Technical Report: Towards Prediction of Platinum Resistance from Expression Levels of Genes Related to Homologous Recombination

Ozan Ozisik, Chiara Facciotto, Antti Häkkinen, Kaiyang Zhang, Manuela Tumiati, Sampsa Hautaniemi, Benno Schwikowski

Technical Report: Towards Prediction of Platinum Resistance from Expression Levels of Genes Related to Homologous Recombination

Ozan Ozisik[1,2,*], Chiara Facciotto[3], Antti Häkkinen[3], Kaiyang Zhang[3], Manuela Tumiati[3], Sampsa Hautaniemi[3], Benno Schwikowski[1,§]

[1]Institut Pasteur (Paris, France), [2]Yildiz Technical University (Istanbul, Turkey),
[3]University of Helsinki (Helsinki, Finland), [*]ozanytu@gmail.com, [§]benno@pasteur.fr

Abstract

The clinical response of high-grade serous ovarian cancer (HGSOC) patients to the standard platinum-based chemotherapy varies widely. While all patients ultimately become resistant, a wide variation in the delay between successive treatments is observed. In an effort to learn more about the molecular basis of resistance, we decided to examine how this variation is manifested in tumor gene expression levels.

The activities of homologous recombination (HR) and Fanconi anemia (FA) pathways are known to correlate with resistance. We therefore hypothesized that the expression of the genes in the above pathways (HR+FA genes) could be linked to resistance in a predictive model that would accurately classify patients into “resistant” or “sensitive” phenotypes. To focus our analysis on the signal of cancer-related cells within bulk data, we applied a computational deconvolution method, and used the epithelial ovarian cancer component to evaluate our hypothesis.

In a dataset from the HERCULES consortium, the accuracy of a classifier based on the HR+FA genes was 0.79, while it was 0.59 in a TCGA dataset. We explored whether any of the HR+FA genes differed significantly between resistant and sensitive patients in the HERCULES dataset. We observed that ABRAXAS1 significantly differed ($p<0.05$ after Benjamini-Hochberg correction, Mann-Whitney U test). Our subsequent accuracy estimate of an ABRAXAS1-only classifier was 0.84. We also found a subgroup of HR-related genes (BRCA1, ABRAXAS1, FANCC, PMS2, RAD50) that led to an accuracy estimate of 0.96.

1. Introduction

In 2019 there will be approximately 22,530 new cases of ovarian cancer diagnosed and 13,980 ovarian cancer deaths in the United States [1]. Ovarian carcinoma are generally classified as Type I or Type II, and different histological subtypes with distinct clinicopathological and molecular genetic features [2]. Type II tumors are primarily composed of high-grade serous carcinomas. These tumors are characterized by late detection, rapid and aggressive progression and poor overall clinical outcome. The standard treatment in high-grade serous ovarian cancer (HGSOC) is aggressive surgery followed by platinum–taxane chemotherapy [3] Although the patients benefit from chemotherapy initially, chemoresistance almost always emerges [2].

In this study we aimed to explore the predictability of platinum resistance in high-grade serous

ovarian cancer (HGSOC) patients using transcriptomic data. Based on the known correlation between the activity of the Homologous Recombination (HR) pathway and primary platinum sensitivity [4], we hypothesized that HR and Fanconi Anemia (FA) related genes (HR+FA genes) could be used to classify new patients as either resistant or sensitive.

2. Materials & Methods

2.1 Datasets

We used HGSOC mRNA expression data obtained from HERCULES consortium and The Cancer Genome Atlas consortium [3].

HERCULES mRNA expression data came from primary tumors. Patients whose primary therapy outcomes were “progressive disease” or “progressive disease after neoadjuvant chemotherapy” were labeled as “resistant”. Patients whose primary therapy outcomes were “complete response” were labeled as “sensitive”. Samples from ascites, adnex and lymph node were removed and samples from fallopian tube, mesenterium, omentum, ovary and peritoneum were used, leaving samples from 7 resistant and 17 sensitive patients. For the patients with multiple samples, expression values were averaged.

TCGA patients were labeled as “resistant” if their recorded platinum-free interval was 180 days or less, and as “sensitive” otherwise. This led to 35 resistant patients and 101 sensitive patients. We used the data from the 35 resistant patients and those 35 sensitive patients with highest (>600) platinum free interval days.

2.2 Processing of bulk mRNA data

To deal with varying proportions of cancer cells in our samples, we used a deconvolution approach to estimate gene expression in epithelial ovarian cancer fraction. We used 16,826 HERCULES single-cell RNA-seq profiles from 18 biopsies (15 patients) to estimate the cell type composition and the constituent expression profiles of each bulk sample. Single cell profiles were classified as ovarian cancer, fibroblast, immune, or unknown cell type using clustering and marker genes, and then used to decompose the bulk samples in these components.

