Mining human prostate cancer datasets: The "camcAPP" shiny app

Mark J Dunning[1], Sarah L Vowler[1], Emilie Lalonde[2,3], Helen Ross-Adams[4], Ian G Mills[5], Andy G Lynch[1], Alastair D Lamb[1,6]

[1]Cancer Research UK Cambridge Institute, Cambridge Biomedical Campus, UK

[2]Department of Medical Biophysics, University of Toronto, Canada

[3]Informatics and Bio-Computing Program, Ontario Institute for Cancer Research, Toronto, Canada

[4]Bioinformatics Unit, Centre for Molecular Oncology, Barts Cancer Institute, Queen Mary University London

[5] Centre for Cancer Research and Cell Biology (CCRCB), Queen's University of Belfast, 97 Lisburn Road, Belfast

[6]Department of Genito-urinary Oncology, Peter MacCallum Cancer Centre, Melbourne, Australia

Keywords:

Prostate cancer; gene signature; prognostication; online interface; big data

Word Count:

968 words

Webpage link to camcAPP shiny app:

bioinformatics.cruk.cam.ac.uk/apps/camcAPP/

[suggest placing in box above / middle of editorial – and hyperlink in online version]

Obtaining access to robust, well-annotated human genomic datasets is an important step in demonstrating 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 two cohorts of men with prostate cancer who had undergone prostatectomy, men from Cambridge, UK and Stockholm, Sweden.[1] 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, processing and assimilating any such dataset into a usable format is daunting and may put off even the more persistent researcher. We also felt that interrogation tools generated to date (e.g. cBioPortal) lack functionality as they either cover too many organ types, or fall short in terms of 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.[1-3] We also included a small landmark series of expression data.[4] These studies are summarised in **Table 1**. We plan to include additional studies in the app as well-annotated datasets become publicly available.

Dataset	**Paper**	**Platform: Gene Expression**	**Platform: Copy Number**	**Primary Tumours**	**Advanced Tumours**
Michigan 2005	Varambally et al (2005)	Affymetrix U133 2.0	N /A	7	6
MSKCC 2010	Taylor et al (2010)	Affymetrix Human 1.0 ST	Agilent 244k	109	19
Michigan 2012	Grasso et al (2012)	Agilent Whole Human 44k	Agilent 105k / 244k	59	32
Stockholm 2015	Ross-Adams et al (2015)	Illumina HT12	Affymetrix SNP 6.0	101	N / A
Cambridge 2015	Ross-Adams et al (2015)	Illumina HT12	Illumina Omni 2.5	125	19

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.

An important finding in our recent study was that prostate cancer could be divided into five distinct molecular subgroups based on stratification by a small number of copy number features which were also associated with expression change. These groups had different clinical outcomes. We wanted the app to allow researchers to determine the mean expression level or copy number status of a single gene or geneset 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 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[5] along with the recently published prostate TCGA dataset.[6]

Here we introduce the camcAPP. This is implemented as a Shiny[7] application in R[8]. Shiny allows the non-specialist bioinformatician to create publication-ready figures and tables through an intuitive interface to the underlying R code. The source code for the entire app is also available through github[9] (**https://github.com/crukci-bioinformatics/camcAPP).** The dplyr[10] package is used throughout for efficient data manipulation, and graphics are generated using ggplot2[11] . The datasets themselves are available via Gene Expression Omnibus. Using the GEOquery[12] package, we downloaded the datasets and converted them into a format compatible with the Bioconductor[13] project. The resulting packages are available as experimental data packages in Bioconductor.

After selecting a dataset of interest, and uploading a list of genes (**Figure 1**), the following analyses can be performed:-

1) Creation of boxplots to visualise the distribution of expression values of each gene for different clinical covariates of interest (**Figure 2**).
2) A survival analysis to assess whether the expression level of each gene can predict relapse (**Figure 1**). The "party" R package[14] is first used to see if the expression level can be split into two distinct groups. If such a separation can be found, then a Kaplan-Meier curve is generated from the survival times of samples in the different groups. If no significant separation of samples is found, the median expression level is used to define groups of samples with high or low expression.
3) Pairwise correlations of the expression level of all specified genes.
4) Construction of a heatmap to assess whether the chosen genes can split the dataset of interest into subgroups. Various methods of clustering and visualisation are supported.
5) Tabulating the number of copy-number amplifications and deletions observed in the Cambridge, Stockholm and MSKCC cohorts, and making a heatmap of copy-number calls (**Figure 2**).[15]

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™. Thus 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.[16] 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.

Figure 1. Examples of "camcaPP" shiny app functions – Gene-list input and Survival analysis.

