INTEGRATING LIPIDOMICS AND PROTEOMICS FOR INCREASED DIAGNOSTIC ACCURACY IN PROSTATE CANCER

Ammara Muazzam[1,2], Lee A. Gethings[3,4,6], Christopher J. Hughes[3], Nyasha Munjoma[3], Robert S. Plumb[5], Joanne Ballantyne[3], Olivier Cexus[6], Fowz Azhar[7], Hardev Pandha[6], Anthony D. Whetton[1,2], Paul A. Townsend[1, 2, 6 §, *] and Nophar Geifman[6, 8, §]

1. Division of Cancer Sciences, University of Manchester, UK; 2. Stoller Biomarker Discovery Centre, University of Manchester, UK; 3. Waters Corporation, Wilmslow, UK; 4.Faculty of Biology & Medicine, University of Manchester, UK; 5. Waters Corporation, Milford, MA, USA; 6. Faculty of Health & Medical Sciences, University of Surrey, UK; 7. Salford Royal NHS Foundation Trust, Salford, Manchester, UK; 8. Centre for Health Informatics, University of Manchester, UK.

Waters
THE SCIENCE OF WHAT'S POSSIBLE.®

HIGHLIGHTS

- **Combined proteomic and lipidomic biomarker signatures to potential improve the accuracy of differentiating between healthy controls and mild/advanced stages of prostate cancer (PCa). AUCs of 0.955 and 0.966 were determined for mild and advanced PCa respectively.**
- **Pathway analysis highlighted a number of pathways exclusively associated with mild/advanced PCa. Acute phase response signalling is one example pathway which was identified for both stages of PCa investigated.**
- **The LipidQuan™ platform and label-free (UDMS^{E})[1] proteomic workflows generated a comprehensive list of candidate biomarkers that not only allowed PCa stages to be differentiated but also provided deep biological insight of the mechanisms which underpin PCa.**

METHODS

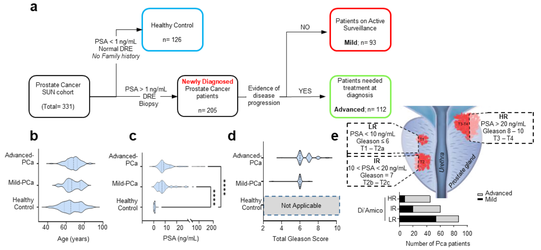

Figure 1. Study overview (a) Patient categories of the SUN cohort, representing healthy controls, mild PCa and advanced PCa groups; (b) Age distribution for each of the groups; (c) PSA level distributions for each of the groups; (d) Total Gleason Scores representing the mild and advanced PCa groups; (e) Di'Amico scores to indicate risk levels for mild and advanced PCa. LR = Low Risk, IR = Intermediate Risk, HR = High Risk

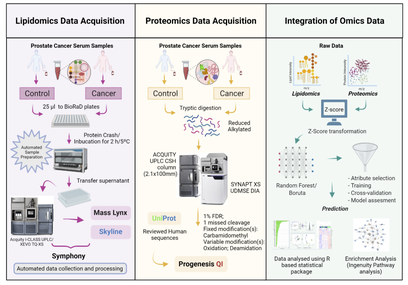

Figure 2. Sample preparation, data acquisition and data analysis pipeline. Sera samples intended for lipidomic analysis were prepared and analysed using the LipidQuan Workflow. Samples were protein precipitated using IPA and LC-MS analysed using the ACQUITY™ I-Class UPLC™ system coupled to a Xevo™ TQ-XS mass spectrometer. The resulting data were then automatically transferred via Symphony™ and processed using TargetLynx™ and Skyline[2] software.

Proteomic analysis consisted of tryptically digesting sera samples prior to LC-MS analysis, which comprised of a ACQUITY Premier I-Class UPLC system coupled to a SYNAPT™ XS mass spectrometer. Peptides were separated over a 15 min gradient using a ACQUITY UPLC CSH™ (2.1x100 mm) Column. Data were subsequently processed using Progenesis™ QI for Proteomics software.

