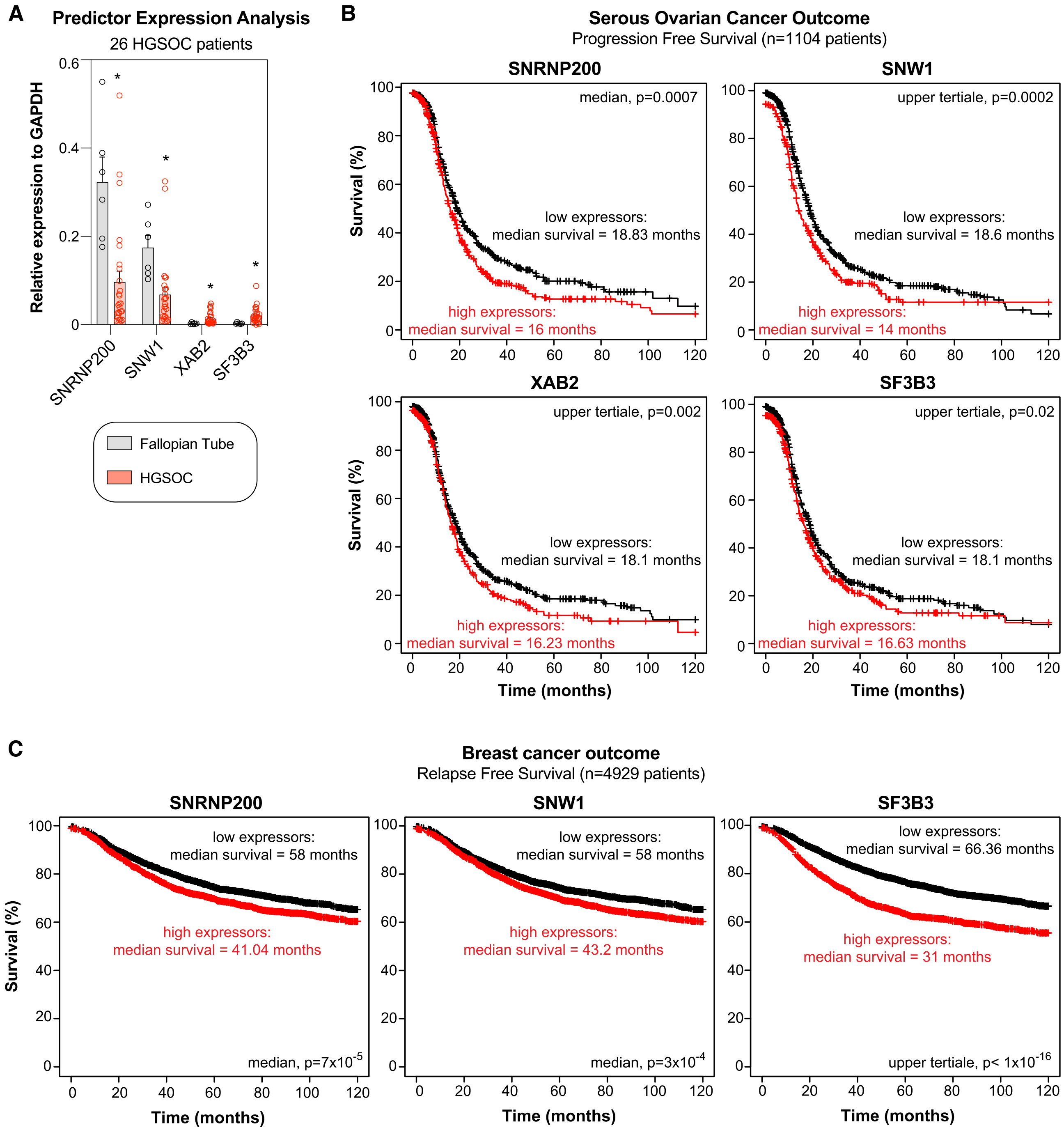


Figure 5. Clinical validation identified SNRNP200 and its co-factors as prognosis factors for both breast and serous ovarian cancers. (**A**) RNA expression analysis of the indicated HR predictions by qPCR in an in-house cohort of HGSOC patients (n = 26 patients). Normal fallopian tube tissues were used as control (n = 6). RNA expression of each gene was normalized to GAPDH. (**B**) Progression free survival analysis based on SNRNP200, SNW1, XAB2 and SF3B3 RNA levels in a cohort of HGSOC patients using KMplot (n = 1104 patients). (**C**) Relapse free survival analysis based on SNRNP200, SNW1 and SF3B3 RNA levels in a cohort of BC patients using KMplot (n = 4929 patients).

