Prediction of Platinum Resistance from Expression Levels of Genes Related to Homologous Recombination

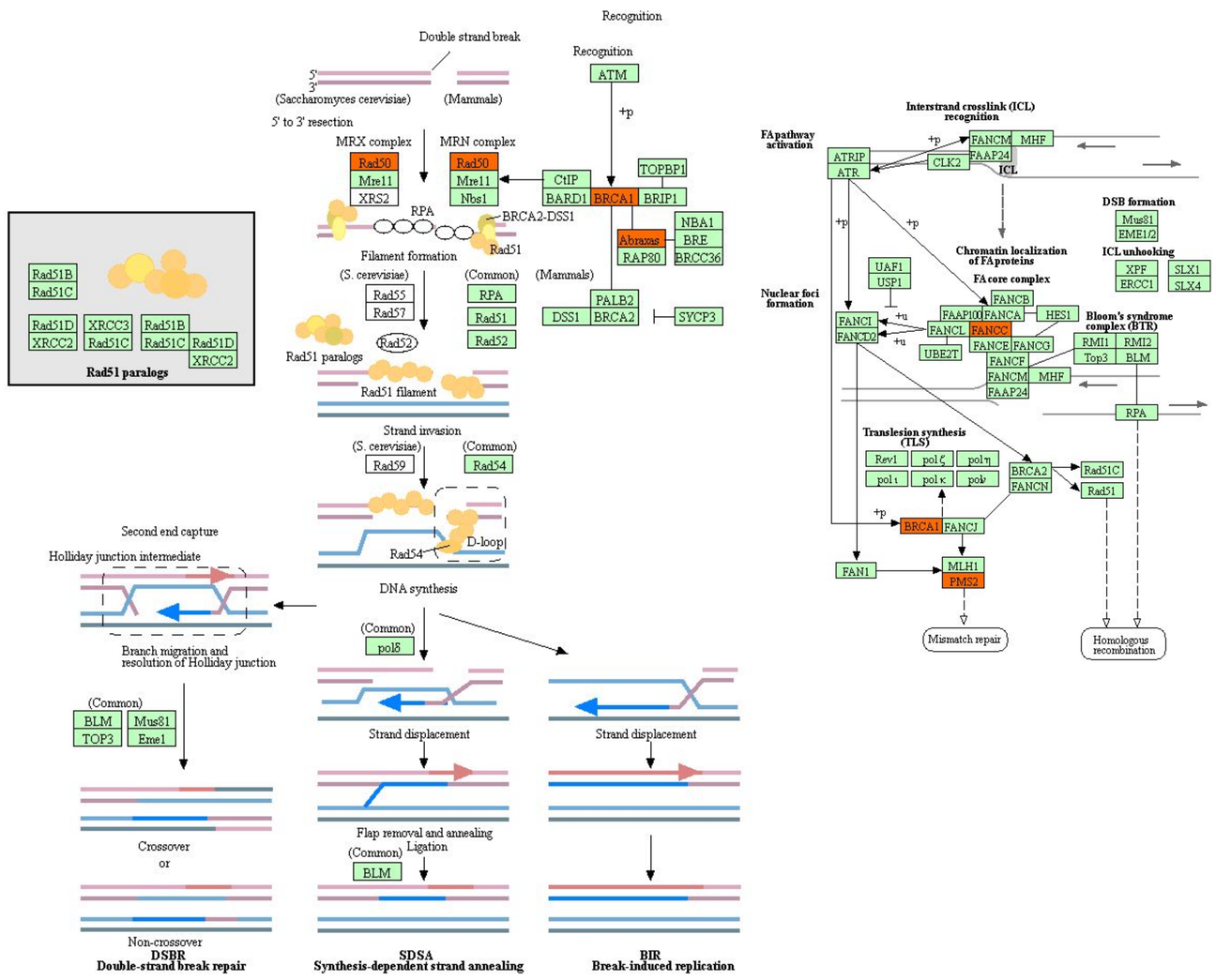
Ozan Ozisik, Chiara Facciotto, Antti Häkkinen, Kaiyang Zhang, Manuela Tumiati, Sampsa Hautaniemi, Benno Schwikowski

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 Computer Engineering Department (Istanbul, Turkey), [3]University of Helsinki (Helsinki, Finland)
[*]ozanytu@gmail.com, [§]benno@pasteur.fr

BACKGROUND

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 [Tumiati et al., Clin Cancer Res 24 (18), 2018], we hypothesized that HR- and Fanconi Anemia (FA)-related genes can be used to classify new patients as either resistant or sensitive.