The acquired data from both platforms was integrated using a bioinformatics method which was subjected to attribute selection, using machine learning algorithms to construct a diagnostic model. Further downstream statistical analysis and enrichment analysis was also performed to estimate their statistical significance across the tested conditions and to determine their role in diseases and functions.

References

1. Distler, U., Kuharev, J., Navarro, P. *et al.* Label-free quantification in ion mobility–enhanced data-independent acquisition proteomics. *Nat Protoc* **11,** 795–812 (2016). https://doi.org/10.1038/nprot.2016.042
2. J Proteome Res. 2020 Apr 3;19(4):1447-1458. doi: 10.1021/acs.jproteome.9b00640. Epub 2020 Mar 26. PMID: 31984744; PMCID: PMC7127945

INTRODUCTION

The diagnosis and prognostication for Prostate cancer (PCa) remains challenging in this relatively common cancer. Diagnosis can involve a number of clinicopathological indicators, including Gleason score and prostate-specific antigen (PSA) for example. PSA which is the most common of the few blood-based protein biomarkers currently available in clinical practice, however, PSA by itself is not accurate especially since there is no reliable PSA range that explicitly signifies the presence of PCa. Studies on potential biomarkers and measurable signatures for the disease thus remain a key area of translational research. Appropriate assay accuracy, precision and sensitivity still need to be attained. Combining multiple levels of molecular information has the potential to improve biomarker panels and help gain a better understanding into the biology underpinning the disease. Within this study, we used a combination of lipid and protein measurements to identify biomarkers that are more beneficial in detecting the disease status of men who are most likely to develop PCa. Data from newly diagnosed PCa patients at various stages of the disease, as well as age-matched controls, were used to generate proteomic and lipidomic profiles. Serum samples were collected from newly diagnosed prostate cancer patients and their age matched healthy individuals. Healthy control (n=126) samples satisfied both a normal digital rectal examination (DRE) and prostate-specific antigen (PSA) levels below 1 ng/mL (<1ng/mL). The inclusion criteria for newly diagnosed prostate cancer patients (n=205) were an abnormal prostate on DRE, symptomatic presentation with high PSA levels and abnormal biopsy; or alternatively, a diagnosis made solely on a steep rise in PSA associated with urinary symptoms. We identified signatures for mild and advanced staged PCa, providing AUC values of 0.955 and 0.966, respectively. Combining lipidomic and proteomic data, provided a striking separation between cancer and non-cancer samples. Importantly, we found that based on the top five biomarkers (i.e., combination of lipids and proteins) provided cumulative AUCs of 0.940 and 0.955 for mild and advanced staged PCa, respectively, suggesting a clear path for translation into clinically meaningful tools.

RESULTS & CONCLUSIONS

Following data processing and z-score transofrmation on the combined proteomic and lipidomic datasets, Random Forest/Boruta were performed prior to training and cross validation of the model. A total of 1115 features (combination of proteins & lipids) were used for the random forest classification. The output of the machine learning models applied to these data are summarised below for mild PCa vs. healthy controls (Figure 3) and advanced PCa vs. healthy controls (Figure 4). In both cases, a variety of proteins and lipids were identified as being discriminating features for categorising patients with either mild or advanced PCa. Enrichment and network analysis was also conducted on the features identified via Boruta/Random Forest for both PCa groups (Figure 5). In order to minimise errors in the models constructed, cross testing was performed. The percentage distribution of missingness across all samples in the lipidomic and proteomic data, indicated very low levels of missigness (Figure 6). Finally, Receiver operating characterising (ROC) curves were generated to highlight the discriminating power based on the significant features identified in both the mild and advanced PCa groups (Figure 7).