ure 5A). Thus, we interrogated a publicly available cohort of more than a thousand HGSOC patients to determine whether the RNA expression of these 4 validated HR genes could correlate with prognosis (74). Remarkably, low expressers of SNRNP200, SNW1, XAB2 and SF3B3 displayed a significantly better progression free survival than their high expressers counterparts (n = 1104 patients; Figure 5B), on average of ∼2 months. Furthermore, HGSOC patients expressing low RNA levels of SNRNP200, XAB2 and SF3B3 have, on average, a better overall survival of 9 months compared to their high expresser counterparts (n = 1207; Supplementary Figure 5A).

To further expand our analysis, we interrogated a publicly available cohort of BC patients (36), and we noted that low

RNA expression of SNRNP200, SNW1 and SF3B3 correlated with a significantly better relapse free survival in BC patients (n = 4929 patients; Figure 5C). If SNRNP200 is truly a HR factor, our model would predict that low expressers of SNRNP200 would respond better to platinum-based regimen and chemotherapy. Indeed, HGSOC patients treated with a platinum-based therapy and expressing low RNA levels of SNRNP200 display an overall better progression free survival than their counterparts (18.27 vs 15.6 months; Supplementary Figure 5B, left panel). Similarly, low expressers of SNRNP200 BC patients treated with chemotherapy have an improved relapse free survival (53.36 months vs 33.4 months; Supplementary Figure 5B, right panel). Overall, these data suggest that the splicing machinery play a key role in the pathobiology of both HGSOC and BC.

DISCUSSION

High-throughput approaches have revolutionized cancer biology, in part by better modelling biological networks (75,76). Here, we report the development of a multi-omics approach to identify novel DNA repair genes relevant for the HR pathway. Integrating 24 distinct datasets (41–45), including our phylogenetic co-evolution profiling approach CladePP (15), into a unique database (HRbase), revealed the central role of the spliceosome in the repair of DSBs, in particular the HR pathway.

RNA and their associated binding proteins (RBPs) have emerged as central elements in the response to DSBs and their resolution (reviewed in (77)). Aside from their canonical contribution in the processing and expression of several DNA repair factors (78,79), RBPs and their structurally diverse RNA-binding domains (80,81), have been more recently involved in the DNA repair process, including the signaling of the break, the remodeling of chromatin at DNA damage sites, DNA:RNA hybrids stabilization, RNA-templated DNA repair and liquid-liquid phase separation (LLPS) (reviewed in (77)). In that regard, the spliceosome machinery exemplifies the versatility of RBPs during the detection, signaling and subsequent repair of DSBs. The splicing of pre-mRNA is a very complex and dynamic process, carried out by several mega-complexes of ribonucleoproteins (RNPs) (82). In fact, the spliceosome is constituted of uridine-rich small nuclear RNPs, named U1, U2, U4/U6 and U5, as well as more than 150 co-factors that assemble on pre-mRNA consensus sequences and perform intron excision and exon ligation. Several splicing factors have been previously involved in the maintenance of genome stability (43,44,83–85). In particular, the spliceosome U2 snRNP factors, such as SF3B3, have been shown to promote genome integrity by regulating the transcription of key DNA repair factors, including BRCA1 and RAD51 (83). Previous work suggest that splicing factors may also be critical in the processing and resolution of R-loops (86). Interestingly, isolation of proteins on nascent DNA coupled to mass spectrometry (iPOND-MS) has recently shown that SNRNP200 accumulates at camptothecin-stalled replication forks (87), suggesting that this splicing factor may play a direct role at stalled replication forks during the replicative stress response. Our data clearly points towards a major contribution of the U5 snRNP complex (SNRPN200, SNW1) in the regulation of homology-directed DNA repair pathways. Structurally, SNRNP200 is very unique as it is the only RNA helicase to contain a tandem array of 2 helicase domains, each of which is made up of 2 RecA-like domains (88–93), suggestive of a reminiscent role in HR.