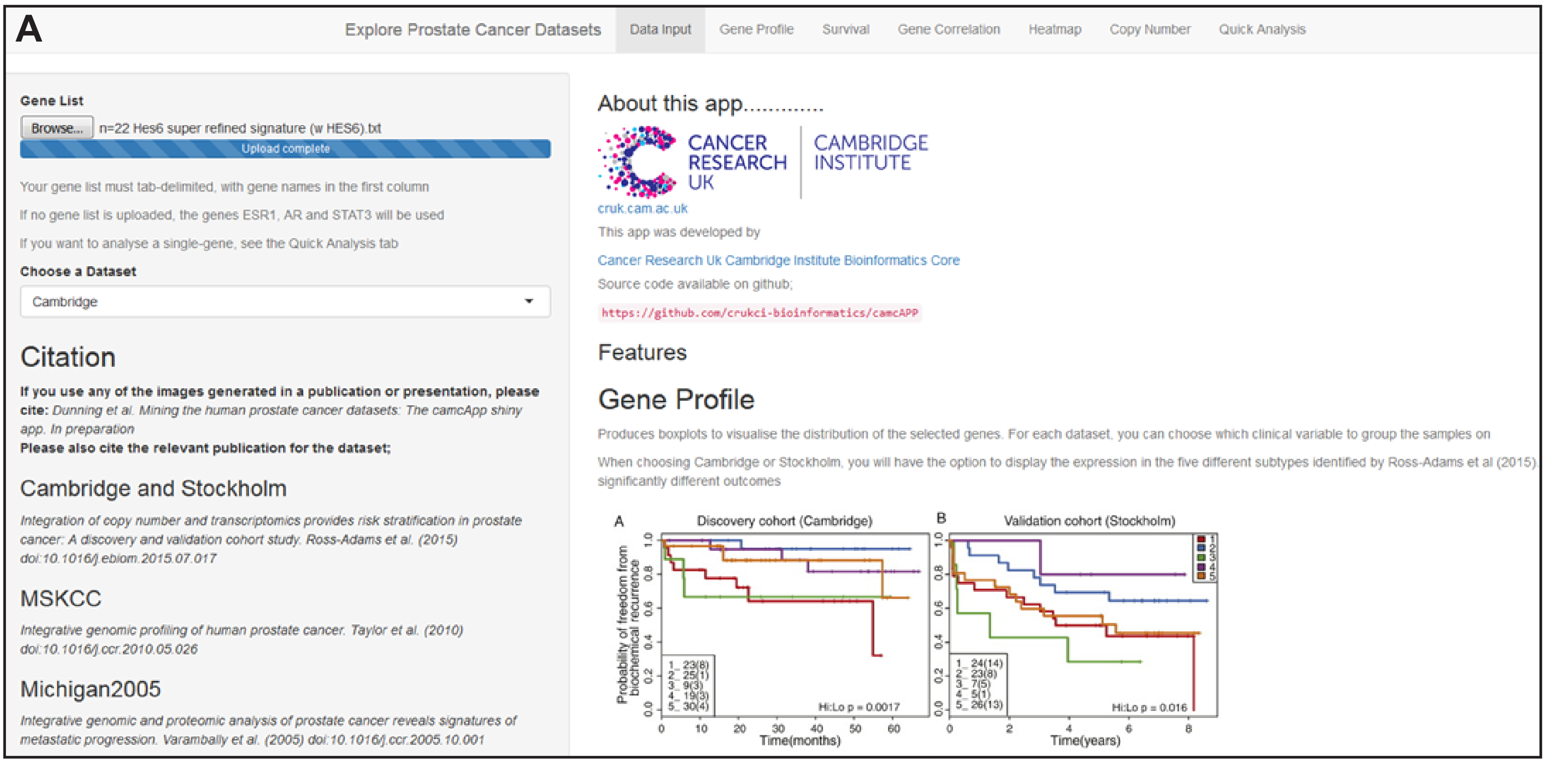
A. Entry point to site including upload point for gene lists and citation information. **B**. Kaplan-Meier biochemical relapse–free survival plots can be created for any selected gene from the input list in any selected dataset. Example of .pdf output modified in Adobe Illustrator^{TM} demonstrated in panel **A** (Ross-Adams et al).[1]

Figure 2. Examples of "camcaPP" shiny app functions – Gene expression and Copy Number plots.

A. Boxplots for gene expression can be created for a list of genes – in this example the five Cambridge subgroups are demonstrated for a number of input genes. **B**. Heatmaps for gene expression can be created from any of the datasets. **C.** Copy number (CN) plots depicting CN gain or loss (or no change = neutral) can also be created. **D.** "Quick Analysis" tab allows rapid spot checks for single genes of interest focussing on relative expression, copy number and survival curves (example shown here is a boxplot showing relative gene expression in CRPC, tumour and benign tissue for HES6, a known driver of castration resistance[17])

References

1. Ross-Adams H, Lamb AD, Dunning MJ, Halim S, Lindberg J, Massie CM, et al. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. *EBioMedicine* 2015;2(9):1133-44. ***GSE70770***
2. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. *Cancer cell* 2010;18(1):11-22. ***GSE21032***
3. Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, et al. The mutational landscape of lethal castration-resistant prostate cancer. *Nature* 2012;487(7406):239-43. ***GSE35988***
4. Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, Tomlins SA, et al. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer cell* 2005;8(5):393-406. ***GSE3325***
5. www.cbioportal.org. cBioPortal for Cancer Genomics. Memorial Sloane Kettering Cancer Centre. , Accessed: 23/03/2016.
6. Robinson D, Van Allen EM, Wu YM, Schultz N, Lonigro RJ, Mosquera JM, et al. Integrative clinical genomics of advanced prostate cancer. *Cell* 2015;161(5):1215-28.
7. Chang W, Cheng J, Allaire JJ, et al.: shiny: Web Application Framework for R. R package version 0.11.1. 2015.
8. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2014.
9. https://github.com/crukci-bioinformatics/camcAPP.
10. Hadley Wickham and Romain Francois (2016). dplyr: A Grammar of Data Manipulation. R package version 0.5.0. https://CRAN.R-project.org/package=dplyr.
11. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.
12. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 2007;23(14):1846-7.
13. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods* 2015;12(2):115-21.
14. Torsten Hothorn, Kurt Hornik and Achim Zeileis (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651--674.
15. Lalonde E, Ishkanian AS, Sykes J, Fraser M, Ross-Adams H, Erho N, et al. Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. *The Lancet. Oncology* 2014;15(13):1521-32.
16. http://www.the-scientist.com/?articles.view/articleNo/43483/title/Big-Data-Problem/. Amanda B. Keener. July 8, 2015. . *The Scientist.* Accessed: 23/3/16.
17. Ramos-Montoya A, Lamb AD, Russell R, Carroll T, Jurmeister S, Galeano-Dalmau N, et al. HES6 drives a critical AR transcriptional programme to induce castration-resistant prostate cancer through activation of an E2F1-mediated cell cycle network. *EMBO molecular medicine* 2014;6(5):651-61.



