RESEARCH ARTICLE

A New Combinatorial Optimization Approach for Integrated Feature Selection Using Different Datasets: A Prostate Cancer Transcriptomic Study

Nisha Puthiyedth[1,2], **Carlos Riveros**[1,2], **Regina Berretta**[1,2], **Pablo Moscato**[1,2]*

1 Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine, Hunter Medical Research Institute, New Lambton Heights, NSW, Australia, **2** School of Electrical Engineering and Computer Science, The University of Newcastle, Callaghan NSW, Australia

* pablo.moscato@newcastle.edu.au

Abstract

Background

The joint study of multiple datasets has become a common technique for increasing statistical power in detecting biomarkers obtained from smaller studies. The approach generally followed is based on the fact that as the total number of samples increases, we expect to have greater power to detect associations of interest. This methodology has been applied to genome-wide association and transcriptomic studies due to the availability of datasets in the public domain. While this approach is well established in biostatistics, the introduction of new combinatorial optimization models to address this issue has not been explored in depth. In this study, we introduce a new model for the integration of multiple datasets and we show its application in transcriptomics.

Methods

We propose a new combinatorial optimization problem that addresses the core issue of biomarker detection in integrated datasets. Optimal solutions for this model deliver a feature selection from a panel of prospective biomarkers. The model we propose is a generalised version of the *(α,β)-k*-Feature Set problem. We illustrate the performance of this new methodology via a challenging meta-analysis task involving six prostate cancer microarray datasets. The results are then compared to the popular RankProd meta-analysis tool and to what can be obtained by analysing the individual datasets by statistical and combinatorial methods alone.

Results

Application of the integrated method resulted in a more informative signature than the rank-based meta-analysis or individual dataset results, and overcomes problems arising from real world datasets. The set of genes identified is highly significant in the context of prostate

OPEN ACCESS

Citation: Puthiyedth N, Riveros C, Berretta R, Moscato P (2015) A New Combinatorial Optimization Approach for Integrated Feature Selection Using Different Datasets: A Prostate Cancer Transcriptomic Study. PLoS ONE 10(6): e0127702. doi:10.1371/journal.pone.0127702

Academic Editor: Holger Fröhlich, University of Bonn, Bonn-Aachen International Center for IT, GERMANY

Received: November 3, 2014

Accepted: April 17, 2015

Published: June 24, 2015

Data Availability Statement: Singh dataset is available at the Broad Institute Cancer Program Legacy Publication Resources website: http://www.broadinstitute.org/mpr/publications/projects/Prostate_Cancer/prostate_normal_N01-N31.CEL.tar.gz; http://www.broadinstitute.org/mpr/publications/projects/Prostate_Cancer/prostate_normal_N32-N62.CEL.tar.gz; http://www.broadinstitute.org/mpr/publications/projects/Prostate_Cancer/prostate_tumor_T01-T30.CEL.tar.gz; http://www.broadinstitute.org/mpr/publications/projects/Prostate_Cancer/prostate_ tumor_T31-T62.CEL.tar.gz. Welsh dataset is available at the Genomics Institute of the Novartis Research Foundation: http://www.stat.cmu.edu/~jiashun/Research/software/HCClassification/Prostate/GNF_prostate_data_CR61_5974.xls. Uma dataset is available at ArrayExpress: (http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-6919/) under accession number E-GEOD-6919. L-2695, L-3044 and L-3289 datasets are available in Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE3933.

Funding: PM is supported by Australian Research Council (ARC, http://www.arc.gov.au/) Future Fellowship FT120100060. This project is partially funded by ARC Discovery Project DP120102576, Australia.

Competing Interests: The authors have declared that no competing interests exist.

cancer. The method used does not rely on homogenisation or transformation of values to a common scale, and at the same time is able to capture markers associated with subgroups of the disease.

Introduction

The extraction of information arising from the integration of multiple datasets and its translation into domain knowledge is a significant problem in several fields. Today, more and more biology and health related studies around the world are engaging in the useful policy of leaving their raw results available for the common good via public domain databases. This open sharing has benefitted the reproducibility of other researchers' findings. The existing online datasets are also becoming very useful for the development of new mathematical and computational approaches for pattern recognition, machine learning and artificial intelligence methods. This healthy practice of sharing data is now being increasingly adopted by governments and scientific journals. The private and public sector is also engaged in "data-mining competitions" in which the datasets are made widely available and crowd-sourced for data analysis. In this new, digital and interconnected global research open data enterprise, this is definitely a good direction for science, research and development and we are confident to affirm that this trend is here to stay.

The term 'meta-analysis' generally refers to an integrated study which aims at developing a consensus of findings from individual studies. Sometimes authors use this term rather loosely meaning just a 'review' of a set of existing studies that are independently obtained but related to a set of common questions of interest [1]. When some conditions are met, an integrated study can help to improve the power of the analysis by increasing the total number of samples under consideration [2]. Meta-analyses are also an important tool when some of the existing studies have conflicting conclusions [3] and the overall aim is to resolve them, if possible. Increasing the detection power of smaller studies by integrating them in a larger study has also become a way to overcome research funding limitations. This is particularly the case in transcriptomics, and there is an undeniable need for new mathematical models and algorithms aimed at extracting information by jointly studying different datasets which often contain information extracted with different and ever-changing technological platforms.

The existence of large number of publicly available transcriptomic studies gives a strong motivation for the development of new mathematical methods that help to extract *panels of biomarkers* by employing several microarray datasets. In spite of the growing number of studies, an overall consensus has yet to be reached about how to do this [4, 5]. Researchers sometimes only highlight the obstacles ahead, for instance, by pointing at the essential differences in microarray platforms, experimental designs, collection procedures for samples, heterogeneities of laboratory protocols and the analysis methods used for the study [6]. Most of the studies are unable to provide a definite answer to the question of interest since too few samples are entered into the study [7]. However, all these confounding issues need to be considered and highlighting them does not diminish the need to develop integrative techniques for joint panel of biomarkers elicitation.

Many studies have shown that it is difficult to obtain a reliable result from a single dataset [8–11]. Even though some researchers may eventually procure the financial resources to conduct studies with large number of samples, leading to greater power to detect individual markers, an integrated study can provide a clearer picture as the final result would look for consensus in a number of individual studies. This shows the necessity for developing combinatorial optimization-based approaches to determine a significant list of genes from multiple platforms when we are looking at a panel that acts together for a discrimination task across several studies.

Multi-platform data integration remains challenging as the datasets from different experiments are not directly comparable due to the factors associated with the generation of the dataset [12]. Some of the challenges are simply technical in nature, for instance the genomic data may come in a wide variety of data formats, thus making direct integration difficult. The datasets can be converted to a common data format before combining them, but this is not always feasible [13]. Several methods have been proposed in the last few years for the meta-analysis of gene expression data to find the set of significant genes among the selected datasets. The existing meta-analysis methods either perform statistics for each dataset or integrate all the selected datasets into a single large dataset to estimate the differential gene expression. A rank based method proposed by Breitling *et al.* [14] and later developed by Hong *et al.* into the RankProd Bioconductor package [15], uses the fold changes between all interclass pair of samples to compute dataset ranks for each gene, then combines ranks with the geometric mean of ranks across sample pairs. MetaArray is another meta-analysis method proposed by Choi *et al.* [16] in which the data is transformed into probability of expression [17] followed by the filtering of genes based on the integrative correlation analysis. Mergemaid [18] is another package for meta-analysis that helps to integrate heterogeneous platform datasets on the basis of user-provided IDs of genes. The standardized regression coefficients and z-scores are used as a measure for the gene selection process form the integrated dataset. Although these methods are capable to select signatures from the integrated dataset of heterogeneous platforms, they are incapable to deal with genes not represented in all the datasets. A recently proposed method called NetSel [19] is a heuristic rank aggregation method for feature selection that can be applied on heterogeneous set of lists. However, RankProd is by far the most popular of those methods, and we have chosen it as a comparison benchmark.

The goal of this article is to present a new method for the integration of microarray gene expression datasets which may have been obtained using different platforms. We do this without needing to transform the values to a common uniform format and range of values. We also propose a new combinatorial optimisation approach to select the best set of common features that can discriminate the given classes. The method is a generalised version of the proven and very successful *(α,β)-k*-Feature Set methodology previously pioneered by our group [20, 21] and we show here how it can be applied to the combined dataset. We benchmark our new method by analysing the integration of six prostate cancer datasets produced using different platforms and highlight its main findings. We deliberately turn our attention to relatively small and also relatively old datasets, somewhat disregarded as potentially “uninteresting” due to the advances of current biotechnologies. We compare the integrated results against the collection of results of individually applying traditional statistical analysis and the *(α,β)-k*-Feature Set methodology to each dataset. We aim to illustrate the potential of secondary analyses of these datasets using the proposed technique.