Our omics data integration analysis further highlighted the clinical relevance of the spliceosome in the progression of both BC and OvC. While the association of defective HR is well established in BC, genomic analyses of HGSOC tumors have so far identified *TP53* mutations as the only common genetic alteration in this subset of OvC (96% of all cases) (94), likely contributing to the high level of genomic instability detected in this histotype. However, dysregulation of the p53 pathway alone appears to be insufficient for the development of HGSOC in mice (95). Furthermore, the low penetrance of OvC in patients carrying germline mutations in *TP53* and developing the Li-Fraumeni syndrome suggests that alteration of the p53 pathway is a pre-requisite for HGSOC initiation (96). Inactivation of the HR is another common genetic feature in HGSOC: germline and somatic mutations in *BRCA1/2* have been detected in up to 50% of HGSOC cases. Alterations in several additional HR genes, including *BARD1*, *BRIP1*, *PALB2*, *RAD51C,* and *RAD51D*, have been reported in HGSOC, though at a much lower frequency (97). Importantly, HR deficiency predicts the therapeutic efficacy of both platinum analogues and PARP inhibitors (PARPi) in BC and HGSOC (98,99). However, PARPi appear to be active beyond HGSOC tumours carrying *BRCA1/2* mutations, in particular those displaying high genomic loss of heterozygosity (LOH) (100), suggesting the presence of additional players in the HR pathway with relevance for the pathobiology of OvC and chemotherapy response. SNRNP200 and its cofactors may be the missing piece lacking in the understanding of HGSOC progression and chemotherapy response.

Altogether, our work demonstrates the direct relevance of using machine learning in the mapping of key molecular pathways and the identification of clinically relevant factors for the diagnosis of pathologies that are still challenging to treat.

DATA AVAILABILITY

Scripts and files have been deposited in GitHub: https://github.com/iditam/data-integration-analysis. RNA-seq data has been deposited in GEO under accession number GSE198887 (https://www.ncbi.nlm.nih.gov/geo/). Flow cytometry data has been deposited in the FlowRepository (https://flowrepository.org/) with the following repository IDs:

- DR-GFP U2OS: FR-FCM-Z559
- DR-GFP HeLa: FR-FCM-Z55A
- SA-GFP: FR-FCM-Z55C
- EJ5-GFP: FR-FCM-Z55G
- Cell cycle:FR-FCM-Z55B

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Cancer Online.

ACKNOWLEDGEMENTS

We are grateful to Daniel Durocher and Jeremy Stark for plasmids and other reagents, to Jennifer Huber, Patrick Gendron and Claudia Dominici for helping with the RNA-seq analysis, and to Amelie Fradet-Turcotte for critical reading.

AuthorContributions: DR, IU and IB performed all the computational analyses presented in this manuscript and analyzed the data under the leadership of DSR. OR designed, performed several functional and clinical validation experiments presented in this manuscript and analyzed the data, with the help of AY and under the supervision of WHG. SF designed several functional validation experiments presented in this manuscript, supervised ZK and analyzed the data, with the help of AS. HK performed the analysis of publicly available BC and OvC datasets, under the supervision of MP. WDF and AZ provided critical insight in the clinical observations presented in this manuscript. YT, DSR and AO conceived the study, designed the research, provided supervision, and wrote the manuscript with input from all the other authors.

FUNDING

OR is a recipient of a research fellowship from the Israel Cancer Research Fund (ICRF). SF is a recipient of a Cole Foundation doctoral scholarship. Work in the AO laboratory was supported by a CCSRI Innovation Grant (#706389), where WDF is the main Principal Investigator, a Transition Grant from the Cole Foundation and an internal Operating Fund from the Sir Mortimer B. Davis Foundation of the Jewish General Hospital. Work in the YT laboratory was supported by the Israel Science Foundation grant no. 1985/13, the Israel Precision Medicine Partnership grant no. 3797/21 and by an Israel Innovation Authority grant (#71420). Work in the WG laboratory was supported by grants from the ICRF, Fonds de la recherche du Québec-Santé (FRQS), the Gloria's Girls Fund, and the Susan and Jonathan Wener Fund.

Conflict of interest statement. None declared.