Gene List

n=22 Hes6 super refined signature (w HES6).txt

Your gene list must tab-delimited, with gene names in the first column

If no gene list is uploaded, the genes ESR1, AR and STAT3 will be used

If you want to analyse a single-gene, see the Quick Analysis tab

Choose a Dataset

Cambridge

Citation

If you use any of the images generated in a publication or presentation, please cite: *Dunning et al. Mining the human prostate cancer datasets: The camcApp shiny app. In preparation*
Please also cite the relevant publication for the dataset;

Cambridge and Stockholm

Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. Ross-Adams et al. (2015) doi:10.1016/j.ebiom.2015.07.017

MSKCC

Integrative genomic profiling of human prostate cancer. Taylor et al. (2010) doi:10.1016/j.ccr.2010.05.026

Michigan2005

Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Varambally et al. (2005) doi:10.1016/j.ccr.2005.10.001

About this app............

cruk.cam.ac.uk

This app was developed by

Cancer Research Uk Cambridge Institute Bioinformatics Core

Source code available on github;

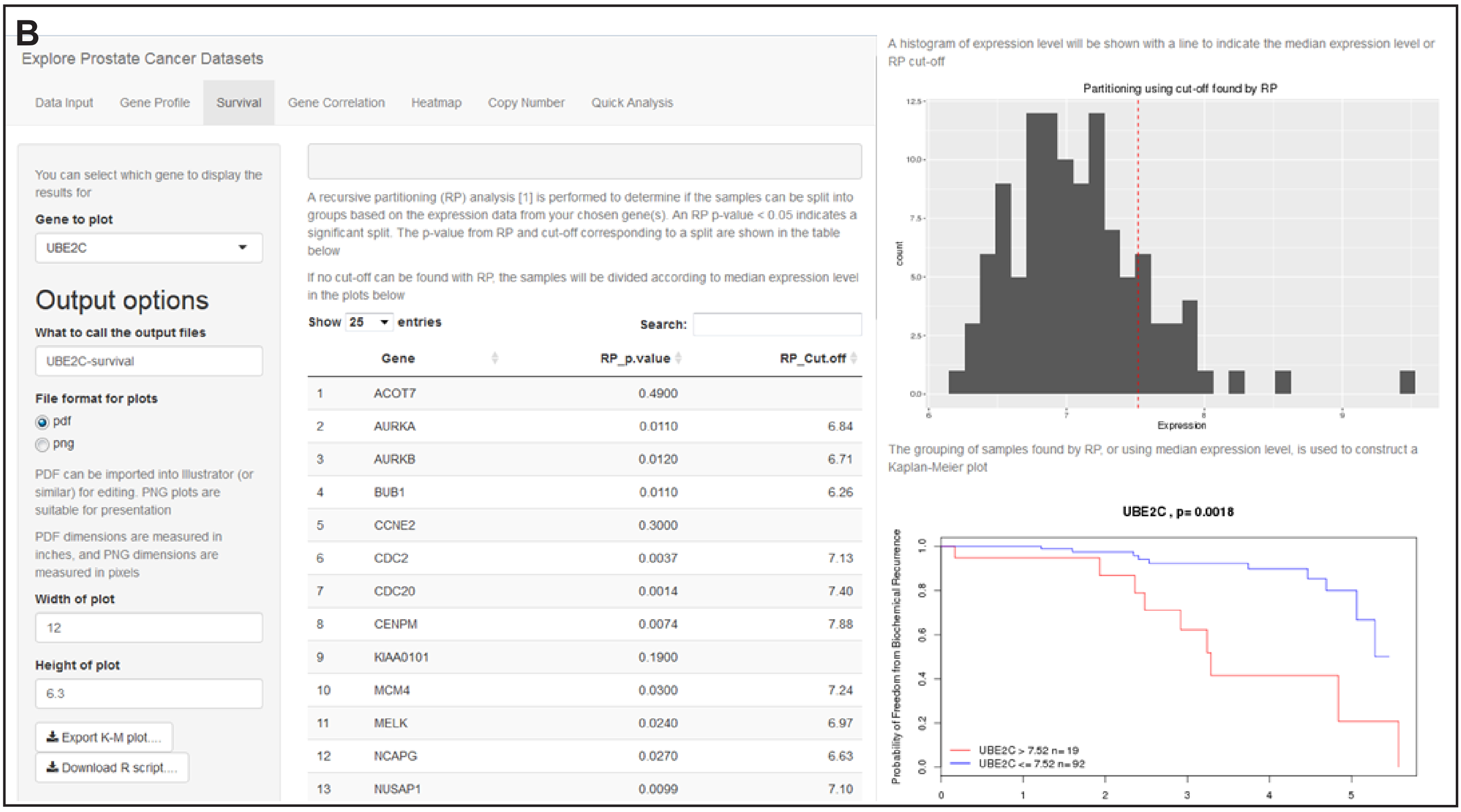
`https://github.com/crukci-bioinformatics/camcAPP`

Features

Gene Profile

Produces boxplots to visualise the distribution of the selected genes. For each dataset, you can choose which clinical variable to group the samples on

When choosing Cambridge or Stockholm, you will have the option to display the expression in the five different subtypes identified by Ross-Adams et al (2015) significantly different outcomes



You can select which gene to display the results for

Gene to plot

UBE2C

Output options

What to call the output files

UBE2C-survival

File format for plots

● pdf
○ png

PDF can be imported into Illustrator (or similar) for editing. PNG plots are suitable for presentation

PDF dimensions are measured in inches, and PNG dimensions are measured in pixels

Width of plot

12

Height of plot

6.3

A recursive partitioning (RP) analysis [1] is performed to determine if the samples can be split into groups based on the expression data from your chosen gene(s). An RP p-value < 0.05 indicates a significant split. The p-value from RP and cut-off corresponding to a split are shown in the table below