The deconvolution method models heterogeneity and adapts to unmatched patients and specificities of each bulk sample. This allows the estimation of specific cancer cell expression profiles free of compositional variation and variation in the expression profiles of adjacent stromal and immune cells.

After deconvolution, we performed normalization and a variance-stabilizing transformation using DESeq2 [5]

2.3 Classification approach

We constructed a classifier using the Nearest Centroid Classification method from the scikit-learn library [6]. Performance was evaluated by repeating randomized 5-fold

cross-validation 10 times, leading to average accuracy estimates across 50 runs.

2.4 Gene selection

We focused our analysis on the combined set of genes that were annotated as either "Homologous Recombination" or "Fanconi Anemia" in Kyoto Encyclopedia of Genes and Genomes (KEGG) [7, 8].

In order to choose a subgroup from HR+FA genes that will lead to better prediction, we used the following approaches independently:

Mann-Whitney U test: We applied Mann-Whitney U test to explore whether any of the HR+FA genes differed between resistant and sensitive patients. We used the most strongly differentially expressed genes for prediction. The degree of differential expression of a gene was quantified by the p-value of the Mann-Whitney U test.

Sequential feature selection (SFS): SFS is a greedy method for feature selection which either starts with an empty feature set and iteratively adds those features that provide the highest increase in empirical classification performance, until no further increase can be achieved (forward selection), or starts with the full feature set and iteratively removes features (backward selection). We used Sequential Forward Selection from MLxtend library [9] to get best two genes among HR+FA genes for prediction. SFS has been applied on each fold of crossvalidation independently.

SFS merging: Genes returned by SFS in each fold of crossvalidation were aggregated and genes that appeared more than 5 times, out of 50 total runs, were selected for the merged set.

3. Results

The balanced accuracy estimates of the nearest centroid classifier, using features obtained by different methods, are given in Table 1.

The accuracy was 0.53 for both HERCULES and TCGA datasets using all genes as features. When HR+FA genes were used, the accuracy increased to 0.79 for the HERCULES dataset and it increased to 0.59 for the TCGA dataset.

In the HERCULES dataset, ABRAXAS1 was found to be significantly differentially expressed between resistant and sensitive patients within HR+FA genes ($p<0.05$ after Benjamini-Hochberg correction, Mann-Whitney U test). Based on this finding we tested an ABRAXAS1-only classifier which gave an accuracy estimate of 0.84.

In the TCGA data, no gene differed significantly, TOP3B was the most differentially expressed gene. The accuracy of a TOP3B-only classifier was 0.63.

The best accuracy estimate was obtained by using features received from SFS merging. In the HERCULES dataset, the gene set ABRAXAS1, FANCC, RAD50, BRCA1, PMS2 resulted in an accuracy of 0.96. In Figure 1 and Figure 2 these genes are highlighted. In the TCGA dataset, the gene set TOP3B, SLX1B, HES1, POLK, RAD51B resulted in an accuracy of 0.73.

Table 1. Accuracy estimates obtained from different candidate gene sets

Method	**HERCULES data**	**TCGA data**
Using all genes	0.53	0.53
HR+FA genes	0.79	0.59
SFS on HR+FA	0.70	0.54
SFS merging on HR+FA	0.96	0.73
10 most strongly differentially expressed HR+FA genes	0.90	0.68
The most strongly differentially expressed HR+FA gene	0.84	0.63

4. Conclusions and Future Directions

In this study we aimed to learn about the molecular basis of platinum-based chemotherapy resistance in high-grade serous ovarian cancer patients. As the activities of homologous recombination and Fanconi anemia pathways are known to correlate with resistance, we hypothesized that the expression of the genes in these pathways could be linked to resistance in a predictive model that would accurately classify patients into “resistant” or “sensitive” phenotypes. We used datasets from HERCULES and TCGA consortia. In two datasets we obtained different predictive gene sets; in the HERCULES dataset the genes were ABRAXAS1, FANCC, RAD50, BRCA1, PMS2, which resulted in an accuracy of 0.96, and in the TCGA dataset the genes were TOP3B, SLX1B, HES1, POLK, RAD51B, which resulted in an accuracy of 0.73. This confirms our hypothesis that Homologous Recombination- and Fanconi Anemia-related transcripts can indeed be used for the prediction of platinum resistance. The differences between the obtained predictive gene sets and between the accuracy scores necessitates further research with other/larger datasets. The obtained gene sets may represent starting points for more detailed statistical and experimental evaluation of the role of these genes in platinum resistance.