REFERENCES

1. Nielsen,F.C., van Overeem Hansen,T. and Sorensen,C.S. (2016) Hereditary breast and ovarian cancer: new genes in confined pathways. *Nat. Rev. Cancer*, **16**, 599–612.
2. Welcsh,P.L. and King,M.C. (2001) BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. *Hum. Mol. Genet.*, **10**, 705–713.
3. Harter,P., Hauke,J., Heitz,F., Reuss,A., Kommoss,S., Marme,F., Heimbach,A., Prieske,K., Richters,L., Burges,A. *et al.* (2017) Prevalence of deleterious germline variants in risk genes including BRCA1/2 in consecutive ovarian cancer patients (AGO-TR-1). *PLoS One*, **12**, e0186043.
4. Shimelis,H., LaDuca,H., Hu,C., Hart,S.N., Na,J., Thomas,A., Akinhanmi,M., Moore,R.M., Brauch,H., Cox,A *et al.* (2018) Triple-Negative Breast Cancer Risk Genes Identified by Multigene Hereditary Cancer Panel Testing. *J. Natl. Cancer Inst.*, **110**, 855–862.
5. Federici,G. and Soddu,S. (2020) Variants of uncertain significance in the era of high-throughput genome sequencing: a lesson from breast and ovary cancers. *J. Exp. Clin. Cancer Res.*, **39**, 46.
6. Fetrow,J.S. and Babbitt,P.C. (2018) New computational approaches to understanding molecular protein function. *PLoS Comput. Biol.*, **14**, e1005756.
7. Lee,D.A., Rentzsch,R. and Orengo,C. (2010) GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. *Nucleic. Acids. Res.*, **38**, 720–737.
8. Knutson,S.T., Westwood,B.M., Leuthaeuser,J.B., Turner,B.E., Nguyendac,D., Shea,G., Kumar,K., Hayden,J.D., Harper,A.F., Brown,S.D. *et al.* (2017) An approach to functionally relevant clustering of the protein universe: Active site profile-based clustering of protein structures and sequences. *Protein Sci.*, **26**, 677–699.
9. de Melo-Minardi,R.C., Bastard,K. and Artiguenave,F. (2010) Identification of subfamily-specific sites based on active sites modeling and clustering. *Bioinformatics*, **26**, 3075–3082.
10. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **96**, 4285–4288.
11. Sadreyev,I.R., Ji,F., Cohen,E., Ruvkun,G. and Tabach,Y. (2015) PhyloGene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles. *Nucleic. Acids. Res.*, **43**, W154–W159.
12. Schwartz,S., Agarwala,S.D., Mumbach,M.R., Jovanovic,M., Mertins,P., Shishkin,A., Tabach,Y., Mikkelsen,T.S., Satija,R., Ruvkun,G. *et al.* (2013) High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell*, **155**, 1409–1421.
13. Tabach,Y., Billi,A.C., Hayes,G.D., Newman,M.A., Zuk,O., Gabel,H., Kamath,R., Yacoby,K., Chapman,B., Garcia,S.M. *et al.* (2013) Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature*, **493**, 694–698.
14. Tabach,Y., Golan,T., Hern\'a,ndez-Hern\'a, ndez,Abrahan, Messer,A.R., Fukuda,T., Kouznetsova,A., Liu,J.-G., Lilienthal,I., Levy,C. *et al.*. (2013) Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. *Mol. Syst. Biol.*, **9**, 692.
15. Sherill-Rofe,D., Rahat,D., Findlay,S., Mellul,A., Guberman,I., Braun,M., Bloch,I., Lalezari,A., Samiei,A., Sadreyev,R. *et al.* (2019) Mapping global and local coevolution across 600 species to identify novel homologous recombination repair genes. *Genome Res.*, **29**, 439–448.
16. Stupp,D., Sharon,E., Bloch,I., Zitnik,M., Zuk,O. and Tabach,Y. (2021) Co-evolution based machine-learning for predicting functional interactions between human genes. *Nat. Commun.*, **12**, 6454.
17. Tsaban,T., Stupp,D., Sherill-Rofe,D., Bloch,I., Sharon,E., Schueler-Furman,O., Wiener,R. and Tabach,Y. (2021) CladeOScope: functional interactions through the prism of clade-wise co-evolution. *NAR Genom Bioinform*, **3**, lqab024.
18. Jackson,S.P. and Bartek,J. (2009) The DNA-damage response in human biology and disease. *Nature*, **461**, 1071–1078.
19. Li,X. and Heyer,W.-D. (2008) Homologous recombination in DNA repair and DNA damage tolerance. *Cell Res.*, **18**, 99–113.
20. San Filippo,J., Sung,P. and Klein,H. (2008) Mechanism of eukaryotic homologous recombination. *Annu. Rev. Biochem.*, **77**, 229–257.
21. O'Driscoll,M. and Jeggo,P.A. (2006) The role of double-strand break repair - insights from human genetics. *Nat. Rev. Genet.*, **7**, 45–54.
22. Walden,H. and Deans,A.J. (2014) The Fanconi anemia DNA repair pathway: structural and functional insights into a complex disorder. *Annu. Rev. Biophys.*, **43**, 257–278.
23. Mladenov,E. and Iliakis,G. (2011) Induction and repair of DNA double strand breaks: the increasing spectrum of non-homologous end joining pathways. *Mutat. Res.*, **711**, 61–72.
24. Torres-Rosell,J., Sunjevaric,I., De Piccoli,G., Sacher,M., Eckert-Boulet,N., Reid,R., Jentsch,S., Rothstein,R., Arag\'o,n., Luis and Lisby,M. (2007) The Smc5–Smc6 complex and SUMO modification of Rad52 regulates recombinational repair at the ribosomal gene locus. *Nat. Cell Biol.*, **9**, 923–931.
25. Escribano-Diaz,C., Orthwein,A., Fradet-Turcotte,A., Xing,M., Young,J.T., Tkac,J., Cook,M.A., Rosebrock,A.P., Munro,M., Canny,M.D. *et al.* (2013) A cell cycle-dependent regulatory circuit composed of 53BP1-RIF1 and BRCA1-CtIP controls DNA repair pathway choice. *Mol. Cell*, **49**, 872–883.