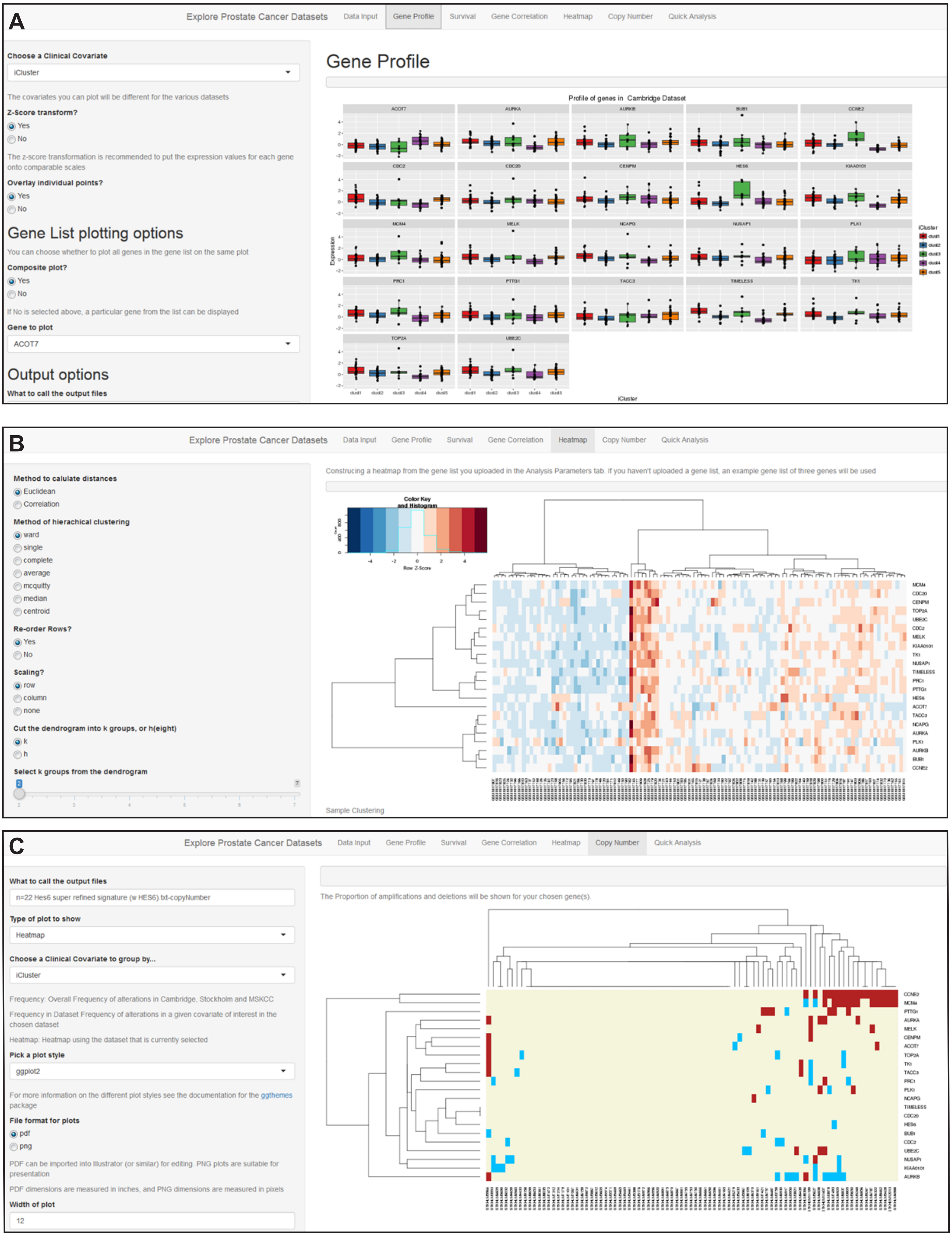
If no cut-off can be found with RP, the samples will be divided according to median expression level in the plots below

Show 25 **entries** **Search:**

	Gene	RP_p.value	RP_Cut.off
1	ACOT7	0.4900	
2	AURKA	0.0110	6.84
3	AURKB	0.0120	6.71
4	BUB1	0.0110	6.26
5	CCNE2	0.3000	
6	CDC2	0.0037	7.13
7	CDC20	0.0014	7.40
8	CENPM	0.0074	7.88
9	KIAA0101	0.1900	
10	MCM4	0.0300	7.24
11	MELK	0.0240	6.97
12	NCAPG	0.0270	6.63
13	NUSAP1	0.0099	7.10

A histogram of expression level will be shown with a line to indicate the median expression level or RP cut-off

The grouping of samples found by RP, or using median expression level, is used to construct a Kaplan-Meier plot



Choose a Clinical Covariate

iCluster

The covariates you can plot will be different for the various datasets

Z-Score transform?

● Yes
○ No

The z-score transformation is recommended to put the expression values for each gene onto comparable scales

Overlay individual points?

● Yes
○ No

Gene List plotting options

You can choose whether to plot all genes in the gene list on the same plot

Composite plot?

● Yes
○ No

If No is selected above, a particular gene from the list can be displayed

Gene to plot

ACOT7

Output options

What to call the output files

Gene Profile

Method to calulate distances

● Euclidean
○ Correlation

Method of hierachical clustering

● ward
○ single
○ complete
○ average
○ mcquitty
○ median
○ centroid

Re-order Rows?

● Yes
○ No

Scaling?