Homologous Recombination and Fanconi Anemia pathways.
Genes identified as predictors on HERCULES dataset are colored in orange red.

RESULTS

In the HERCULES dataset, **ABRAXAS1** was found to be significantly differentially expressed between resistant and sensitive patients, and within HR+FA genes (p_{adj}<0.05 after Mann-Whitney U test and Benjamini-Hochberg correction).
The accuracy of an ABRAXAS1-based classifier was 0.84.
In the TCGA data, no gene differed significantly, **TOP3B** was the most differentially expressed gene.
The accuracy of a TOP3B-based classifier was 0.63.

The best accuracy score was obtained by merging frequently selected features on different runs.
In the **HERCULES** dataset, the gene set **ABRAXAS1**, **FANCC**, **RAD50**, **BRCA1**, **PMS2** resulted in an accuracy of **0.96**.
In the **TCGA** data, the gene set **SLX1B**, **HES1**, **POLK**, **TOP3B**, **RAD51B** resulted in an accuracy of **0.73**.

Method	HERCULES	TCGA
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
Top 10 genes according to Mann Whitney U test from HR+FA	0.90	0.68
Top gene according to Mann Whitney U test from HR+FA	0.84	0.63

Recall values for resistant and sensitive patients were similar in general except the case of using all genes for the HERCULES dataset, in which accuracy was 0.53, recall for resistant patients was 0.72 and recall for sensitive patients was 0.34.

MATERIALS & METHODS

Datasets

HERCULES mRNA expression data
- Samples came from primary tumors.
- Patients whose primary therapy outcomes were “progressive disease” or “progressive disease after NACT” were classified as resistant.
- Patients whose primary therapy outcomes were “complete response” were classified as sensitive.
- Samples from Ascites, Adnex, Lymph Node were ignored.
- This led to 7 resistant and 17 sensitive patients.
- For the patients with multiple samples, counts were averaged.

TCGA mRNA expression data
- Patients were marked as “Resistant” or “Sensitive”, based on their stated platinum-free interval (less or more than 180 days).
- This led to 35 resistant and 35 sensitive patients.

Gene selection
We focused our analysis on the “HR+FA” set of genes in the KEGG database that were annotated with either “Homologous Recombination” or “Fanconi Anemia”.

Deconvolution, other preprocessing of bulk mRNA data
To deal with the varying proportions of cancer cells in our samples, we used a deconvolution approach to estimate gene expression in the epithelial ovarian cancer fraction. We used 16,826 HERCULES single-cell RNA-seq profiles from 18 biopsies (15 patients) to estimate the 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.

Classification approach
We constructed a classifier using the HR+FA genes and the Nearest Centroid Classification method. Performance was evaluated by repeating 5-fold cross-validation 10 times, leading to average accuracy scores from 50 runs.

Feature selection
Sequential feature selection (SFS): SFS was applied to HR+FA genes. For each iteration in 50 runs (10 times run of 5-fold cross-validation), two genes that gave the highest score on training folds (using a Nearest Centroid Classifier) were selected.
SFS merging: Genes that were frequently (>5 times in 50 runs) selected by SFS were selected for the merged set.
Mann-Whitney U test: We explored whether any of the HR+FA genes differed between resistant and sensitive patients.

CONCLUSIONS AND FUTURE DIRECTIONS

- **Predictors based on all measured genes leads to performance close to random.**
- **SFS merging increases predictive performance on HERCULES and TCGA datasets.**
- **Certain small gene sets from the HR+FA pathways lead to good predictive performance in HERCULES data.**
- **Predictive accuracy is better in HERCULES data.**

Questions for further research

- Does the good performance of SFS merging replicate in other/larger datasets?
- What is the role of the identified HR and FA genes in chemotherapy resistance?
- What are the technical or biological limitations that prevent the replication of our classifiers on TCGA data?
- How can the performance of our classifiers be enhanced towards clinically actionable predictions?

This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 667403 for HERCULES