26. Chapman,J.R., Taylor,M.R. and Boulton,S.J. (2012) Playing the end game: DNA double-strand break repair pathway choice. *Mol. Cell*, **47**, 497–510.
27. Aparicio,T., Baer,R. and Gautier,J. (2014) DNA double-strand break repair pathway choice and cancer. *DNA Repair (Amst.)*, **19**, 169–175.
28. Moldovan,G.-L. *et al.* (2009) How the fanconi anemia pathway guards the genome. *Annu. Rev. Genet.*, **43**, 223.
29. Chen,T.Q. and Guestrin,C. (2016) XGBoost: A Scalable Tree Boosting System. In: *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*. pp.785–794.
30. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res*, **12**, 2825–2830.
31. Ribeiro,M.T., Singh,S. and Guestrin,C. (2016) In: *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*. pp.1135–1144.
32. Eisenman,R.L. (1967) A profit-sharing interpretation of Shapley value for N-person games. *Behav. Sci.*, **12**, 396–398.
33. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
34. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
35. Findlay,S., Heath,J., Luo,V.M., Malina,A., Morin,T., Coulombe,Y., Djerir,B., Li,Z., Samiei,A., Simo-Cheyou,E. *et al.* (2018) SHLD2/FAM35A co-operates with REV7 to coordinate DNA double-strand break repair pathway choice. *EMBO J.*, **37**, e100158.
36. Gyorffy,B. (2021) Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. *Comput Struct Biotechnol J*, **19**, 4101–4109.
37. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
38. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
39. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinf.*, **12**, 323.
40. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
41. Elia,A.E., Boardman,A.P., Wang,D.C., Huttlin,E.L., Everley,R.A., Dephoure,N., Zhou,C., Koren,I., Gygi,S.P. and Elledge,S.J. (2015) Quantitative Proteomic Atlas of Ubiquitination and Acetylation in the DNA Damage Response. *Mol. Cell*, **59**, 867–881.
42. Matsuoka,S., Ballif,B.A., Smogorzewska,A., McDonald,E.R., Hurov,K.E., Luo,J., Bakalarski,C.E., Zhao,Z., Solimini,N., Lerenthal,Y. *et al.* (2007) ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science*, **316**, 1160–1166.
43. Adamson,B., Smogorzewska,A., Sigoillot,F.D., King,R.W. and Elledge,S.J. (2012) A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. *Nat. Cell Biol.*, **14**, 318–328.
44. Herr,P., Lundin,C., Evers,B., Ebner,D., Bauerschmidt,C., Kingham,G., Palmai-Pallag,T., Mortusewicz,O., Frings,O., Sonnhammer,E. *et al.* (2015) A genome-wide IR-induced RAD51 foci RNAi screen identifies CDC73 involved in chromatin remodeling for DNA repair. *Cell Discov*, **1**, 15034.
45. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic. Acids. Res.*, **43**, D447–D452.
46. Tabach,Y., Billi,A.C., Hayes,G.D., Newman,M.A., Zuk,O., Gabel,H., Kamath,R., Yacoby,K., Chapman,B., Garcia,S.M. *et al.* (2013) Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature*, **493**, 694–698.
47. Stelzer,G., Plaschkes,I., Oz-Levi,D., Alkelai,A., Olender,T., Zimmerman,S., Twik,M., Belinky,F., Fishilevich,S., Nudel,R. *et al.* (2016) VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. *BMC Genomics*, **17**, 444.
48. Zimmermann,M., Murina,O., Reijns,M.A.M., Agathanggelou,A., Challis,R., Tarnauskaite,Z., Muir,M., Fluteau,A., Aregger,M., McEwan,A. *et al.* (2018) CRISPR screens identify genomic ribonucleotides as a source of PARP-trapping lesions. *Nature*, **559**, 285–289.
49. Kannouche,P.L. and Lehmann,A.R. (2004) Ubiquitination of PCNA and the polymerase switch in human cells. *Cell Cycle*, **3**, 1011–1013.
