Barts Cancer Institute
Queen Mary, University of London

Transcriptomics integrative analysis of microarray and next-generation sequencing data

Jacek Marzec, Jun Wang, Jude Fitzgibbon, Yong-Jie Lu and Claude Chelala

BACKGROUND

The microarray and recently introduced next generation sequencing (NGS) technologies became powerful tools in cancer research. The advance of these high-throughput technologies and the facility in data storage is leading to an ever-increasing amount of large-scale functional genomic datasets available in the public domain. These data are only useful if they could be effectively integrated and further explored. While this makes cross-platform analysis an attractive research tool that could provide a comprehensive characterisation of cancer complexity and drive gene prioritisation for experimental validation, open source tools for integrating and analysing data obtained from different platforms and technologies are still lacking.

We present a robust bioinformatics framework for integrative analysis of high-throughput transcriptomics data derived from various microarray and NGS platforms applied to prostate cancer data available within the public domain. The extensive volume of data used in this study provides the most comprehensive insight into a prostate cancer transcriptional space to date.

METHODS AND RESULTS

We collected four RNA sequencing (RNA-seq) and 19 array-based gene expression datasets from 11 different platforms comprising 2,343 transcriptomics profiles of tumour-free prostate, high-grade prostatic intraepithelial neoplasia (HGPIN), primary prostate cancer, prostate cancer with metastasis and multiple-site metastatic prostate cancer tissues (Table 1). Curated datasets were subjected to proposed analytical framework for cross-platform integrative analysis (Figure 1).

After data quality control and pre-processing we performed differential expression analysis between respective biological groups (Figure 2) followed by integrative and survival analyses. We found top differentially expressed genes contributing to the transition from normal or HGPIN to localised prostate cancer and eventually to metastatic disease. Among them are a number of genes previously reported to be implicated in prostate cancer development and progression as well as new potential prognostic biomarkers not identified by individual studies (Figure 3 and 4).

Technology	Platform	Tumour-free prostate	HGPIN	Primary tumour	Metastatic tumour (prostate)	Metastatic tumour (multiple-site)
RNA-seq	Illumina HiSeq 2000 RNA	65	-	358	-	-
	Illumina Genome Analyzer II RNA	10	-	20	-	-
Microarray	Affymetrix Human Exon 1.0 ST	82	-	234	28	-
	Affymetrix Human Genome U133 Plus 2.0	158	25	145	-	6
	Affymetrix Human Genome U133A 2.0	20	-	69	-	29
	Affymetrix Human Genome U133A	95	-	90	75	-
	Affymetrix Human Genome U133B	13	-	44	-	-
	Affymetrix Human Genome U95Av2	175	-	153	-	25
	Affymetrix Human Genome U95B	75	-	63	-	25
	Affymetrix Human Genome U95C	75	-	63	-	25
	Illumina HumanHT-12 v3.0	39	-	59	-	-
	Total number of samples	807	25	1,298	103	110

Table 1. Overview of platforms and gene expression profiles used for integrative analysis. HGPIN – high-grade prostatic intraepithelial neoplasia

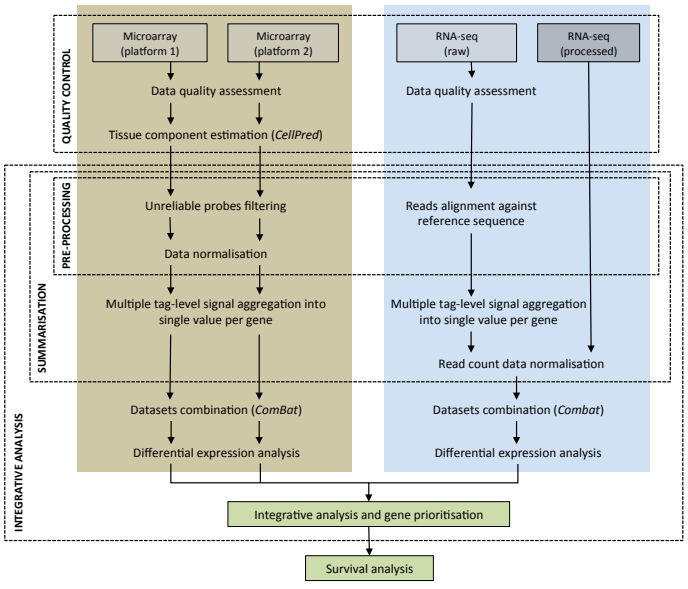

Figure 1. Overview of the gene expression data integration and analysis pipeline.