The structure of the article is as follows; the materials and methods employed in this paper are explained in detail in Section 2; in Section 3 we present our results by applying the proposed integration and feature selection method on prostate cancer datasets. In Section 4 we present some discussion on the basis of the result. Section 5 gives a conclusion of this study and future directions.

Table 1. Summary of datasets used in this study.

Name	Plat	Series	NS	Norm	PT	Met	Probes	EF
Singh [22]	Affymetrix [HG-U95Av2]	N/A	102	50	52	0	12558	1519
Welsh [23]	Affymetrix [HG-U95Av2]	N/A	55	9	25	21	12560	2429
Uma [24]	Affymetrix [HG-U95B]	E-GEOD-6919	80	17	63	0	37691	3484
L-2695 [25]	SHBB	GSE3933	26	9	13	4	44161	4288
L-3044 [25]	SHCQ	GSE3933	41	16	23	2	43009	4082
L-3289 [25]	SHBW	GSE3933	45	16	26	3	43009	4953

Name is the name assigned to the study throughout this paper. **Plat** is the platform details of each dataset. **Series** is the Gene Expression Omnibus Series identifier for the dataset. **NS** is the original number of samples in the study, of which **Norm** are the number of healthy tissue samples, **PT** are the number of primary tumour samples, **Met** is the number of metastasis samples present in each dataset, **Probes** is the number of probes present in each dataset, **EF** is the number of probes present after entropy filtering.

doi:10.1371/journal.pone.0127702.t001

Materials and Methods

2.1 Datasets

The six publicly available prostate cancer gene expression datasets used in this study were collected from Gene Expression Omnibus (GEO) or from the original source. The details of all the datasets in this work are summarised in Table 1.

The selected datasets have been generated using two different platforms. The gene expression levels of three of them were measured using cDNA two-channel arrays and the other three using Affymetrix arrays. The datasets are named according to the name of the first author of the published article. As shown in, the last three datasets are collected form the same article, so the datasets have been named with the first author's initial and the GEO platform number (eg. L-2695). Details of the datasets are as follow.

In [22], Singh et al. introduced an outcome prediction model to distinguish between tumour and normal samples. The dataset used in this study contains 102 tissue samples collected after radical prostatectomy. The sample consists of 50 normal samples and 52 primary prostate cancer samples. This dataset was generated using Affymetrix HG-U95A v2 (GPL8300) arrays.

The second dataset has been contributed by Welsh et al. [23] in 2001. The study investigates a therapeutic approach to differentiate the tumour and normal samples. The dataset contains 55 samples that are hybridised to HG-U95A v2 (GPL8300) arrays. The samples are of 25 primary tumour and 9 normal tissues and the rest of the samples were taken from different donors with different types of cancers.

The third dataset has been published by Uma et al. in 2007 [24]. This study introduces an experimental design to address the differences in cellular content between primary and metastatic tumours. The dataset contains 63 tumour tissue samples and 17 normal tissue samples and has been produced using Affymetrix HGU95Av2 arrays.

Lapointe et al. [25] introduced a hierarchical clustering technique to distinguish tumour from normal samples and to identify the subclasses of prostate cancer in 2004. This study was performed using three different datasets produced using cDNA two-channel arrays; the first Lapointe dataset (L-2695) contains 26 samples (13 primary tumour tissue, 9 normal tissue and 4 metastasis tissue samples). The second Lapointe dataset (L-3044), with a total sample count of 41, has 23 primary tumour samples, 16 normal samples and 2 metastasis samples. The third dataset (L-3289) contains a total of 45 samples, of which 26 are primary tumour, 16 normal and 3 metastasis samples.

We have restricted our study only to those samples which originate in either primary tumours or normal tissue. The total numbers of samples are then 319, of which 202 are primary tumours and the rest are from normal tissue.

2.2 Integration method

The direct integration of microarray gene expression data from multiple platforms is, in principle, greatly facilitated when there exists commonality between the platforms used. However different gene expression platforms will target genes or transcripts differently by using different sets of probes. There may be many probes mapping the same gene due to duplicate spotted probes in microarray chips. On the other hand, there may be a single probe that maps to several genes (or loci) if the specificity of the probe sequence is not good enough. These probes must be discarded from the preliminary analysis as it is difficult to analyse these multiple genes. In addition, the interpretation of the results via Gene Ontology or pathway-informed databases could be compromised by the multiple mapping problems. In addition to these difficulties, we may also face the problem that one probe targeting different regions of the same gene could be indirectly monitoring possible different abundances of protein isoforms. This many-to-many nature of the mapping problem makes it difficult to take a simplistic approach to the essentially different maps that platforms produce by their probe sets.

In this contribution, we map at the gene level. In order to map the probes across the platforms in Table 1 to genes, we have used a simple alignment policy, explained below; with no distinction of isoforms and also ignored the mentioned problems. The probes were mapped using the hg19-GRCh37 version of the Genome Browser's table produced by the Genome Reference Consortium to avoid the misnaming and misalignment of genes. In order to obtain a relatively large number of probes that could be used in the final integrated dataset, we collected those that satisfy any of the given three conditions:

- Where the probes are targeting the same sequence
- Where the targeting sequences are overlapping
- Where the targeting sequences are at a distance of at most 1000 base pairs

The probes from each dataset have been mapped to genes and the associated transcription start and end position of the targeting genes compared according to the conditions mentioned above. Whenever there is a common targeting gene for different probes from multiple datasets, we consider the different combinations of those probes in the combined dataset. Similarly, if the features (the transcription start and end sequences) have an overlap between them, or are at a distance of at most 1000bp, the combination of those probes is also selected to be part of the combined dataset. The selected list of combination of probes is given in the Supplementary Materials (S1 Table). Each unique combination of probes from different datasets becomes a feature in the combined dataset.

2.3 Feature selection method

Initially, we used Fayyad and Irani's entropy-based heuristic on each individual dataset to remove uninformative features. This univariate selection mechanism is a pre-processing step related to the Minimum Description Length Principle (MDL) [26]. The purpose of using this step in this method is twofold: it removes features that are not significantly different in healthy and disease samples (thus it helps by reducing the dimensionality of the problem), and second it helps discretise the values (which in turn facilitate the combinatorial approach).

In this contribution we propose and analyse a new combinatorial approach to select a set of *k* significant features that can explain the multi-platform integrated datasets. We call this problem the Coloured *(α,β)-k*-Feature Set problem. The approach is a generalised version of the *(α, β)-k*-Feature Set problem methodology [27, 28] which is a supervised feature selection method to select a significant set of features that can collectively separate the sample groups. The method has been successfully used in several studies by Moscato et al. for finding biomarkers for different diseases [20, 21, 28–34].

The *(α,β)-k*-Feature Set problem provides a significant set of genes that collectively maximise the inter-class discrimination and the intra-class coherency [33]. The method seeks to differentiate all sample pairs which belong to different classes by selecting a minimum set of genes that do not necessarily present a uniform expression level across samples in each class but collectively provide the maximum amount of evidence. In contrast, rank methods that score and order genes by their differential expression across the classes bring gene sets that may not work together as a signature, particularly in complex diseases whose molecular characterisation may present subgroups.

The mentioned feature selection method works well with a single uniform dataset, but not for an integrated dataset. The Coloured *(α,β)-k*-Feature Set problem handles the integrated dataset in a consistent manner and selects features that differentiate sample pairs across the datasets. The application of an *(α,β)-k*-Feature Set problem based method for meta-analysis thus helps provide the best set of features from the combined dataset, allowing researchers to reveal the genetic pathways that take part in the development of the disease.

Here we more formally present the decision versions of the generalization of the *k*-Feature Set problem called the *(α,β)-k*-Feature Set problem, the Coloured *(α,β)-k*-Feature Set problem and the Generalised *(α,β)-k*-Feature Set problem. In what follows, let $\mathbb{B}$ represent the set of binary values, i.e. $\mathbb{B} = \{0, 1\}$; let *n* be the number of features and *m* the number of samples, *p* be the number of sample groups (i.e., different platforms/cohorts/datasets) and the tuple *y* be the class labels of the samples.

2.3.1 (α,β)-k-Feature Set.