50. Tham,K.C., Kanaar,R. and Lebbink,J.H. (2016) Mismatch repair and homeologous recombination. *DNA Repair (Amst.)*, **38**, 75–83.
51. Chakraborty,U. and Alani,E. (2016) Understanding how mismatch repair proteins participate in the repair/anti-recombination decision. *FEMS Yeast Res.*, **16**, fow071.
52. Chakraborty,U., George,C.M., Lyndaker,A.M. and Alani,E. (2016) A Delicate Balance Between Repair and Replication Factors Regulates Recombination Between Divergent DNA Sequences in Saccharomyces cerevisiae. *Genetics*, **202**, 525–540.
53. Surtees,J.A., Argueso,J.L. and Alani,E. (2004) Mismatch repair proteins: key regulators of genetic recombination. *Cytogenet. Genome Res.*, **107**, 146–159.
54. Elliott,B. and Jasin,M. (2001) Repair of double-strand breaks by homologous recombination in mismatch repair-defective mammalian cells. *Mol. Cell. Biol.*, **21**, 2671–2682.
55. Rajesh,P., Litvinchuk,A.V., Pittman,D.L. and Wyatt,M.D. (2011) The homologous recombination protein RAD51D mediates the processing of 6-thioguanine lesions downstream of mismatch repair. *Mol. Cancer Res.*, **9**, 206–214.
56. Lin,S.Y., Liang,Y. and Li,K. (2010) Multiple roles of BRIT1/MCPH1 in DNA damage response, DNA repair, and cancer suppression. *Yonsei Med. J.*, **51**, 295–301.
57. Kaur,E., Agrawal,R. and Sengupta,S. (2021) Functions of BLM Helicase in Cells: Is It Acting Like a Double-Edged Sword?*Front Genet*, **12**, 634789.
58. Spies,M. and Fishel,R. (2015) Mismatch repair during homologous and homeologous recombination. *Cold Spring Harb. Perspect. Biol.*, **7**, a022657.
59. Matos,J. and West,S.C. (2014) Holliday junction resolution: regulation in space and time. *DNA Repair (Amst.)*, **19**, 176–181.
60. Sparks,J.L., Chistol,G., Gao,A.O., Raschle,M., Larsen,N.B., Mann,M., Duxin,J.P. and Walter,J.C. (2019) The CMG Helicase Bypasses DNA-Protein Cross-Links to Facilitate Their Repair. *Cell*, **176**, 167–181.
61. Lydeard,J.R., Lipkin-Moore,Z., Sheu,Y.J., Stillman,B., Burgers,P.M. and Haber,J.E. (2010) Break-induced replication requires all essential DNA replication factors except those specific for pre-RC assembly. *Genes Dev.*, **24**, 1133–1144.
62. Rahman,M.M., Mohiuddin,M., Shamima Keka,I., Yamada,K., Tsuda,M., Sasanuma,H., Andreani,J., Guerois,R., Borde,V., Charbonnier,J.B. *et al.* (2020) Genetic evidence for the involvement of mismatch repair proteins, PMS2 and MLH3, in a late step of homologous recombination. *J. Biol. Chem.*, **295**, 17460–17475.
63. Wood,K., Tellier,M. and Murphy,S. (2018) DOT1L and H3K79 Methylation in Transcription and Genomic Stability. *Biomolecules*, **8**, 11.
64. Pierce,A.J., Hu,P., Han,M., Ellis,N. and Jasin,M. (2001) Ku DNA end-binding protein modulates homologous repair of double-strand breaks in mammalian cells. *Genes Dev.*, **15**, 3237–3242.
65. Stark,J.M., Pierce,A.J., Oh,J., Pastink,A. and Jasin,M. (2004) Genetic steps of mammalian homologous repair with distinct mutagenic consequences. *Mol. Cell. Biol.*, **24**, 9305–9316.
66. Bennardo,N., Cheng,A., Huang,N. and Stark,J.M. (2008) Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. *PLos Genet.*, **4**, e1000110.
67. Gupta,R., Somyajit,K., Narita,T., Maskey,E., Stanlie,A., Kremer,M., Typas,D., Lammers,M., Mailand,N., Nussenzweig,A. *et al.* (2018) DNA Repair Network Analysis Reveals Shieldin as a Key Regulator of NHEJ and PARP Inhibitor Sensitivity. *Cell*, **173**, 972–988.
68. Olivieri,M., Cho,T., Alvarez-Quilon,A., Li,K., Schellenberg,M.J., Zimmermann,M., Hustedt,N., Rossi,S.E., Adam,S., Melo,H. *et al.*