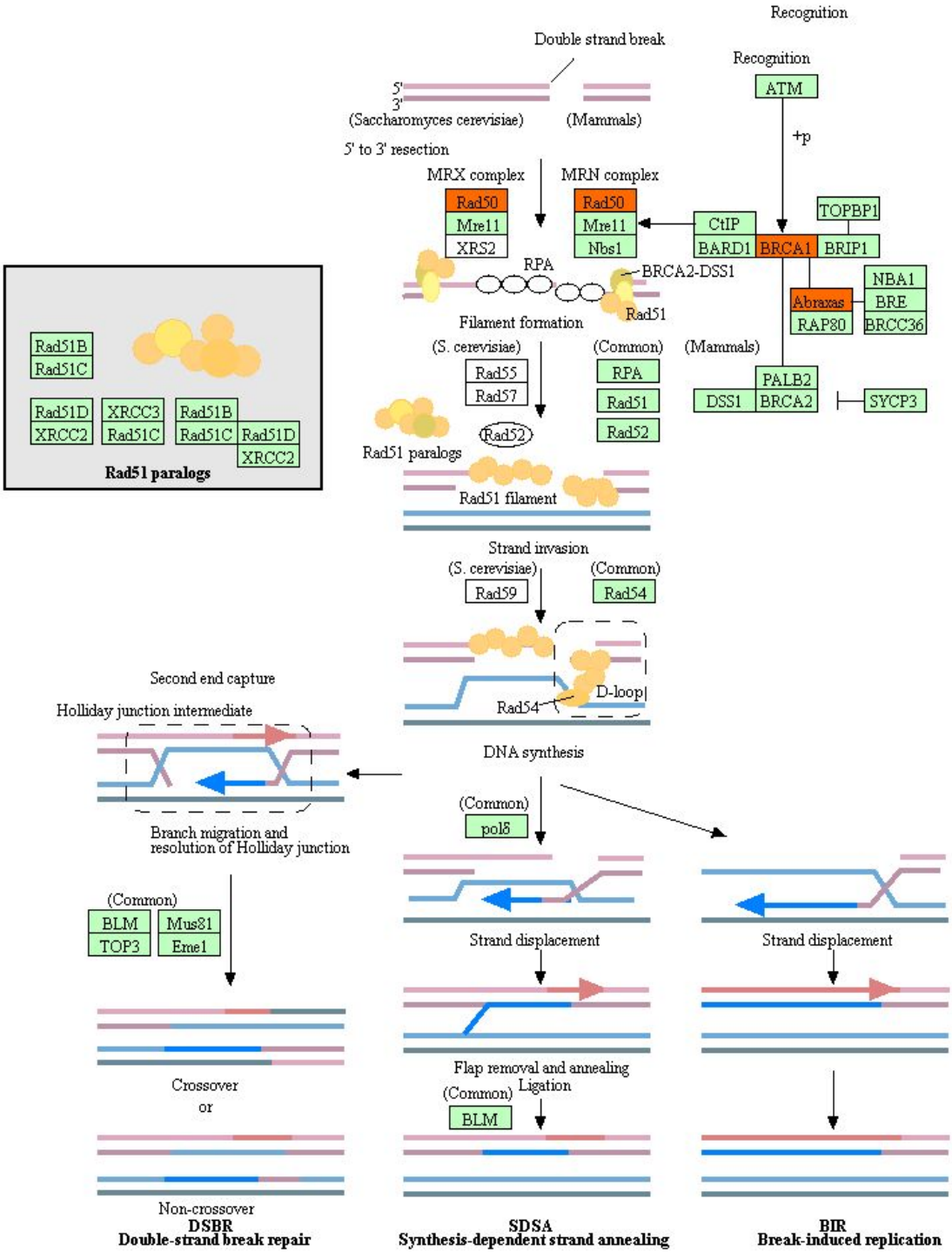

Figure 1. Homologous Recombination pathway. Genes selected by SFS merging (ABRAXAS1, RAD50, BRCA1) are highlighted.