Instance:. A set $X = \{x_i | x_i \in \mathbb{B}^n \wedge 1 \leq i \leq m\}$, a tuple $y \in B^m$, integers $\alpha > 0, \beta \geq 0, k > 0$

Parameters:. α, β and k

Question:. Is there a set $I \subseteq \{1,..., n\}$ with $|I| \leq k$ such that for all $i, j \in \{1,..., m\}$

- If $y_i \neq y_j$ there exists $I^{\alpha}_{i,j} \subseteq I$ with $|I^{\alpha}_{i,j}| \geq \alpha$ such that $x_{i,s} \neq x_{j,s}$ for all $s \in I^{\alpha}_{(i,j)}$,
- If $y_i = y_j$ there exists $I^{\beta}_{i,j} \subseteq I$ with $|I^{\beta}_{(i,j)}| \geq \beta$ such that $x_{i,s} = x_{j,s}$ for all $s \in I^{\beta}_{(i,j)}$?

Detailed explanation of safe reduction rules that help to reduce the dimensionality of the *(α, β)-k* Feature Set problem are given in [20, 32].

2.3.2 Coloured *(α,β)-k*-Feature Set.

Instance:. A set $X = \{x_i | x_i \in \mathbb{B}^n \wedge 1 \leq i \leq m\}$, a colouring function c: $\{1,..., m\} \rightarrow \{1,..., p\}$, a tuple $y \in \mathbb{B}^m$, integers $\alpha > 0, \beta \geq 0, k > 0$

Parameters:. α, β and k

Question:. Is there a set $I \subseteq \{1,..., n\}$ with $|I| \leq k$ such that for all $i, j \in \{1,..., m\}$ where $c(i) = c(j)$

- If $y_i \neq y_j$ there exists $I^{\alpha}_{i,j} \subseteq I$ with $|I^{\alpha}_{i,j}| \geq \alpha$ such that $x_{i,s} \neq x_{j,s}$ for all $s \in I^{\alpha}_{(i,j)}$,
- If $y_i = y_j$ there exists $I^{\beta}_{i,j} \subseteq I$ with $|I^{\beta}_{(i,j)}| \geq \beta$ such that $x_{i,s} = x_{j,s}$ for all $s \in I^{\beta}_{(i,j)}$?

In words, the Coloured *(α,β)-k*-Feature Set problem instance is constructed from a collection of individual *(α,β)-k*-Feature Set instances with common features, where the comparison of feature values is limited to sample pairs formed from each individual instance. The "coloured" name stems from assuming samples in each individual instance are coloured with the same unique colour, then only same coloured samples can be combined in pairs.

It is evident that the same set of data reduction rules presented in [21] for the *(α,β)-k*-Feature Set problem applies to an instance of the Coloured *(α,β)-k*-Feature Set problem, as the latter is formally equivalent to a larger instance of an *(α,β)-k*-Feature Set problem by an appropriate relabelling of samples.

2.3.3 Generalised *(α,β)-k*-Feature Set.

In the most general form appropriate for meta-analysis of datasets with common features, the *(α,β)-k*-Feature Set problem can be stated as follows:

Instance:. A set $X = \{x_i | x_i \in \mathbb{B}^n \wedge 1 \leq i \leq m\}$, a function $g : \{1, \ldots, m\} \times \{1, \ldots, m\} \rightarrow \mathbb{B}$, a tuple $y \in \mathbb{B}^m$, integers $\alpha > 0, \beta \geq 0, k > 0$

Parameters:. α, β and k

Question:. Is there a set $I \subseteq \{1,\ldots, n\}$ with $|I| \leq k$ such that for all $i, j \in \{1,\ldots, m\}$ where $g(i, j) = 1$

- If $y_i \neq y_j$ there exists $I^{\alpha}_{i,j} \subseteq I$ with $|I^{\alpha}_{i,j}| \geq \alpha$ such that $\mathrm{x}_{\mathrm{i,s}} \neq \mathrm{x}_{\mathrm{j,s}}$ for all $s \in I^{\alpha}_{(i,j)}$,
- If $y_i = y_j$ there exists $I^{\beta}_{i,j} \subseteq I$ with $|I^{\beta}_{(i,j)}| \geq \beta$ such that $\mathrm{x}_{\mathrm{i,s}} = \mathrm{x}_{\mathrm{j,s}}$ for all $s \in I^{\beta}_{(i,j)}$?

The Generalised *(α,β)-k*-Feature Set problem has been devised to deal with the more general situation in which some samples in one sample group may be compared to samples in another sample group, for example. The binary function $g(i, j)$ indicates when feature values for a given arbitrary sample pair (i, j) can be compared.

In all previous formulations, the samples have been presented as an array of n+1 binary values, although this is not strictly necessary. The class label can be a categorical variable taking values over a (typically small) set of categories or classes. The features can have values of any kind, as long as there exists a meaningful comparison able to decide if any two values are considered equal or different.

2.3.4 Coloured *(α,β)-k*-Feature Set as an Integer Programming Problem.

Next, we present the Coloured *(α,β)-k*-Feature Set problem as an Integer Programming optimisation problem. Let p, n, m and y be as given before. As the sample groups are disjoint, there are no common samples between any two of them. For any sample j and any feature $s \in \{1,\ldots, n\}$, let $c_j \in \{1,\ldots, p\}$ be the sample group it belongs to, and x_{js} the value of the feature for the sample. For any sample pair (i, j) let

$$a_{ijs} = \begin{cases} 1 \; if \; y_i \neq y_j \; and \; c_i = c_j \; and \; x_{is} \neq x_{js} \\ 0 \; otherwise \end{cases}$$

and

$$b_{ijs} = \begin{cases} 1 \; if \; y_i = y_j \; and \; c_i = c_j \; and \; x_{is} = x_{js} \\ 0 \; otherwise \end{cases}$$

The objective function and constraints for the Coloured *(α,β)-k*-Feature Set problem integer programming optimisation models are given below, where the binary variable f_s is 1 if the feature *s* is selected to the feature set, and 0 otherwise. The problem seeks the minimum of:

$$k = min\sum_{s=1}^{n} f_s \qquad (1)$$

subject to the conditions:

$$\sum_{s=1}^{n} a_{ijs} f_s \geq \alpha \quad \forall (i,j) \qquad (2)$$

$$\sum_{s=1}^{n} b_{ijs} f_s \geq \beta \quad \forall (i,j), \qquad (3)$$

where:

$$f_s \in \{0,1\}$$

A Coloured *(α,β)-k*-Feature Set problem instance can have more than one optimal solution with k features in each. This multiplicity is resolved by a subsequent optimisation problem which searches for the solution of size k with maximum cover. We then define the optimal solution of the Coloured *(α,β)-k*-Feature Set problem as the one that maximises:

$$V = max\sum_{s=1}^{n} e_s f_s \qquad (4)$$

subject to the conditions:

$$\sum_{s=1}^{n} f_s = k \qquad (5)$$

$$\sum_{s=1}^{n} a_{ijs} f_s \geq \alpha \quad \forall (i,j) \qquad (6)$$

$$\sum_{s=1}^{n} b_{ijs} f_s \geq \beta \quad \forall (i,j), \qquad (7)$$

where:

$$f_s \in \{0,1\}$$

In Eq 4, the cover e_s is the number of pairs of samples that feature *s* covers, and can be specified as:

$$e_s = \sum_{i,j \in \{1,\ldots,m\}} (a_{ijs} + b_{ijs})$$

The solution of the optimisation problem (1–3) requires the specification of the parameters α and β. One way of requiring a robust solution of the problem is to specify α as large as possible. This value is determined by the instance of the problem, and is equal to the minimum number of features that differentiate any sample pair of different class labels. Once the value of *k* is obtained with $\beta = 0$, we can then repeatedly solve the problem (4–7) for increasingly large values of β in (7), until the problem becomes unfeasible. The last feasible solution is the signature sought.

A final note about the computational complexity of this family of problems. The *(α,β)-k*-Feature Set problem is at least as complex as the classical *k*-Feature Set problem, which is NP-complete [35, 36]. The *(α,β)-k*-Feature Set problem is not only NP-complete, but W[2]-complete [37, 38].

2.4 t-test

In order to benchmark against traditional statistical methods, we perform a t-test analysis of the individual datasets. The t-test is a statistical significance test method used here to select genes that exhibit differential gene expression between two different conditions [39], in our case normal vs. primary tumour, above a certain *p*-value level of confidence. The procedure of *t*-test is described below:

Let S_1 and S_2 be the mean values of a particular gene in the two different class labels 1 and 2, of sizes m_1 and m_2. The *t*-statistic for this particular gene is computed as:

$$t = \frac{S_1 - S_2}{X\sqrt{\frac{1}{m_1} + \frac{1}{m_2}}}$$

where X is the pooled sample variance

$$X = \sqrt{\frac{m_1 x_1^2 + m_1 x_2^2}{m_1 + m_2 - 2}}$$

Here x_1^2 and x_2^2 are the variance of replicated observations in each condition and $n_1 + n_2 - 2$ is the number of degrees of freedom. In our study we used the 'genefilter' Bioconductor package [40] with a chosen *p*-value of 10^{-4} to perform our *t*-test.