(2020) A Genetic Map of the Response to DNA Damage in Human Cells. *Cell*, **182**, 481–496.

69. Frank,D.N., Roiha,H. and Guthrie,C. (1994) Architecture of the U5 small nuclear RNA. *Mol. Cell. Biol.*, **14**, 2180–2190.
70. Luck,K., Kim,D.K., Lambourne,L., Spirohn,K., Begg,B.E., Bian,W., Brignall,R., Cafarelli,T., Campos-Laborie,F.J., Charloteaux,B. *et al.* (2020) A reference map of the human binary protein interactome. *Nature*, **580**, 402–408.
71. Rolland,T., Tasan,M., Charloteaux,B., Pevzner,S.J., Zhong,Q., Sahni,N., Yi,S., Lemmens,I., Fontanillo,C., Mosca,R. *et al.* (2014) A proteome-scale map of the human interactome network. *Cell*, **159**, 1212–1226.
72. Sahni,N., Yi,S., Taipale,M., Fuxman Bass,J.I., Coulombe-Huntington,J., Yang,F., Peng,J., Weile,J., Karras,G.I., Wang,Y. *et al.* (2015) Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, **161**, 647–660.
73. Bowtell,D.D. (2010) The genesis and evolution of high-grade serous ovarian cancer. *Nat. Rev. Cancer*, **10**, 803–808.
74. Gyorffy,B., Lanczky,A. and Szallasi,Z. (2012) Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. *Endocr. Relat. Cancer*, **19**, 197–208.
75. Hasin,Y., Seldin,M. and Lusis,A. (2017) Multi-omics approaches to disease. *Genome Biol.*, **18**, 83.
76. Yan,J., Risacher,S.L., Shen,L. and Saykin,A.J. (2018) Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief Bioinform*, **19**, 1370–1381.
77. Klaric,J.A., Wust,S. and Panier,S. (2021) New Faces of old Friends: Emerging new Roles of RNA-Binding Proteins in the DNA Double-Strand Break Response. *Front. Mol. Biosci.*, **8**, 668821.
78. Wickramasinghe,V.O. and Venkitaraman,A.R. (2016) RNA Processing and Genome Stability: Cause and Consequence. *Mol. Cell*, **61**, 496–505.
79. Mikolaskova,B., Jurcik,M., Cipakova,I., Kretova,M., Chovanec,M. and Cipak,L. (2018) Maintenance of genome stability: the unifying role of interconnections between the DNA damage response and RNA-processing pathways. *Curr. Genet.*, **64**, 971–983.
80. Gerstberger,S., Hafner,M. and Tuschl,T. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.
81. Hentze,M.W., Castello,A., Schwarzl,T. and Preiss,T. (2018) A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.*, **19**, 327–341.
82. Lee,Y. and Rio,D.C. (2015) Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu. Rev. Biochem.*, **84**, 291–323.
83. Tanikawa,M., Sanjiv,K., Helleday,T., Herr,P. and Mortusewicz,O. (2016) The spliceosome U2 snRNP factors promote genome stability through distinct mechanisms; transcription of repair factors and R-loop processing. *Oncogenesis*, **5**, e280.
84. Onyango,D.O., Lee,G. and Stark,J.M. (2017) PRPF8 is important for BRCA1-mediated homologous recombination. *Oncotarget*, **8**, 93319–93337.
85. Savage,K.I., Gorski,J.J., Barros,E.M., Irwin,G.W., Manti,L., Powell,A.J., Pellagatti,A., Lukashchuk,N., McCance,D.J., McCluggage,W.G. *et al.* (2014) Identification of a BRCA1-mRNA splicing complex required for efficient DNA repair and maintenance of genomic stability. *Mol. Cell*, **54**, 445–459.