CONCLUSIONS

Systematic cross-platform integrative analysis performed on large patient cohort allows development of more accurate gene signatures, provides more comprehensive biological insight and allows identification of novel diagnostic and prognostic biomarkers not apparent from individual experiments.

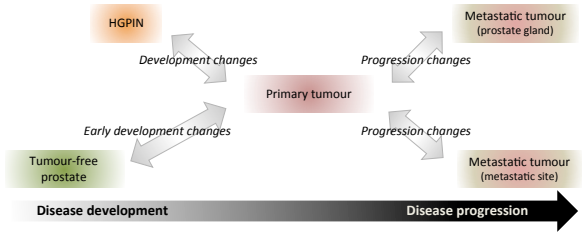

Figure 2. Presented biological comparisons were performed to detect aberrantly expressed genes associated with prostate cancer development and progression to aggressive disease.

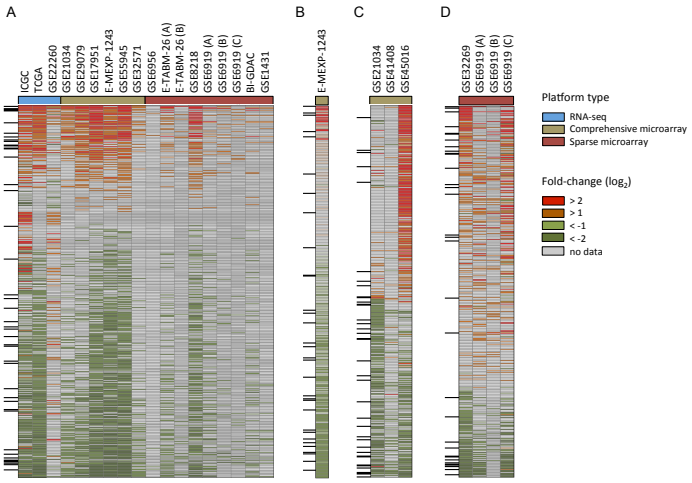

Figure 3. Heat maps of 500 genes with highest evidence of aberrant expression between (A) primary prostate cancer and tumour-free prostate, (B) primary prostate cancer and HGPIN, (C) prostate cancer with metastasis and primary prostate cancer, (D) multiple-site metastatic and primary prostate cancer groups. The heat map colours correspond to $\log_2$ fold-change values between investigated groups computed for individual datasets. Bright-grey cells indicate $\log_2$ fold-change values between -1 and 1. Dark-grey cells correspond to genes not present on a given platform or removed by non-specific gene filtering. Datasets are shown at the top and are ordered accordingly to platform comprehensiveness, with the most comprehensive platforms located on the left side of each heat map. Genes are represented by rows and are arranged based on average $\log_2$ fold-change values across all datasets. Genes previously reported to be implicated in prostate cancer are depicted on the left side as black blocks.

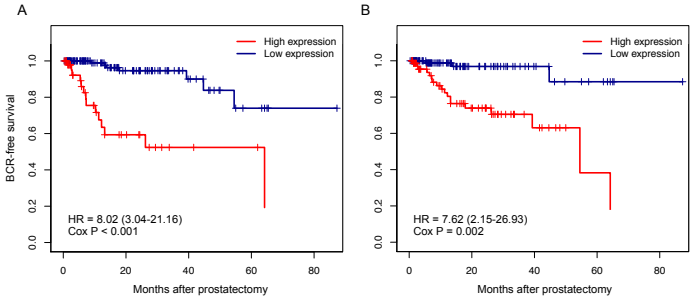

Figure 4. Kaplan-Meier survival plots for (A) *MKI67* and (B) *TWIST1* genes, which aberrant expression significantly correlated with biochemical recurrence-free (BCR-free) survival after adjusting for effects of other clinical parameters. *MKI67* is broadly used diagnostic marker in various cancers, also related to distant metastasis and mortality of prostate cancer. *TWIST1* encodes transcription factor and was previously reported to be involved in prostate cancer progression and metastasis.

Contact: j.marzec@qmul.ac.uk
Web: www.bci.qmul.ac.uk

Queen Mary University of London

EPSRC Engineering and Physical Sciences Research Council