2.5 RankProd

We compare our results to those obtained by another popular meta-analysis method. RankProd is a non-parametric meta-analysis tool introduced by Hong et al. [15] to detect differentially expressed genes. It arguably is the most widely used gene expression meta-analysis method, and is provided as a Bioconductor package that modifies and extends the rank product method proposed by Breitling et al. [14]. Fold Change (FC) is used as scoring criteria to rank and compare genes within each dataset. An overall ranked gene list is produced by aggregating the individual ranks across datasets.

A pair-wise fold change (*p*FC) is computed for each gene *g* within a given dataset *k* as,

$$T_1^g/C_1^g, T_1^g/C_2^g, \ldots, T_2^g/C_1^g, \ldots, T_{n_{Tk}}^g/C_{n_{Ck}}^g$$

in which T_j^g and C_l^g are the expression values of gene *g* for sample *j* (belonging to experimental condition *T*–e.g. "tumour") and *l* (belonging to experimental condition *C*–e.g. "control"), and n_{T_k} and n_{C_k} are the number of replicates which produce a total of $K_k = n_{T_k} \times n_{C_k}$ *p*FC values per gene. Then the corresponding *p*FC ratios are ranked and are denoted as r_{gi}, where $g = 1, \ldots,$ G represents the number of genes and $i = 1, \ldots, K_k$ represents the pairwise comparison between samples. The rank product of each gene *g* is defined as the geometric mean,

$$RP_g = \left(\prod_i^K r_{gi}\right)^{1/K}$$

Expression values for each gene within each datasets is independently permuted *L* times and produce $RP_g^{(l)}$ where $l = 1, \ldots, L$ by repeating the above steps. A reference distribution is obtained from all $RP_g^{(l)}$ and the adjusted p-value and the false discovery rate for each gene calculated.

In this study, the datasets are combined in terms of common genes across the platforms. We have applied RankProd on the combined dataset to select genes associated to the condition being investigated.

2.6 Robustness

To evaluate the robustness of our method with respect to perturbations in the data we have performed a series of experiments. The presence of noise in the gene expression data is difficult to estimate, as it depends on platform-specific factors as well as experimental conditions. However, the final manifestation of perturbations in the datasets would be a change in the composition of the set of probes that pass the MDL criterion. We have thus analysed the robustness of the final integration results with respect to varying compositions of the individual datasets, for different perturbation models, inspired by the 'leave one out' approach. Specifically, we have modelled the following setups: a) removal of one, two and five genes from the combined dataset, and b) removal of one gene from one and two individual datasets. In order to estimate the worst case scenario, all genes were restricted to those that appear in our final signature as expressed in all six datasets. In each case, all combined probes corresponding to the chosen gene (s) are removed. An integrated signature is then obtained and compared with our original signature. The procedure is repeated 10 times for the a) case and 5 times for the b) case, with random selection of gene(s) and dataset(s), and average results reported.

Results

As we have mentioned before, to evaluate the applicability and usefulness of the proposed method, we have selected primary tumours and normal samples from six prostate cancer datasets measured with different platforms. Since the method proposed in this work is a generalization of the *(α,β)-k*-Feature Set approach for probe set selection, the most natural comparison is to evaluate their results by contrasting them to those that are obtained by applying *(α,β)-k*-Feature Set individually. This means that we need to solve the feature set problem for each of the datasets and find the individual gene signatures that discriminate the sample classes. We then apply our proposed method, Coloured *(α,β)-k*-Feature Set problem. This experimental scenario has been designed to observe the benefit of an integrated approach against the comparison of gene lists obtained by analysing each of the individual experiments by separate.

For completeness we include the results of the individual datasets *t*-test analysis. In this way, we compare the benefits of our integrated approach against a frequently used feature selection method based on a univariate statistical test.

To evaluate the relative performance of the proposed method, we compare results obtained by the RankProd method. As explained in the introduction, many meta-analysis methods are not applicable in the general conditions of a multi-platform meta-analysis. RankProd is a popular choice that is able to do it, and somewhat similar in the sense that ranks genes based on the comparison of values for pairs of samples of different class labels.

3.1 Individual *(α,β)-k*-Feature Set problem results

The application of the *(α,β)-k*-Feature Set methodology consists of a pre-filtering step and the solution of a combinatorial optimisation problem. The pre-filtering selects features based on the class information content and discard less informative features thus reducing the

Table 2. The results of the numerical solution of the *(α,β)-k*-Feature Set problem on each of the six individual datasets.

Dataset	Feat.No	After EF	*α*	*β*	*k* (signature size)
Singh	12558	1519	215	329	754
Welsh	12560	2429	1188	1068	1768
Uma	37691	3484	881	1079	1857
L-2695	44161	4288	2266	2421	3533
L-3044	43009	4028	966	862	1800
L-3289	43009	4953	1397	1216	2696

Dataset is the short name used in this paper for the dataset. **Feat. No** is the initial number of features (probes) present in the dataset, **After EF** is the number of features after applying entropy filtering, ***α*** and ***β*** are the values for the parameters *α* and *β* for any feasible solution, and ***k* (signature size)** is the number of probes in the resulting solution to the individual *(α,β)-k*-Feature Selection problem for the dataset. For method details refer to Materials and Methods.

doi:10.1371/journal.pone.0127702.t002

dimensionality for the subsequent combinatorial problem. Details of the methods are provided in Section 0. The characteristics of the individual dataset result are given in Table 2.

Each dataset resulted in molecular signatures with a large number of genes (provided in the S2 Table). Surprisingly the number of common genes between them is only seven and they represent a negligible overlap of results between all experiments. This shows the need of an integrative method as it would be infeasible to come up with any form of statistical support that could link these genes to putative pathways that could be deregulated. On the positive side, however, all seven genes in the overlap already have reported association with prostate cancer. The list and literature references are given in Table 3.

3.2 *t*-test Results

In order to have a baseline for comparison with another methodology common in practice, a *t*-test was conducted on each of the six individual datasets to compare the gene expression levels of normal and primary tumour samples. We compare this with our individual *(α,β)-k-* Feature Selection results. The method is explained in detail in Section 2 and the individual dataset results are given in Table 4.

Large numbers of genes are filtered out from each dataset using the *t*-test approach (the resulted list of genes is provided in S3 Table). The only genes common to the results of all six experiments are EPCAM (epithelial cell adhesion molecule), also a known marker [17, 41–48], SOX4, EEF2 and AMACR. This shows even less genes in common than the overlap between

Table 3. List of common genes among all the individual dataset results from Table 2.

Gene Symbol	Gene Name	Reference
EEF2	Eukaryotic Translation Elongation Factor 2	[82, 83]
SPG20	Spastic Paraplegia 20	No associated reference
ERG	Erythroblastosis Virus E26 Oncogene Homolog	[89, 90]
AMACR	Alpha-Methylacyl-CoA Racemase	[59, 91]
SOX4	SRY (Sex determining Region Y)-box 4	[64, 90]
APOC1	Apolipoprotein C-I	[92, 93]
GUCY1A3	Guanylate Cyclase 1, soluble, alpha 3	[94]

Gene Symbol is the official gene symbols. **Gene Name** is the expanded gene name. **Reference** is the reference for each gene which shows the relation with prostate cancer.

doi:10.1371/journal.pone.0127702.t003

Table 4. *t*-test results on individual dataset.

Dataset	Feat.No	Signature size
Singh	1519	616
Welsh	2429	717
Uma	3484	690
L-2695	4288	286
L-3044	4028	654
L-3289	4953	647

Dataset is the short name used in this paper for the dataset. **Feat.No** is the number of features (probes) present in the dataset before applying t-test, and **Signature size** is the number of genes in the resulting solution for each dataset. For method details refer to Section 2.

doi:10.1371/journal.pone.0127702.t004

individual *(α,β)–k*-Feature Selection signatures. The *t*-test result is consistent with the individual *(α,β)-k*-Feature Selection results as SOX4, EEF2 and AMACR are also present in the overlapping genes of individual *(α,β)-k*-Feature Selection results.

3.3 Coloured *(α,β)-k*-Feature Set problem results

To apply the proposed feature selection method we prepared a combined dataset collecting the entropy filtered probes from the individual studies. This ensures that selected probes in each individual study carry some differential expression information with respect to the sample classes, and also provides a well-defined discretization which respects the individual study conditions.

Probes in one platform were matched to probes in another platform based on gene names and genomic positions as explained in Section 0. The combined dataset contains 319 samples and 16157 combined probes. Out of these, 1405 contain values for all six datasets and 10729 for three or more datasets which is annotated to 1454 unique genes. The number of combined probes covering only one dataset was 3425. This uneven cover of datasets is due to some probes being discarded by the entropy filtering as uninformative only in some datasets and not in others. However, the large number of combined probes with values in three or more datasets indicates a good level of coverage after dataset integration.