● row
○ column
○ none

Cut the dendrogram into k groups, or h(eight)

● k
○ h

Select k groups from the dendrogram

2

Construcing a heatmap from the gene list you uploaded in the Analysis Parameters tab. If you haven't uploaded a gene list, an example gene list of three genes will be used

Sample Clustering

What to call the output files

n=22 Hes6 super refined signature (w HES6).txt-copyNumber

Type of plot to show

Heatmap

Choose a Clinical Covariate to group by...

iCluster

Frequency: Overall Frequency of alterations in Cambridge, Stockholm and MSKCC

Frequency in Dataset Frequency of alterations in a given covariate of interest in the chosen dataset

Heatmap: Heatmap using the dataset that is currently selected

Pick a plot style

ggplot2

For more information on the different plot styles see the documentation for the ggthemes package

File format for plots

● pdf
○ png

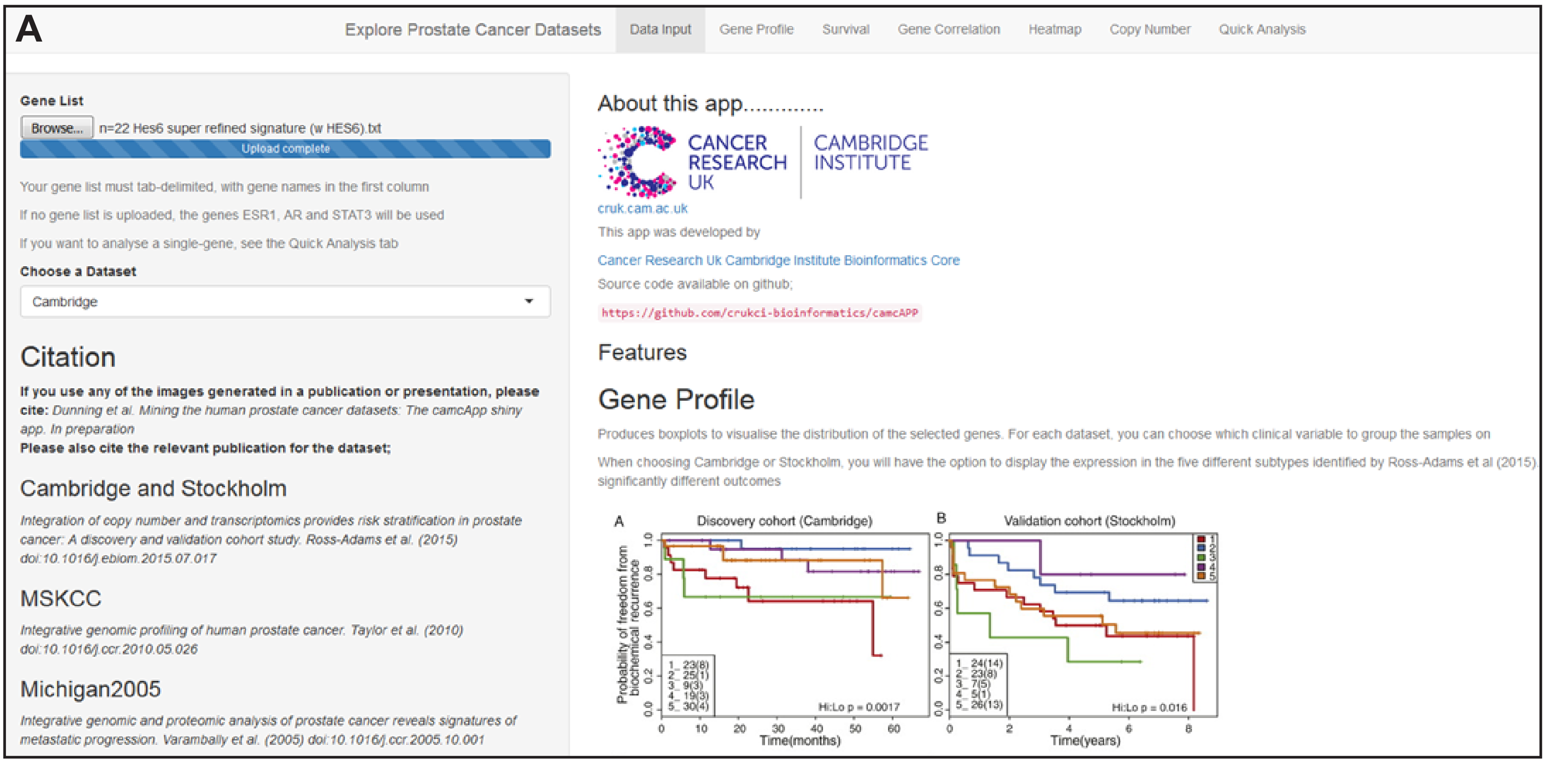
PDF can be imported into Illustrator (or similar) for editing. PNG plots are suitable for presentation

PDF dimensions are measured in inches, and PNG dimensions are measured in pixels

Width of plot

12

The Proportion of amplifications and deletions will be shown for your chosen gene(s).



Gene List

n=22 Hes6 super refined signature (w HES6).txt

Your gene list must tab-delimited, with gene names in the first column

If no gene list is uploaded, the genes ESR1, AR and STAT3 will be used

If you want to analyse a single-gene, see the Quick Analysis tab

Choose a Dataset

Cambridge

Citation

If you use any of the images generated in a publication or presentation, please cite: *Dunning et al. Mining the human prostate cancer datasets: The camcApp shiny app. In preparation*
Please also cite the relevant publication for the dataset;

Cambridge and Stockholm

Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. Ross-Adams et al. (2015) doi:10.1016/j.ebiom.2015.07.017

MSKCC

Integrative genomic profiling of human prostate cancer. Taylor et al. (2010) doi:10.1016/j.ccr.2010.05.026

Michigan2005

Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Varambally et al. (2005) doi:10.1016/j.ccr.2005.10.001

About this app............

cruk.cam.ac.uk

This app was developed by

Cancer Research Uk Cambridge Institute Bioinformatics Core

Source code available on github;

