EBioMedicine 17 (2017) 5–6

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

In Focus

# Mining Human Prostate Cancer Datasets: The "camcAPP" Shiny App☆

Mark J. Dunning [a], Sarah L. Vowler [a], Emilie Lalonde [b,c], Helen Ross-Adams [d], Paul Boutros [b,c], Ian G. Mills [e], Andy G. Lynch [a], Alastair D. Lamb [a,f,g,]*

[a] Cancer Research UK Cambridge Institute, Cambridge Biomedical Campus, UK
[b] Department of Medical Biophysics, University of Toronto, Canada
[c] Informatics and Bio-Computing Program, Ontario Institute for Cancer Research, Toronto, Canada
[d] Bioinformatics Unit, Centre for Molecular Oncology, Barts Cancer Institute, Queen Mary University London, UK
[e] Centre for Cancer Research and Cell Biology (CCRCB), Queen's University of Belfast, 97 Lisburn Road, Belfast, UK
[f] Department of Genito-urinary Oncology, Peter MacCallum Cancer Centre, Melbourne, Australia
[g] Academic Urology Group, Department of Surgery, University of Cambridge, UK

ARTICLE INFO

*Article history:*
Received 22 February 2017
Accepted 22 February 2017
Available online 2 March 2017

*Keywords:*
Prostate cancer
Gene signature
Prognostication
Online interface
Big data

Obtaining access to robust, well-annotated human genomic datasets is an important step in demonstrating the relevance of experimental findings and, often, in generating the hypotheses that led to those experiments being conducted in the first place. We recently published data from the CamCaP Study Group which comprised two cohorts of men with prostate cancer who had undergone prostatectomy in Cambridge, UK and Stockholm, Sweden (Ross-Adams et al., 2015). We considered how we might best share our output with those who wish to interrogate the data with their own ideas, gene lists and clinical questions. We recognised that finding, down-loading, pre-processing and assimilating any such dataset into a usable format is daunting and may put off many researchers. We also felt that interrogation tools generated to date (*e.g. cBioPortal*) lack functionality as they either cover too many organ types, or are limited in the extent, precision and tumour-site specificity of their clinical annotation. We therefore determined to produce an accessible web-based platform that would permit straightforward interrogation of these datasets with individual gene identifiers or gene sets. Furthermore, we decided to include additional 'publicly-accessible' human prostate cancer sets in order to increase the number of samples available and provide a degree of validation of any observations made across independent cohorts. We included a number of prominent publicly available sets with both gene expression and copy number data leading to a cohort of almost 500 men (Ross-Adams et al., 2015; Taylor et al., 2010; Grasso et al., 2012). We also included a small landmark series of expression data (Varambally et al., 2005). These studies are summarised in Table 1. We plan to include additional studies in the app as well-annotated datasets become publicly available.

An important finding in our recent study was that prostate cancer could be divided into five distinct molecular subgroups based on stratification with a small number of copy number features which were also associated with RNA-expression change. These groups had different clinical outcomes. We wanted the app to allow researchers to determine the mean RNA-expression level or copy number status of a single gene or gene-set in prostates from men divided either according to clinical categories (Gleason score, biochemical relapse status or tumour type) or according to molecular subgroups. These subgroups could either be pre-defined molecular groups published in the relevant papers, or *de novo* subgroups generated by hierarchical clustering based on an uploaded geneset.

We searched for other web-tools that are already available for this purpose. Although no such site exists for assessment of subgroup patterns or combined expression and copy number profiles, the Memorial Sloane Kettering Cancer Centre (MSKCC) and Michigan data (Table 1) can be analysed as part of cBioPortal (cBioPortal for Cancer Genomics, n.d.) along with the recently published prostate TCGA dataset (Robinson et al., 2015).

Here we introduce the camcAPP (http://bioinformatics.cruk.cam.ac.uk/apps/camcAPP/); a bespoke web interface to multiple prostate cancer genomics datasets. The interface was created with Shiny (https://www.rstudio.com/products/shiny/), and allows the non-specialist Bioinformatician to create publication-ready figures and tables through an intuitive interface to the underlying computer code.

After selecting a dataset of interest, and uploading a list of genes, the following analyses can be performed:

1) Boxplots and analysis of variance for expression of genes of interest grouped by clinical group, sample type, Gleason grade of copy-

☆ Web address: bioinformatics.cruk.cam.ac.uk/apps/camcAPP/.
* Corresponding author.
*E-mail address:* Alastair.Lamb@cruk.cam.ac.uk (A.D. Lamb).

http://dx.doi.org/10.1016/j.ebiom.2017.02.022

**Table 1**
Summary of studies included in the camcAPP at initial release. Primary Tumours = tissue taken from radical prostatectomy specimens in men with confirmed organ-confined disease. Advanced Tumours = tissue from channel transurethral resection of the prostate (chTURP) or prostatectomy in men with metastatic disease.

| Dataset | Paper | Platform: gene expression | Platform: copy number | Primary tumours | Advanced tumours | Clinical covariates |
|---|---|---|---|---|---|---|
| Michigan 2005 | Varambally et al. (2005)<br>GSE3325 | Affymetrix U133 2.0 | N/A | 7 | 6 | Sample Group |
| MSKCC 2010 | Taylor et al. (2010)<br>GSE21032 | Affymetrix Human 1.0 ST | Agilent 244k | 109 | 19 | Gleason, Copy Number Cluster |
| Michigan 2012 | Grasso et al. (2012)<br>GSE35988 | Agilent Whole Human 44k | Agilent 105k/244k | 59 | 32 | Sample Group |
| Stockholm 2015 | Ross-Adams et al. (2015)<br>GSE70770 | Illumina HT12 | Affymetrix SNP 6.0 | 101 | N/A | iCluster, Sample Group |
| Cambridge 2015 | Ross-Adams et al. (2015)<br>GSE70770 | Illumina HT12 | Illumina Omni 2.5 | 125 | 19 | iCluster, Gleason, Sample Group |

number cluster (N.B. not all covariates available for all data sets) (see Supplementary Fig. 2).