We then applied Coloured *(α,β)-k*-Feature Set selection methodology in the combined dataset and obtained a resulting list of 3190 combined probes with a maximum of α and β value of 612 and 776 respectively, which corresponded to 1788 unique genes. The resulted number of probes for selected number of datasets and their corresponding number of genes are given in Table 5 (The list of genes can be found in S4 Table).

Table 5. Result of Coloured *(α,β)-k*-Feature Set selection methodology.

No of Datasets	No of Combined Probes	No of Genes
Four or more	2272	327
Five or more	1806	186
Six	792	120

No of Datasets is the considered number of datasets to find the coverage. **No of Combined Probes** is the resulted number of features after applying Coloured *(α,β)-k*-Feature Set selection methodology and **No of Genes** is the number of genes corresponds to the number of combined probes.

doi:10.1371/journal.pone.0127702.t005

An gene ordering algorithm, presented in [49], has been applied on this set of genes to generate a heatmap that brings out the correlation between the resulted genes and is shown in Fig 1, heatmap for the 186 genes that cover five or more datasets and Fig 2 for the 120 genes that cover all six datasets, respectively.

If we consider the genes that appear in the overlap of the *t*-test, *(α,β)-k*-Feature Selection and Coloured *(α,β)-k*-Feature Selection results individually, we get very few genes in the case of *t*-test and the individual *(α,β)-k*-Feature Selection, but Coloured *(α,β)-k*-Feature Selection gives 120 unique genes (The list of genes and the details can be found in S5 Table). That shows a significant difference in the number of common genes from 7 to 120. The number of overlapping genes in different method is given in Table 6.

3.4 RankProd Result

The RankProd ordered the genes by increasing pfp (percentage of false positive likelihood) value, and the top genes with a 0.05 pfp cut-off from both up and down regulated list of genes were used for the comparison. This resulted in a list of 1883 genes from the combined dataset (the list of genes can be found in S6 Table).

The comparison between Coloured *(α,β)-k*-Feature Set methodology result (120 genes) with the RankProd result shows that 80 out of our 120 genes are present in the top listed genes of RankProd result of the combined dataset. This signals a high level of agreement between the two meta-analysis methods. The comparison of Coloured *(α,β)-k*-Feature Set problem result and RankProd result is given in Table 7. All genes in RankProd result also appearing in Coloured *(α,β)-k*-Feature Set result are marked in a supplementary material table (S6 Table). In addition to the common 80 genes, notice there are also genes marked as four or five datasets in Coloured *(α,β)-k* because they have been filtered by the entropy filtering as non-informative for one or two datasets. This increases the agreement to 260 genes out of 327 (almost 80%) appearing in four or more datasets for Coloured *(α,β)-k*.

However, further analysis with RankProd including genes missing in one or more datasets places these genes at the top of the list, making further analysis difficult. Similarly, when genes with sparse missing values are included, these genes artificially escalate in the ranked lists towards the significant side as more missing values are introduced. This shows the inability of this method to deal with two frequent situations found in microarray datasets.

3.5 Robustness

To evaluate the robustness of the proposed method, we performed a sensitivity analysis by iteratively removing one or a set a genes at a time and compared the result with the original result. Summary results are given in Table 8. On average, our results remain the same for more than 97% of the signature list, while the signatures size remain essentially the same (less than 0.5% increase in the worst case). This point to a highly robust result which does not depend on a (small) set of genes, even if they are on the high coverage set.

3. 6 Functional and Pathway Analysis

Functional and pathway analysis has been performed on these 120 genes for further validation of our results. We have used DAVID [50] and STRING [51] for the functional annotation of the association between these genes. Functional annotation of these 120 genes clustered as 8 functionally related groups. Most of the genes in each group are related with prostate cancer and the most known genes in relation with prostate cancer with the clusters of genes can be found in S7 Table.

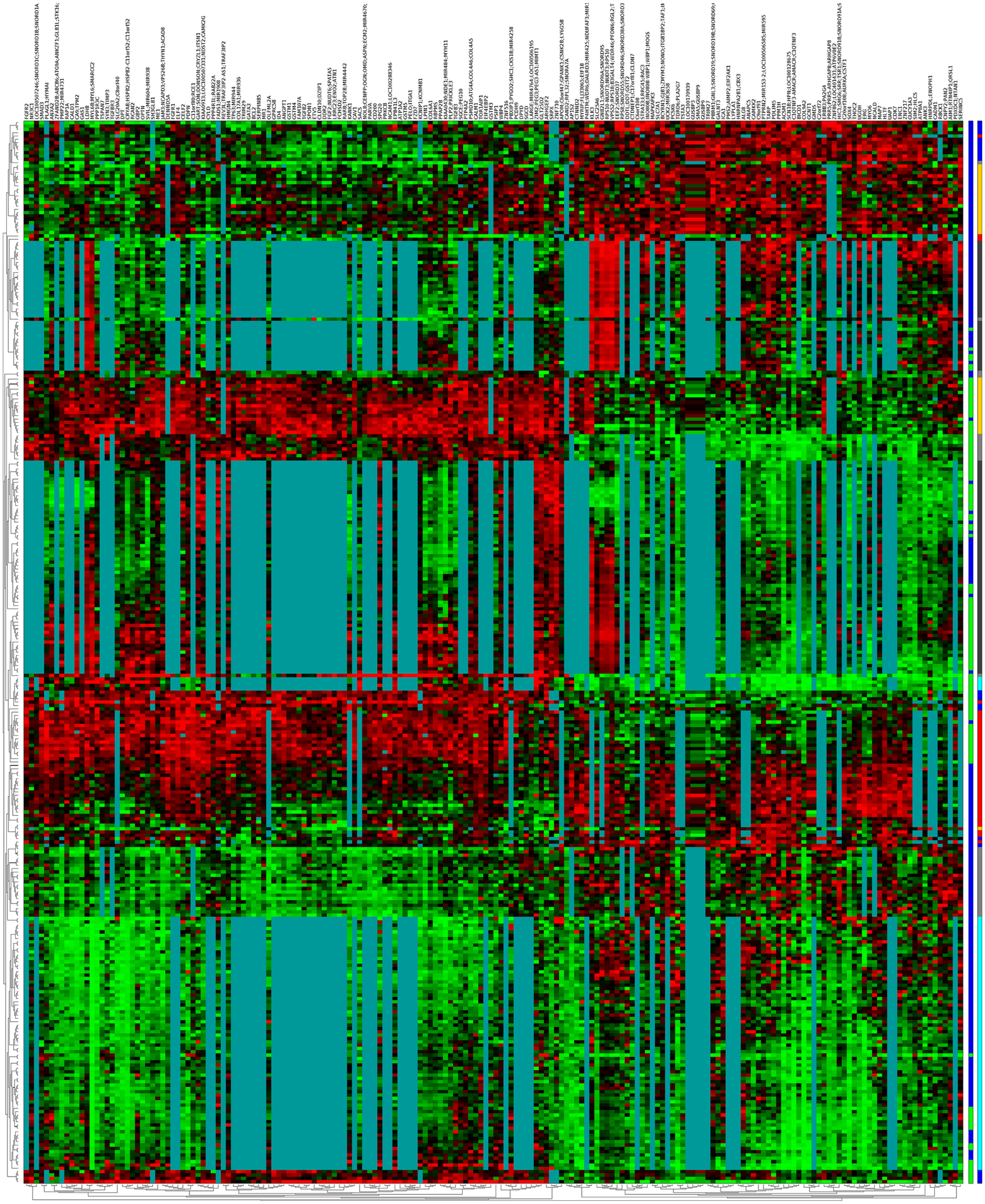

Fig 1. Heatmap for the Coloured *(α,β)-k*-Feature Selection resulted genes that cover five or more datasets. It contains 186 up and down regulated genes (columns). The genes are ordered using a memetic algorithm introduced by Moscato et al. in [49]. The blocks of greenish blue colour represent the absence of gene values in particular datasets. The first colour bar at the right indicates Primary Tumour (blue) and Normal [13] samples. The second colour bar represents each sample group in different colour. L-2695 (blue), L-3044 (red), L-3289 (orange), Welsh (grey), Uma (cyan) and Singh (dark grey).