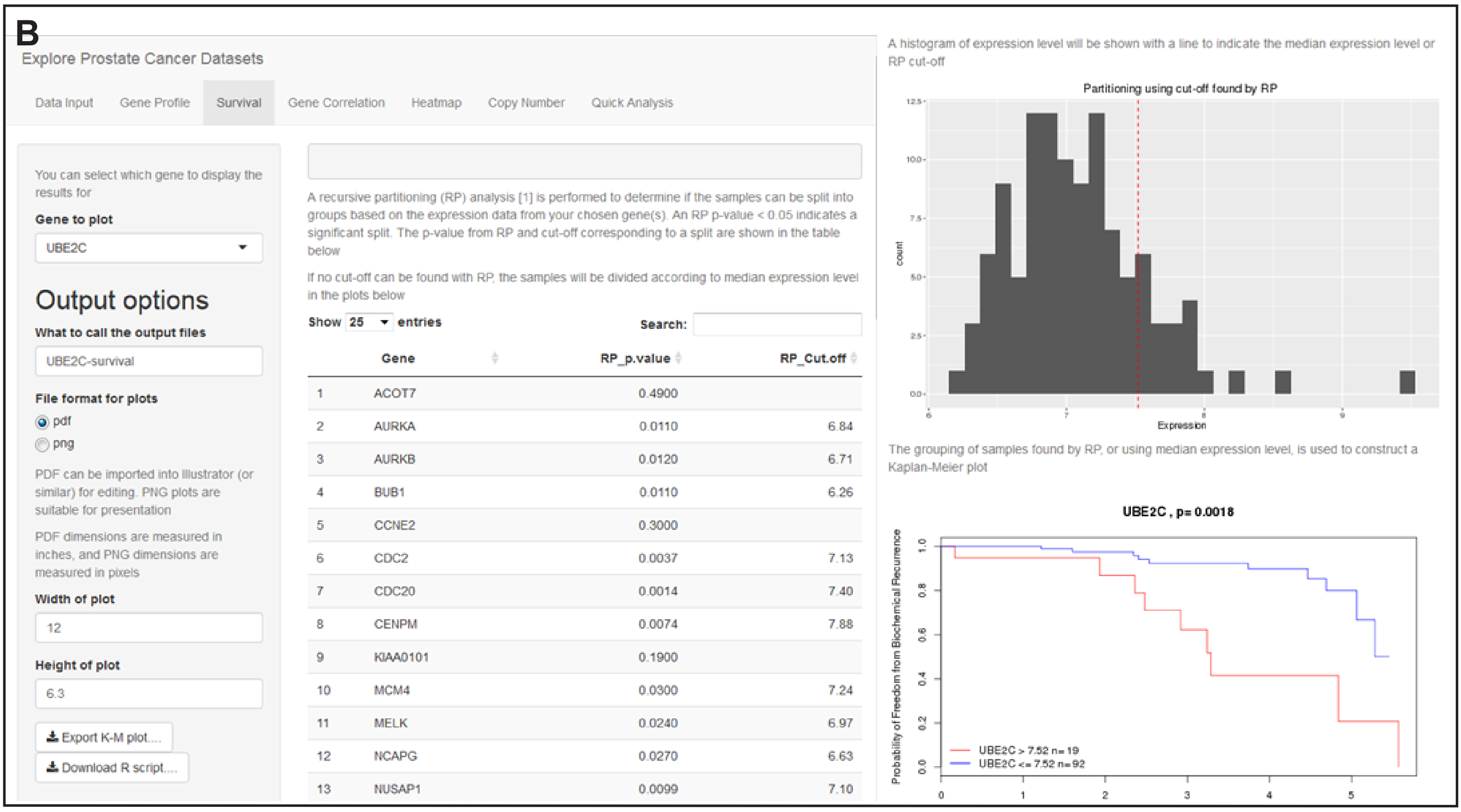
`https://github.com/crukci-bioinformatics/camcAPP`

Features

Gene Profile

Produces boxplots to visualise the distribution of the selected genes. For each dataset, you can choose which clinical variable to group the samples on

When choosing Cambridge or Stockholm, you will have the option to display the expression in the five different subtypes identified by Ross-Adams et al (2015) significantly different outcomes



You can select which gene to display the results for

Gene to plot

UBE2C

Output options

What to call the output files

UBE2C-survival

File format for plots

● pdf
○ png

PDF can be imported into Illustrator (or similar) for editing. PNG plots are suitable for presentation

PDF dimensions are measured in inches, and PNG dimensions are measured in pixels

Width of plot

12

Height of plot

6.3

A recursive partitioning (RP) analysis [1] is performed to determine if the samples can be split into groups based on the expression data from your chosen gene(s). An RP p-value < 0.05 indicates a significant split. The p-value from RP and cut-off corresponding to a split are shown in the table below