86. Crossley,M.P., Bocek,M. and Cimprich,K.A. (2019) R-Loops as Cellular Regulators and Genomic Threats. *Mol. Cell*, **73**, 398–411.
87. Ribeyre,C., Zellweger,R., Chauvin,M., Bec,N., Larroque,C., Lopes,M. and Constantinou,A. (2016) Nascent DNA Proteomics Reveals a Chromatin Remodeler Required for Topoisomerase I Loading at Replication Forks. *Cell Rep.*, **15**, 300–309.
88. Zhang,X., Yan,C., Hang,J., Finci,L.I., Lei,J. and Shi,Y. (2017) An Atomic Structure of the Human Spliceosome. *Cell*, **169**, 918–929.
89. Bertram,K., Agafonov,D.E., Liu,W.T., Dybkov,O., Will,C.L., Hartmuth,K., Urlaub,H., Kastner,B., Stark,H. and Luhrmann,R. (2017) Cryo-EM structure of a human spliceosome activated for step 2 of splicing. *Nature*, **542**, 318–323.
90. Haselbach,D., Komarov,I., Agafonov,D.E., Hartmuth,K., Graf,B., Dybkov,O., Urlaub,H., Kastner,B., Luhrmann,R. and Stark,H. (2018) Structure and Conformational Dynamics of the Human Spliceosomal B(act) Complex. *Cell*, **172**, 454–464.
91. Zhan,X., Yan,C., Zhang,X., Lei,J. and Shi,Y. (2018) Structures of the human pre-catalytic spliceosome and its precursor spliceosome. *Cell Res.*, **28**, 1129–1140.
92. Zhang,X., Yan,C., Zhan,X., Li,L., Lei,J. and Shi,Y. (2018) Structure of the human activated spliceosome in three conformational states. *Cell Res.*, **28**, 307–322.
93. Zhang,X., Zhan,X., Yan,C., Zhang,W., Liu,D., Lei,J. and Shi,Y. (2019) Structures of the human spliceosomes before and after release of the ligated exon. *Cell Res.*, **29**, 274–285.
94. Masoodi,T., Siraj,S., Siraj,A.K., Azam,S., Qadri,Z., Parvathareddy,S.K., Tulbah,A., Al-Dayel,F., AlHusaini,H., AlOmar,O. *et al.* (2020) Genetic heterogeneity and evolutionary history of high-grade ovarian carcinoma and matched distant metastases. *Br. J. Cancer*, **122**, 1219–1230.
95. Kim,J., Coffey,D.M., Ma,L. and Matzuk,M.M. (2015) The ovary is an alternative site of origin for high-grade serous ovarian cancer in mice. *Endocrinology*, **156**, 1975–1981.
96. Toss,A., Tomasello,C., Razzaboni,E., Contu,G., Grandi,G., Cagnacci,A., Schilder,R.J. and Cortesi,L. (2015) Hereditary ovarian cancer: not only BRCA 1 and 2 genes. *Biomed. Res. Int.*, **2015**, 341723.
97. Suszynska,M., Ratajska,M. and Kozlowski,P. (2020) BRIP1, RAD51C, and RAD51D mutations are associated with high susceptibility to ovarian cancer: mutation prevalence and precise risk estimates based on a pooled analysis of ∼30,000 cases. *J Ovarian Res*, **13**, 50.
98. Jiang,X., Li,X., Li,W., Bai,H. and Zhang,Z. (2019) PARP inhibitors in ovarian cancer: Sensitivity prediction and resistance mechanisms. *J. Cell. Mol. Med.*, **23**, 2303–2313.
99. Livraghi,L. and Garber,J.E. (2015) PARP inhibitors in the management of breast cancer: current data and future prospects. *BMC Med.*, **13**, 188.
100. Pawlyn,C., Loehr,A., Ashby,C., Tytarenko,R., Deshpande,S., Sun,J., Fedorchak,K., Mughal,T., Davies,F.E., Walker,B.A. *et al.* (2018) Loss of heterozygosity as a marker of homologous repair deficiency in multiple myeloma: a role for PARP inhibition?*Leukemia*, **32**, 1561–1566.