doi:10.1371/journal.pone.0127702.g001

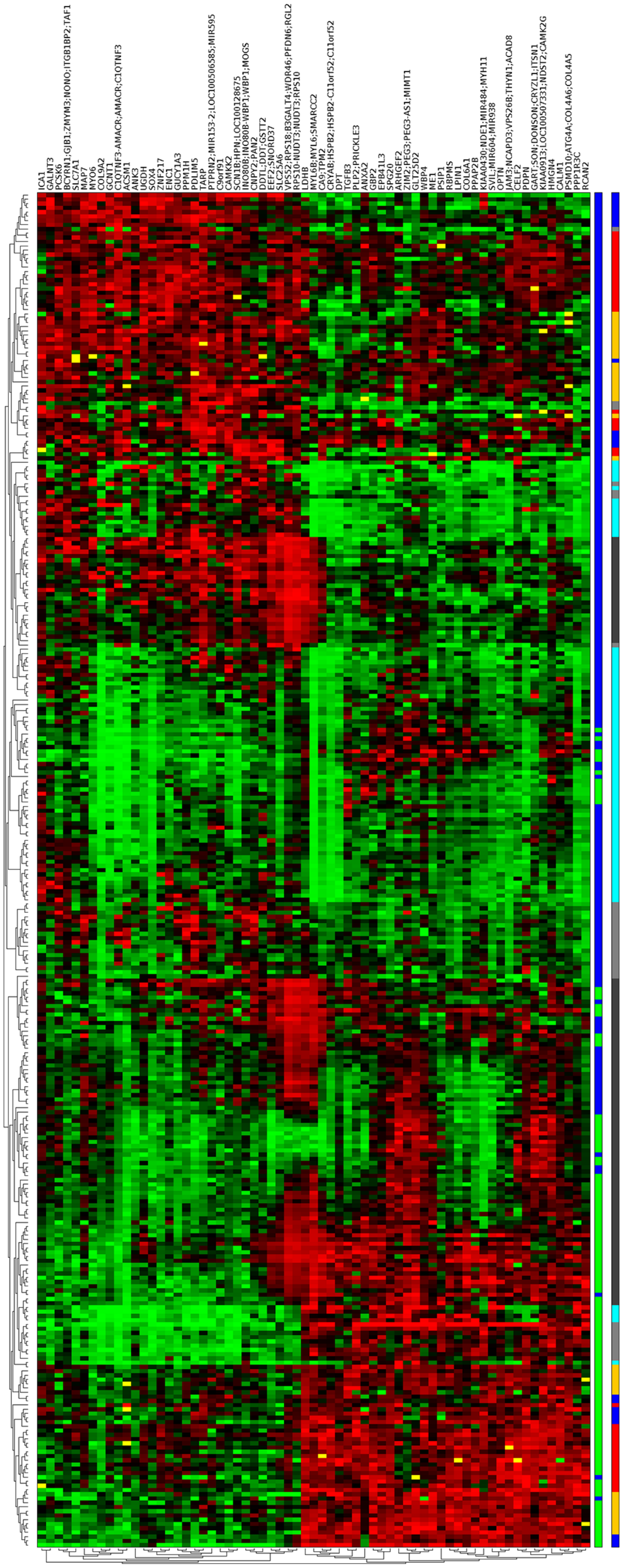

Fig 2. Heatmap for the Coloured *(α,β)-k*-Feature Selection resulted genes that cover six datasets. There are 120 up and down regulated genes (columns) which are differentially expressed between normal and tumour classes. The two colour bars at the right represent the ordering of samples and sample groups, respectively, as explained in Fig 1.

doi:10.1371/journal.pone.0127702.g002

Table 6. Overlapping genes in t-test, Coloured *(α,β)-k* and *(α,β)-k*-Feature Selection.

Number of Datasets	*t*-test	*(α,β)-k*-Feature Selection	Coloured *(α,β)-k*-Feature Selection
Six	4	7	120
Five or more	22	57	327
Four or more	36	139	623

Number of Datasets shows the number of datasets considered to find the overlapping. ***t*-test** gives the number of overlapping genes in t-test results for the considered datasets, ***(α,β)-k* Feature Selection** gives the number of overlapping genes between individual *(α,β)-k* feature selection result for each case. **Coloured *(α,β)-k*-Feature Selection** gives the number of common genes in the result of Coloured *(α,β)-k*-feature selection considered case of datasets. For method details refer to Section 2.

doi:10.1371/journal.pone.0127702.t006

To find the prostate cancer related pathways, we performed a pathway analysis using databases like DAVID [50], KEGG [52], FatiGO [53]. The resulted pathways with significant p-value are given in Table 9. A Pubmed search confirmed that all the resulting pathways are related to prostate cancer. Our analysis also identified several other genes that have no related publications in relation with prostate cancer.

Discussion

The microarray technology has a tremendous impact on cancer research in assessing the presence of cancer cells in patient tissues. The rapid acquisition of microarray data makes it possible to integrate this large amount of data across a range of platforms. In this study, we identified robust cancer gene expression signatures common to all datasets. The comparison of our proposed method with individual study results highlights the advantages of meta-analysis over individual studies. The comparison of our method with one of the state of the art methods shows the robustness of this method.

Results of *(α,β)-k*-Feature Set selection for each individual dataset provide signatures of reasonable size capable of discriminating between primary tumours and normal samples. However even though individual signatures consist of a large number of features, the number of common genes is limited to seven and is too little for further analysis. The result of Coloured *(α,β)-k*-Feature Set problem shows a vast difference in the number of resulting genes and is reliable for further analysis. The combined dataset contains 10729 out of 16157 combinations of probes from three or more datasets. That confirms we get a good coverage on all the datasets by using the proposed method of integration. Furthermore, the Coloured *(α,β)-k*-Feature Set problem results show that around 2272 out of 3190 resulted features cover four or more datasets.

Table 7. Comparison of Coloured *(α,β)-k*-Feature Set problem result and RankProd result.

	RankProd			No of CABK resulted genes		
Dataset	No of genes as input	pfp Cut off	No of resulted genes	Six datasets (120)	Five datasets (327)	Four datasets(623)
Combined dataset	6929	0.05	1883	80	169	260
	6929	0.01	1484	58	140	214

RankProd is the result of RankProd for Combined dataset with 0.05 and 0.01 pfp (percentage of false positive likelihood cut-off). **No of CABK resulted genes** is the number of genes resulted from Coloured *(α,β)-k*-Feature Set problem which covered six, five and more, four and more datasets.

doi:10.1371/journal.pone.0127702.t007

Table 8. Result of sensitivity analysis.

	Case a			Case b	
	Exp-1 (1 gene)	**Exp-2 (2 genes)**	**Exp-3 (5 genes)**	**Exp-4 (1 gene / 1DS)**	**Exp-5 (1 gene / 2DS)**
Average Signature Length	3203.1 (28.68)	3201.4 (28.21)	3204.9 (9.48)	3190.2 (0.45)	3190.8(1.30)
Average % Overlap with Original	97.42 (3.31)	97.62 (3.03)	98.04 (0.51)	99.41 (0.52)	99.15 (0.63)
Average Number of New Features	46.1 (11.47)	50.5 (17.67)	77.3 (17.97)	18.6 (16.80)	27.6 (20.98)
Average Cover of New Features	3.6	3.17	3.4	1.47	1.4
Average Signature length variation	0.41%	0.36%	0.47%	0.01%	0.03%

Case a is the result of sensitivity analysis after removing one gene (**Exp-1**), two genes (**Exp-2**) and five genes (**Exp-3**) from the combined dataset. **Case b** gives the result of sensitivity analysis after removing one gene from one (**Exp-4**) and two (**Exp-5**) individual datasets. For more details refer to Section 2. Values in parenthesis are the standard deviations for the 10 repetitions (Case a) and 5 repetitions (Case b).

doi:10.1371/journal.pone.0127702.t008

Even though the *(α,β)-k*-Feature Set methodology and *t*-test provided good results on individual datasets, a large number of genes have been eliminated from the common set of genes which may include potential biomarkers. So when we performed the data integration instead of considering common genes on individual dataset results, the number of resulting genes is significantly increased. This shows that the proposed method makes it possible to uncover robust biomarkers by increasing the sample size to a sufficient level and helps to capture the consistent features that might have been masked because of the limitations of individual studies. As we have a reasonable number of genes, they can also provide more information about prostate cancer.

The result of Coloured *(α,β)-k*-Feature Set problem evidences a high level of agreement with the top listed genes of RankProd result, where almost 80% of our signature is included in RankProd's result. However, RankProd results are considerably larger in size, hindering interpretation. Additionally, as mentioned before, RankProd artificially reduces the rank of any gene with missing values (escalating its position to the significant side of the list), which: i) restricts

Table 9. The top 14 resulted pathways from pathway analysis.