2) A recursive partitioning-based survival analysis and Kaplan-Meier plots on a gene-by-gene basis (Supplementary Fig. 3).
3) Pairwise-correlations of gene expression across whole studies and within clinical subgroups.
4) Clustering and heatmaps of gene expression data, with options to interrogate associations between clinical covariates and newly-derived clusters (Supplementary Fig. 4).
5) Tabulating the number of copy-number amplifications and deletions observed across whole studies or within a particular clinical covariate, and making a heatmap of copy-number calls (Supplementary Fig. 5) (Lalonde et al., 2014).

One of the challenges in constructing such a tool is delivering an output format that is readily transferable to slides for presentations or panels of a figure for publication. We recognise that this is, in part, a matter of axis typesetting and plot configuration but also of delivering an output file which permits further adjustment of the figure in, for example, Adobe Illustrator™. To this end, all plots can be exported as PDF or PNG files with configurable dimensions. Furthermore, for those that are well-versed in R, the code to produce a particular plot can be downloaded and modified as required. A further challenge that we seek to address with this interface is merging datasets for combined analysis. We hope to offer this option in due course, as we include further datasets that include samples analysed on compatible platforms.

Strategies to address the Big Data problem have focussed on making the ever-increasing volume of genomic data accessible to scientists and on opening up the possibility of engaging non-specialists (Keener, 2015). This approach embodies a responsible attitude to science both in terms of patient input and financial resource and we believe that tools such as this are an important step to maximising the value of these landmark studies. We take pleasure in making this platform available to the prostate cancer community by means of this 'In Focus' article in EBioMedicine, a journal that we believe champions this responsible approach to genomic data both in cancer genomics (Kerns et al., 2016) and further afield (Taudien et al., 2016).

## Funding

Core CRUK funding: MD, AGL, ADL.

Academy of Medical Sciences Clinical Lecturer Starter Grant SGCL11 (prinicipal funder of this work): ADL.

No other funding source to declare. The funding bodies have not been involved in any way in the design, completion or writing of the project.

## Authors' contribution

Study concept: MD, AGL, ADL.
Study design: MD, AGL, ADL.
CamCaP Study Group leads: ADL, HRA.
Data programming: MD, SV, AGL.
Programme contributions: EL.
Beta testing: IGM, EL, PB.
Oversight: AGL, ADL.

## Disclosure

The authors have no conflicts of interest to declare.

## Acknowledgements

The authors would like to thank Ola Bratt, Christof Kastner, Rajesh Nair and Declan Murphy for their critical and creative advice in review of the manuscript and/or design. We also acknowledge David Neal as the lead and principal investigator for the CamCaP Study Group.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ebiom.2017.02.022. This includes a 'manual' for the Shiny App which can also be downloaded from the app itself.

## References

cBioPortal for Cancer Genomics. Memorial Sloane Kettering Cancer Centre, www.cbioportal.org, (Accessed: 23/03/2016).

Grasso, C.S., Wu, Y.M., Robinson, D.R., Cao, X., Dhanasekaran, S.M., Khan, A.P., et al., 2012. The mutational landscape of lethal castration-resistant prostate cancer. Nature 487 (7406), 239–243.

Keener, Amanda B., July 8, 2015. The Scientist http://www.the-scientist.com/?articles.view/articleNo/43483/title/Big-Data-Problem/ (Accessed: 23/3/16).

Kerns, S.L., Dorling, L., Fachal, L., Bentzen, S., Pharoah, P.D., Barnes, D.R., et al., 2016. Meta-analysis of genome wide association studies identifies genetic markers of late toxicity following radiotherapy for prostate cancer: radiogenomics consortium. EBioMedicine 10:150–163. http://dx.doi.org/10.1016/j.ebiom.2016.07.022.

Lalonde, E., Ishkanian, A.S., Sykes, J., Fraser, M., Ross-Adams, H., Erho, N., et al., 2014. Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. Lancet Oncol. 15 (13), 1521–1532.

Robinson, D., Van Allen, E.M., Wu, Y.M., Schultz, N., Lonigro, R.J., Mosquera, J.M., et al., 2015. Integrative clinical genomics of advanced prostate cancer. Cell 161 (5), 1215–1228.

Ross-Adams, H., Lamb, A.D., Dunning, M.J., Halim, S., Lindberg, J., Massie, C.M., et al., 2015. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: a discovery and validation cohort study. EBioMedicine 2 (9), 1133–1144.

Taudien, S., Lausser, L., Giamarellos-Bourboulis, E.J., Sponholz, C., Schöneweck, F., Felder, M., et al., 2016. Genetic factors of the disease course after sepsis: rare deleterious variants are predictive. EBioMedicine 12:227–238. http://dx.doi.org/10.1016/j.ebiom.2016.08.037.

Taylor, B.S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B.S., et al., 2010. Integrative genomic profiling of human prostate cancer. Cancer Cell 18 (1), 11–22.

Varambally, S., Yu, J., Laxman, B., Rhodes, D.R., Mehra, R., Tomlins, S.A., et al., 2005. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Cancer Cell 8 (5), 393–406.