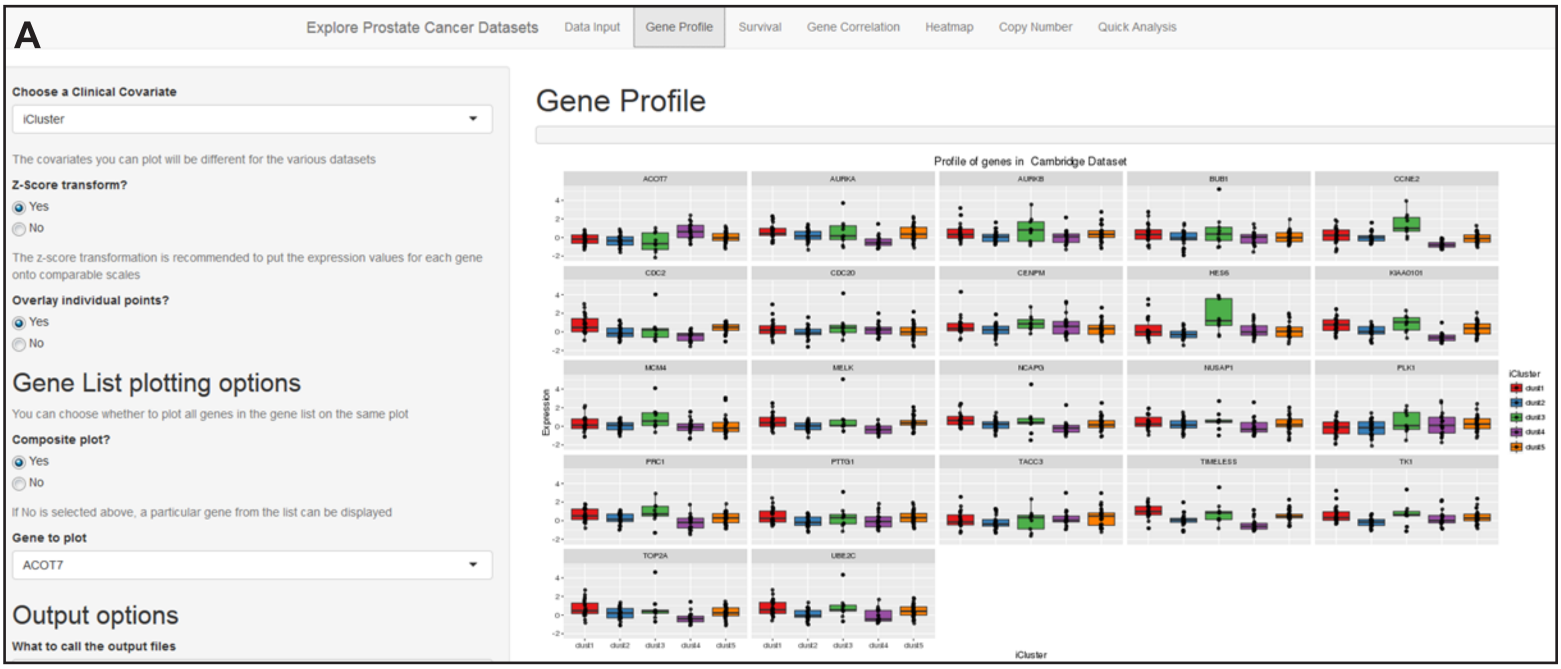
If no cut-off can be found with RP, the samples will be divided according to median expression level in the plots below

Show 25 **entries** **Search:**

	Gene	RP_p.value	RP_Cut.off
1	ACOT7	0.4900	
2	AURKA	0.0110	6.84
3	AURKB	0.0120	6.71
4	BUB1	0.0110	6.26
5	CCNE2	0.3000	
6	CDC2	0.0037	7.13
7	CDC20	0.0014	7.40
8	CENPM	0.0074	7.88
9	KIAA0101	0.1900	
10	MCM4	0.0300	7.24
11	MELK	0.0240	6.97
12	NCAPG	0.0270	6.63
13	NUSAP1	0.0099	7.10

A histogram of expression level will be shown with a line to indicate the median expression level or RP cut-off

The grouping of samples found by RP, or using median expression level, is used to construct a Kaplan-Meier plot



Choose a Clinical Covariate

iCluster

The covariates you can plot will be different for the various datasets

Z-Score transform?

● Yes

○ No

The z-score transformation is recommended to put the expression values for each gene onto comparable scales

Overlay individual points?

● Yes

○ No

Gene List plotting options

You can choose whether to plot all genes in the gene list on the same plot

Composite plot?