Pathway name	Pathway Classification	P-value	Reference
Integrin signalling pathway	Cell communication	1.03E-08	[85, 95]
Smooth Muscle Contraction	Organismal Systems; Circulatory system	5.98E-08	[96, 97]
Oxytocin signalling pathway	Organismal Systems; Endocrine system	1.23E-08	[98–100]
Collagen biosynthesis and modifying enzymes	Metabolism; Amino acid metabolism	1.13E-07	[88, 101]
Axon guidance	Development	1.34E-06	[102]
Gap junction trafficking	Cell communication	3.12E-06	[103, 104]
Protein digestion and absorption	Organismal Systems; Digestive system	3.6E-06	[101]
Ras activation	Regulation of translation and transcription	3.46E-05	[105, 106]
regulation of pgc-1a	Cell motility	3.61E-05	[107]
Assembly of collagen fibrils and other multimeric structures	Metabolism	3.68E-05	[108]
CREB phosphorylation	Metabolism; Energy metabolism	6.31E-05	[109]
Syndecan-1-mediated signalling events	Genetic Information Processing	6.3E-05	[110]
NCAM1 interactions	Signal Transduction	6.72E-05	[111]
regulators of bone mineralization	Metabolism	6.7E-05	[112]

Pathway Name is the name of the pathways. **Pathway Classification** is the class of each pathway. **P-value** is the respective p-value for each pathway. **Reference** is the papers which show the relation of each pathway with prostate cancer.

doi:10.1371/journal.pone.0127702.t009

applicability to the genes represented in all platforms, and ii) introduces non-linear rank scaling in the presence of scattered missing values. In contrast, Coloured *(α,β)-k*-Feature Set methodology automatically deals with any amount of missing values (that is, a gene may not be present in a dataset but still be significant to explain a large number of sample pairs in the other datasets), providing a more reliable result. Although not used in our investigation, Coloured *(α,β)-k*-Feature Set methodology allows for weights to be assigned to genes and samples independently, and account for an external perceived relative confidence in each experimental condition, if so desired.

The sensitivity analysis shows a high level of consistency with the original solution. Each step of the sensitivity analysis confirms that the proposed method is not relying on a single or a small set of genes. The results of the analysis also show that the significance of the gene is not dependent on a single dataset. The consistency of the results shows the robustness of our proposed method and validates the findings.

It is not surprising that most of the signature genes have been reported to be related with prostate cancer. For instance, AMACR [54–59], HPN [60–62], SOX4 [63–67], DAXX [68, 69], EPB41L3 [70–72], CXCR3 [73–79], TGFB3 [80, 81], EEF2 [82, 83] are the most well-known biomarkers for prostate cancer. As defined by the Gene Ontology Consortium, most of the resulted genes are involved in cell cycle (MYH11), regulation of transcription (SOX4, SMARCC2, ZIM2, PDLIM5, ZNF217, PSIPI, ACRC, PEG3, TAF1, ZMYM3), receptor activity (JAM3, TAPBP, COL4A5, CXCR3, COL4A6, HPN, COL9A2, PTPRN2, COL6A1) and other biological activities like transportation, cell adhesion and cell organisation (the list of genes with the related literature references can be found in S8 Table).

Most of the genes mentioned above and in S8 Table are highly correlated with prostate cancer. However we could find only some of them in the individual dataset results. We have also uncovered genes which participate in the same pathway class as genes related to prostate cancer, but have not yet been reported in relation to prostate cancer. For instance, the gene NUDT3 is not yet reported in relation to prostate cancer, but NUDT3 participates in the Collagen biosynthesis and modifying enzymes pathway which has been identified as a prostate cancer related pathway [84]. This indirectly suggests that this gene may also have some influence on cancer development.

Interestingly, the most significant pathway overrepresented in our results is the Integrin signalling pathway and focal adhesion. Integrins are transmembrane receptors and play an important role in cell survival, proliferation, migration, gene expression, and activation of growth factor receptors. Studies show that integrins are down regulated in the transition from normal prostate tissue to primary localized prostate cancer [85]. From our resulted list of genes COL4A5, COL4A6, COL6A1 and ITGB1BP2 are participating in integrin signalling pathway.

Smooth muscles found in the walls of reproductive tract of male and female which is made up of actin and myosin, together have the capacity to contract and relax. The prostate helps to control urine flow and ejaculations, via contractions and relaxation of its smooth muscle layers. The uncontrolled contraction of prostate smooth muscle may result in urinary tract problems in addition to prostate growth [86]. The smooth muscle contraction pathway has already been reported related with prostate cancer [87]. From our list of genes MYH11, MYL6, MYL6B and GUCY1A3 are related with smooth muscle contraction.

Collagen biosynthesis is the biosynthetic pathway responsible for collagen production. Studies have shown that the Gleason sum is increasing with decreasing cancer collagen content [88]. From our list of genes TGFB, COL4A5, COL4A6, COL6A1 and COL9A2 are related with collagen biosynthesis.

The outcomes of our work support the claim that the proposed method is a viable meta-analysis method for feature selection. The functional and pathway analysis results show that the Coloured (α,β)-k-Feature Set approach is capable of uncovering genes with significant and biologically relevant functions that other, non-integrative methods fail to identify.

Conclusion

We have presented the Coloured (α,β)-k-Feature Set problem as a combinatorial optimisation approach for multi-platform integration analysis without the need for normalisation of the data between datasets. The results indicate that the method is capable of providing highly significant signatures, even where the individual datasets before integration are small and thus lacking informational content. The method is generic and does not depend on inherent properties of gene expression data, allowing it to be potentially applied to any dataset where the notions of features, class based classification and equality between feature values is meaningful. In applying this methodology to an integrated prostate cancer dataset we have identified potential novel prostate cancer associated pathways and genes. As the number of cancer datasets increases we will be able to use this novel and robust method to combine more cancer datasets and identify more candidate pathways and genes.

Supporting Information

S1 Table. List of combination of probes. List of combination of probes resulted by applying the integration method. The probes are selected according to the conditions and the selected probe ID is given with the corresponding dataset name. The table also contains the gene names correspond to each combination of probe. (XLS)
(XLS)

S2 Table. Individual *(α,β)-k*-Feature Set problem Results. The list of genes resulted by applying *(α,β)-k*-Feature Set methodology on individual datasets. Single XLS contains six worksheets, one for each dataset result. The worksheets are names according to the dataset name.
(XLS)

S3 Table. *t*-test result. List of genes resulted after applying *t*-test on each dataset. XLS contains six worksheets, for each dataset. Each worksheet is named according to the dataset name.
(XLS)

S4 Table. Coloured *(α,β)-k*-Feature Set problem Result. The list of 3190 combined probes and 1788 genes resulted after applying the Coloured *(α,β)-k*-Feature Set methodology on the combined dataset. Also another worksheet with the annotation result of 1788 genes.
(XLS)

S5 Table. List of common genes resulted from Coloured (α,β)-k-Feature Set problem. An XLS file contains the list of 120 genes which are common in all the six datasets and the annotation details with the heatmap in sheet 1 and the list of 186 genes common in five or more datasets with annotation details and heatmap in sheet 2.
(XLS)

S6 Table. RankProd Result. The list of RankProd resulted genes with related rank, pfp and p-value.
(XLS)

S7 Table. Result of functional analysis. The details of eight clusters resulted from functional analysis using DAVID.
(XLS)

S8 Table. List of 120 genes with related literature references. Word document with the list of 120 genes which are common in all six datasets and the related literature references (DOCX)

Acknowledgments

PM is supported by Australian Research Council (ARC, http://www.arc.gov.au/) Future Fellowship FT120100060. This project is partially funded by ARC Discovery Project DP120102576, Australia.

Author Contributions

Conceived and designed the experiments: NP CR PM. Performed the experiments: NP CR. Analyzed the data: NP CR RB PM. Contributed reagents/materials/analysis tools: NP CR. Wrote the paper: NP CR RB PM. Developed software: CR.