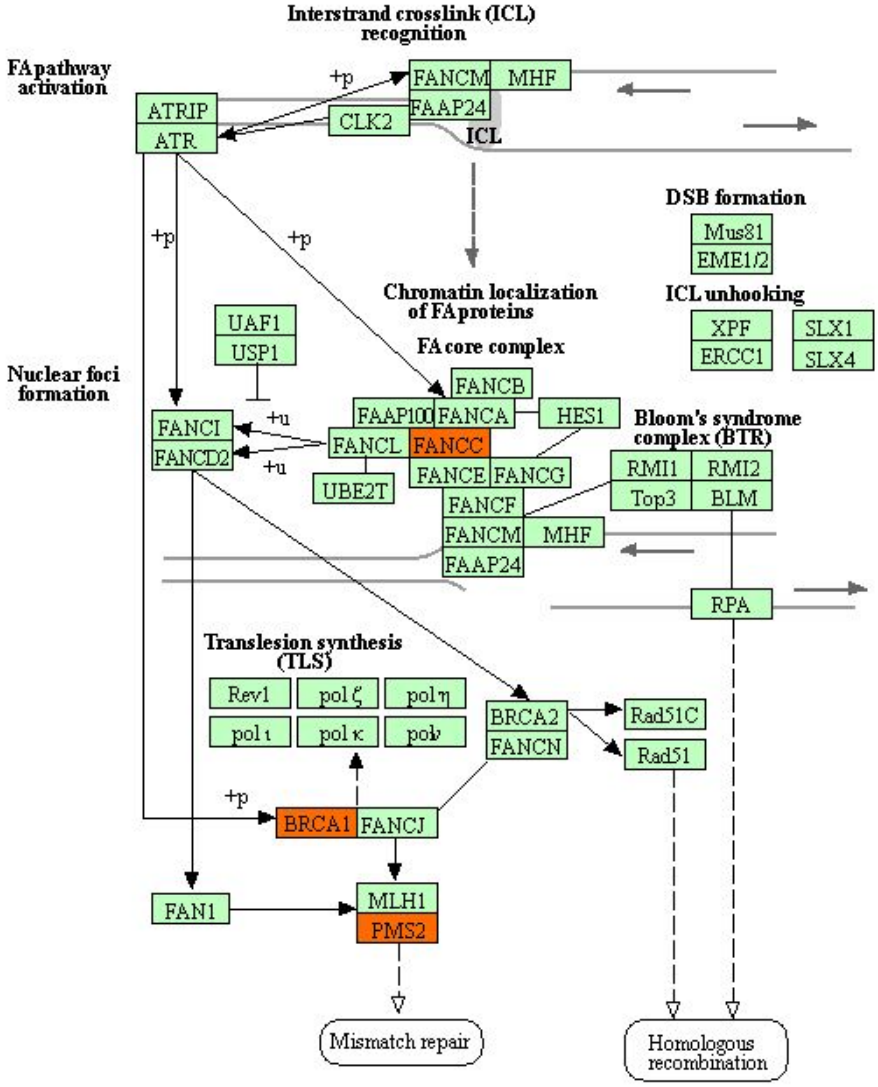

Figure 2. Fanconi Anemia pathway. Genes selected by SFS merging are highlighted.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 667403 for HERCULES. Ozan Ozisik's studies in Institut Pasteur was funded by The Embassy of France in Turkey, HERCULES project funding and Yildiz Technical University.

References

1. Siegel, RL, Miller, KD, Jemal, A **(2019)**. Cancer statistics, 2019. *CA Cancer J Clin.*, 69, 1:7-34.

2. Kurman, RJ, Shih, IeM **(2016)**. The Dualistic Model of Ovarian Carcinogenesis: Revisited, Revised, and Expanded. *Am. J. Pathol.*, 186, 4:733-47.

3. The Cancer Genome Atlas Research Network **(2011)**. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474, 7353:609-15.

4. Tumiati, M, Hietanen, S, Hynninen, J, Pietilä, E, Färkkilä, A, Kaipio, K, Roering, P, Huhtinen, K, Alkodsi, A, Li, Y, Lehtonen, R, Erkan, EP, Tuominen, MM, Lehti, K, Hautaniemi, SK, Vähärautio, A, Grénman, S, Carpén, O, Kauppi, L **(2018)**. A Functional Homologous Recombination Assay Predicts Primary Chemotherapy Response and Long-Term Survival in Ovarian Cancer Patients. *Clin. Cancer Res.*, 24, 18:4482-4493.

5. Love, MI, Huber, W, Anders, S **(2014)**. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15, 12:550.

6. Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, Grisel, O, Blondel, M, Prettenhofer, P, Weiss, R, Dubourg, V, Vanderplas, J, Passos, A, Cournapeau, D, Brucher, M, Perrot, M, Duchesnay, E **(2011)**. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

7. Kanehisa, M, Goto, S **(2000)**. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28, 1:27-30.

8. Kanehisa, M, Sato, Y, Furumichi, M, Morishima, K, Tanabe, M **(2019)**. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.*, 47, D1:D590-D595.

9. Raschka, S **(2018)**. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software*, 3, 24:638.