● Yes

○ No

If No is selected above, a particular gene from the list can be displayed

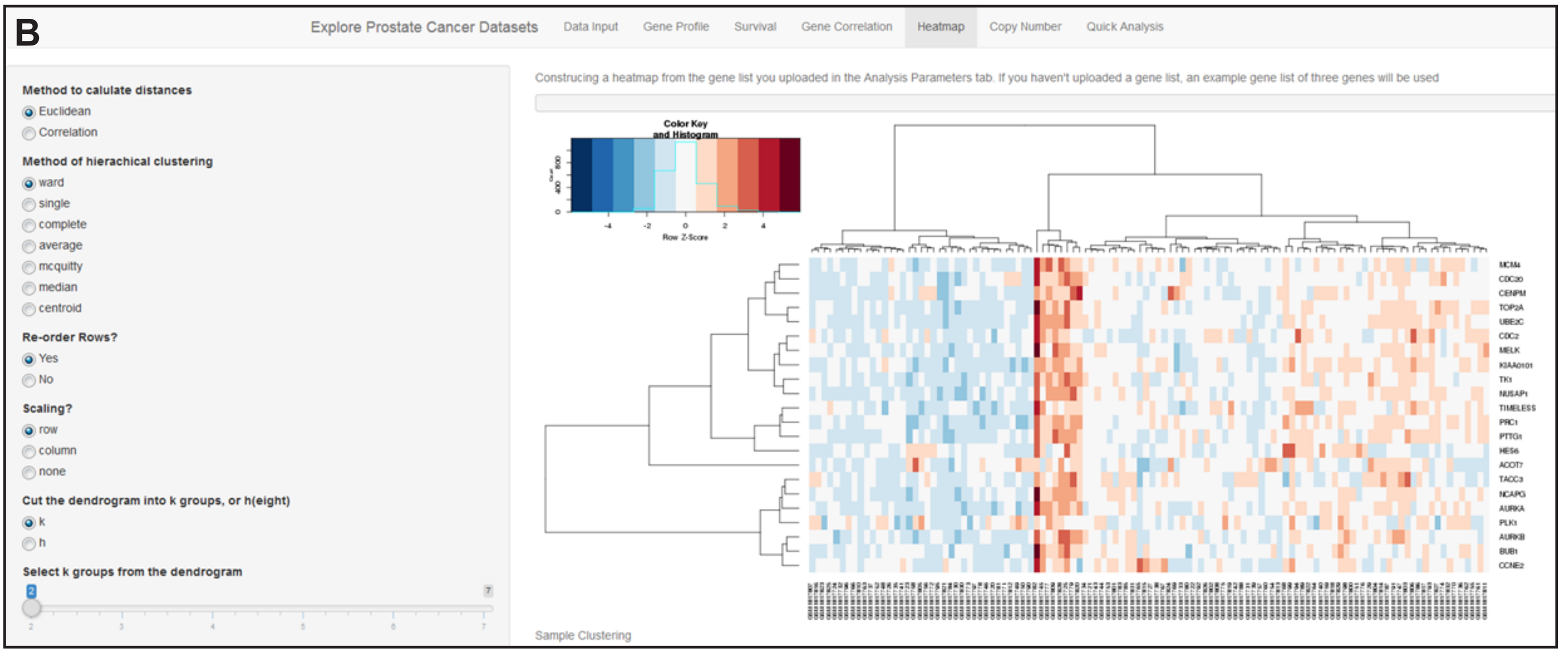
Gene to plot

ACOT7

Output options

What to call the output files

Gene Profile



Method to calulate distances

● Euclidean

○ Correlation

Method of hierachical clustering

● ward

○ single

○ complete

○ average

○ mcquitty

○ median

○ centroid

Re-order Rows?

● Yes

○ No

Scaling?

● row

○ column

○ none

Cut the dendrogram into k groups, or h(eight)

● k

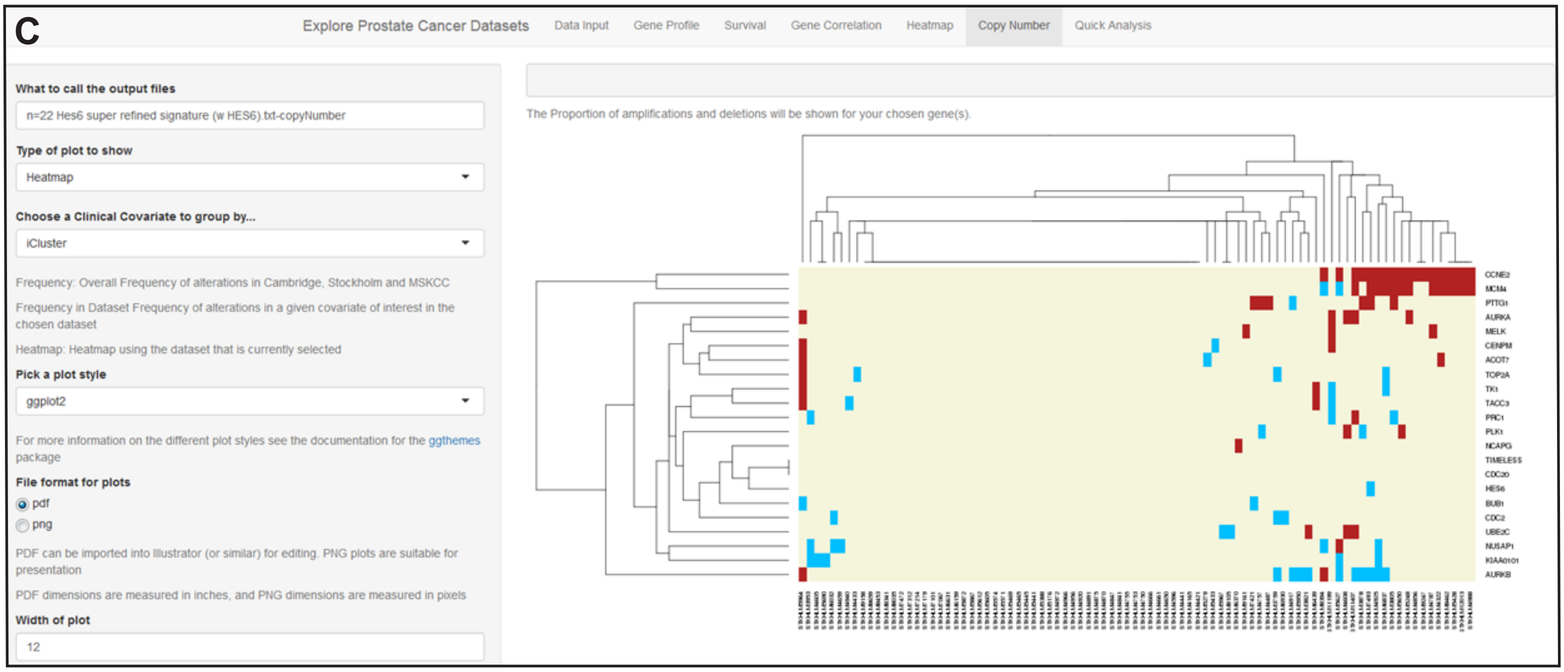
○ h

Select k groups from the dendrogram

2

Construing a heatmap from the gene list you uploaded in the Analysis Parameters tab. If you haven't uploaded a gene list, an example gene list of three genes will be used

Sample Clustering



What to call the output files

n=22 Hes6 super refined signature (w HES6).txt-copyNumber

Type of plot to show

Heatmap

Choose a Clinical Covariate to group by...

iCluster

Frequency: Overall Frequency of alterations in Cambridge, Stockholm and MSKCC

Frequency in Dataset Frequency of alterations in a given covariate of interest in the chosen dataset

Heatmap: Heatmap using the dataset that is currently selected

Pick a plot style

ggplot2

For more information on the different plot styles see the documentation for the ggthemes package

File format for plots

● pdf

○ png

PDF can be imported into Illustrator (or similar) for editing. PNG plots are suitable for presentation

PDF dimensions are measured in inches, and PNG dimensions are measured in pixels

Width of plot

12

The Proportion of amplifications and deletions will be shown for your chosen gene(s).