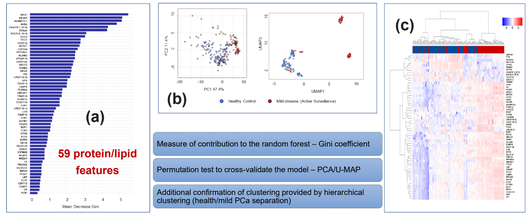

Figure 3. Mild PCa vs. Healthy Controls: Machine learning models resulted in 59 protein/lipid features as being the most significant, ranked by Gini coefficient. The Mean Decrease Gini (importance) score of all molecules contributing to the construction of the predictive model are shown in (a), with KIF5C, KIF28P, ADAMTSL1, SHBG. Permutation tests to cross validate the model were conducted in the form of unsupervised PCA and U-MAP (b), shows distinct separation between both groups. Hierarchial custering (c) further highlights distinct clusters between mild PCa and healthy controls.

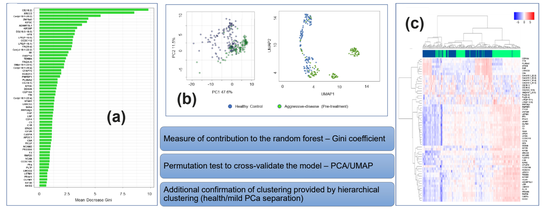

Figure 4. Advanced PCa vs. Healthy Controls: Machine learning models resulted in a larger number of lipid features as being significant (Gini coefficient) when compared with the mild PCa analysis. The Mean Decrease Gini (importance) score of all molecules contributing to the construction of the predictive model are shown in (a), with CE(18:2) and XRCC5 being the highest ranking. Permutation tests to cross validate the model were conducted in the form of unsupervised PCA and U-MAP (b), shows distinct separation between both groups. Hierarchial custering (c) further highlights distinct clusters between advanced PCa and healthy controls.

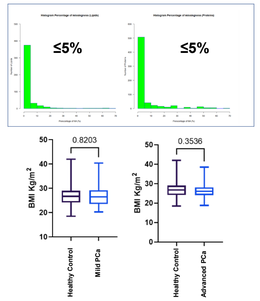

Figure 6. In order to avoid potentially amplifying any noise in the dataset and potentially overfitting the models, the distribution of missing values was assessed (upper bar chart) for the total number of proteins and lipids used. The percentage of missingness calculated within the datasets was <5% in both cases. Body Mass Index (BMI) which can be a potential co-founder linked with PCa was also assessed to ensure that the group separations observed, are not influenced by BMI. Evaluation of the patient data indicates that the mean BMI across all groups is similar and therefore no BMI adjustment was necessary.

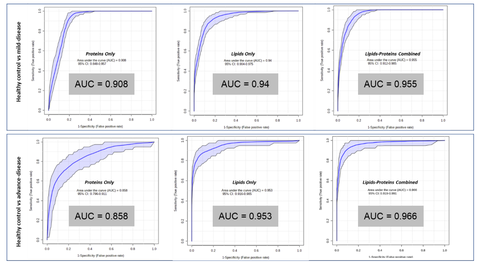

Figure 7. A comparative analysis of Area under ROC curve of the model built using raw protein quantification data, raw lipid quantification data and z-score normalised integrated proteomic-lipidomic dataset for both mild (upper plots) and advanced (lower plots) PCa.

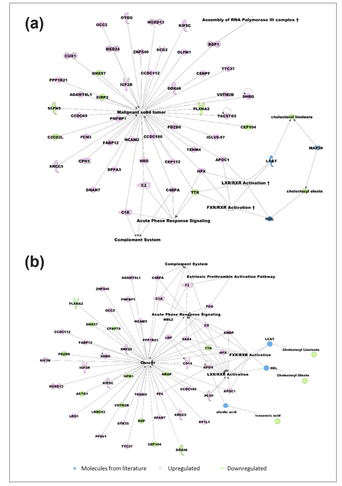

Figure 5. Enrichment analysis of Boruta-identified features for mild (a) and advanced (b) PCa. Under expressed molecules are represented in green, whilst those which are over expressed, are presented in pink. Molecules with no direct link to the network were excluded from the analysis.

TO DOWNLOAD A COPY OF THIS POSTER, VISIT WWW.WATERS.COM/POSTERS