References

1. Normand S-LT. Meta-analysis: formulating, evaluating, combining, and reporting. Statistics in Medicine. 1999; 18(3):321–59. doi: 10.1002/(SICI)1097-0258(19990215)18:3<321::AID-SIM28>3.0.CO;2-P PMID: 10070677.
2. Guerra R, Goldstein DR. Meta-analysis and combining information in genetics and genomics. Boca Raton: CRC Press; 2010. xxiii, 335 p. p.
3. Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics. 2008; 24(3):374–82. doi: 10.1093/bioinformatics/btm620 PMID: 18204063.
4. Nelson PS. Predicting prostate cancer behavior using transcript profiles. The Journal of Urology. 2004; 172(5, Supplement):S28—S33. PMID: 15535439
5. Granlund AvB, Flatberg A, Ostvik AE, Drozdov I, Gustafsson BI, Kidd M, et al. Whole Genome Gene Expression Meta-Analysis of Inflammatory Bowel Disease Colon Mucosa Demonstrates Lack of Major Differences between Crohn's Disease and Ulcerative Colitis. PLoS ONE. 2013; 8(2):e56818. doi: 10.1371/journal.pone.0056818 PMID: 23468882; PubMed Central PMCID: PMC3572080.
6. Kothapalli R, Yoder S, Mane S, Loughran T. Microarray results: how accurate are they? BMC Bioinformatics. 2002; 3(1):22.
7. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 2005; 21(2):171–8. PMID: 15308542
8. Cahan P, Rovegno F, Mooney D, Newman JC, St Laurent G III, McCaffrey TA. Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. Gene. 2007; 401(1–2):12–8. doi: 10.1016/j.gene.2007.06.016 PMID: 17651921; PubMed Central PMCID: PMC2111172.
9. Mukherjee S, Tamayo P, Rogers S, Rifkin R, Engle A, Campbell C, et al. Estimating dataset size requirements for classifying DNA Microarray data. Journal of Computational Biology. 2003; 10(2):119–42. doi: 10.1089/106652703321825928 PMID: 12804087.
10. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101(25):9309–14. PMID: 15184677
11. Xu L, Geman D, Winslow R. Large-scale integration of cancer microarray data identifies a robust common cancer signature. BMC Bioinformatics. 2007; 8(1):275.
12. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, et al. Multiple-laboratory comparison of microarray platforms. Nat Meth. 2005; 2(5):345–50. doi: 10.1038/nmeth756 PMID: 15846361; PubMed Central PMCID: PMC10.1038/nmeth756.
13. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CMT, Beyene J. Data Integration in Genetics and Genomics: Methods and Challenges. Human Genomics and Proteomics. 2009; 1(1). doi: 10.4061/2009/869093 PMID: 20948564; PubMed Central PMCID: PMC2950414.

14. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS letters. 2004; 573 (1–3):83–92. doi: 10.1016/j.febslet.2004.07.055 PMID: 15327980.

15. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics. 2006; 22 (22):2825–7. doi: 10.1093/bioinformatics/btl476 PMID: 16982708.

16. Choi H, Shen R, Chinnaiyan AM, Ghosh D. A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. BMC Bioinformatics. 2007; 8:364. doi: 10.1186/1471-2105-8-364 PMID: 17900369; PubMed Central PMCID: PMC2246152.

17. Rybalov M, Ananias HJ, Hoving HD, van der Poel HG, Rosati S, de Jong IJ. PSMA, EpCAM, VEGF and GRPR as imaging targets in locally recurrent prostate cancer after radiotherapy. International journal of molecular sciences. 2014; 15(4):6046–61. Epub 2014/04/15. doi: 10.3390/ijms15046046 PMID: 24727373; PubMed Central PMCID: PMC4013614.

18. Cope L, Zhong X, Garrett E, Parmigiani G. MergeMaid: R tools for merging and cross-study validation of gene expression data. Stat Appl Genet Mol Biol. 2004; 3:Article29. doi: 10.2202/1544-6115.1046 PMID: 16646808.

19. Cutillo L, Carissimo A, Figini S. Network selection: a method for ranked lists selection. PLoS One. 2012; 7(8):e43678. doi: 10.1371/journal.pone.0043678 PMID: 22937075; PubMed Central PMCID: PMC3427185.

20. Berretta R, Mendes A, Moscato P, editors. Integer Programming Models and Algorithms for Molecular Classification of Cancer from Microarray Data. Twenty-Eighth Australasian Computer Science Conference (ACSC2005); 2005; Newcastle, Australia: ACS.

21. Moscato P, Berretta R, Hourani MA, Mendes A, Cotta C. Genes Related with Alzheimers Disease: A Comparison of Evolutionary Search, Statistical and Integer Programming Approaches. In: Rothlauf F, Branke Jr, Cagnoni S, Corne D, Drechsler R, Jin Y, et al., editors. Applications of Evolutionary Computing. Lecture Notes in Computer Science. 3449: Springer Berlin Heidelberg; 2005. p. 84–94.

22. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell. 2002; 1:203–9-. PMID: 12086878

23. Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodriguez J, Moskaluk CA, et al. Analysis of Gene Expression Identifies Candidate Markers and Pharmacological Targets in Prostate Cancer. Cancer Research. 2001; 61:5974–8-. PMID: 11507037

24. Uma C, Ma C, Dhir R, Bisceglia M, Lyons-Weiler M, Liang W, et al. Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. BMC Cancer. 2007; 7:64-. PMID: 17430594

25. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci U S A. 2004; 101(3):811–6-. doi: 10.1073/pnas.0304146101 PMID: 14711987; PubMed Central PMCID: PMCLapointe, Jacques Li, Chunde Higgins, John P van de Rijn, Matt Bair, Eric Montgomery, Kelli Ferrari, Michelle Egevad, Lars Rayford, Walter Bergerheim, Ulf Ekman, Peter DeMarzo, Angelo M Tibshirani, Robert Botstein, David Brown, Patrick O Brooks, James D Pollack, Jonathan R.

26. Fayyad U, Irani K, editors. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. Proceedings of the 13th International Joint Conference on Artificial Intelligence; 1993; Chambery, France.

27. Cotta C, Sloper C, Moscato P. Evolutionary Search of Thresholds for Robust Feature Set Selection: Application to the Analysis of Microarray Data. In: Raidl Gn, Cagnoni S, Branke Jr, Corne D, Drechsler R, Jin Y, et al., editors. Applications of Evolutionary Computing. Lecture Notes in Computer Science. 3005: Springer Berlin Heidelberg; 2004. p. 21–30.

28. Berretta R, Costa W, Moscato P. Combinatorial Optimization Models for Finding Genetic Signatures from Gene Expression Datasets. In: Keith J, editor. Bioinformatics. Methods in Molecular Biologyâ„¢. 453: Humana Press; 2008. p. 363–77.

29. Hourani M, Mendes A, Berretta R, Moscato P. Genetic biomarkers for brain hemisphere differentiation in Parkinson's Disease. AIP Conference Proceedings. 2007; 952(1):207–16. PMID: WOS:000252100400022.

30. Gomez Ravetti M, Berretta R, Moscato P. Novel Biomarkers for Prostate Cancer Revealed by (α,β)—k Feature Sets. In: Abraham A, Hassanien A-E, Snasel V, editors. Foundations of Computational Intelligence Volume 5. Studies in Computational Intelligence. 205: Springer Berlin Heidelberg; 2009. p. 149–75.

31. Gomez Ravetti M, Moscato P. Identification of a 5-Protein Biomarker Molecular Signature for Predicting Alzheimer's Disease. PLoS ONE. 2008; 3(9):e3111. doi: 10.1371/journal.pone.0003111 PMID: 18769539; PubMed Central PMCID: PMC2518833.

32. Berretta R, Mendes A, Moscato P. Selection of Discriminative Genes in Microarray Experiments Using Mathematical Programming. Journal of Research & Practice in Information Technology. 2007; 39(4):287–99. PMID: WOS:000250650600005.

33. Milward EA, Moscato P, Riveros C, Johnstone DM. Beyond Statistics: A New Combinatorial Approach to Identifying Biomarker Panels for the Early Detection and Diagnosis of Alzheimer's Disease. Journal of Alzheimers Disease. 2014; 39(1):211–7. doi: 10.3233/Jad-131424 PMID: WOS:000329506000020.

34. Riveros C, Mellor D, Gandhi KS, McKay FC, Cox MB, Berretta R, et al. A transcription factor map as revealed by a genome-wide gene expression analysis of whole-blood mRNA transcriptome in multiple sclerosis. PLoS One. 2010; 5(12):e14176. doi: 10.1371/journal.pone.0014176 PMID: 21152067; PubMed Central PMCID: PMC2995726.

35. Downey RG, Fellows MR. Parameterized Complexity: Springer Verlag; 1999.

36. Flum J. Parameterized Complexity Theory: Springer; 2006.

37. Davies S, Russell S. NP-completeness of searches for smallest possible feature sets. 1994 AAAI Fall Symposium on Relevance. Menlo Park, CA: The AAAI Press; 1994. p. 37–9.

38. Cotta C, Moscato P. The k-Feature Set problem is W[2]-complete. Journal of Computer and System Sciences. 2003; 67(4):686–90. doi: 10.1016/S0022-0000(03)00081-3 PMID: WOS:000187019400003.

39. Mary-Huard T, Picard F, Robin S. Introduction to statistical methods for microarray data analysis. Mathematical and Computational Methods in Biology. 2006:56–126.

40. Gentleman RCV, Huber W, Hahne F. genefilter: methods for filtering genes from microarray experiments. R Package Version 1.24.2. ed2009.

41. Guo C, Liu H, Zhang B-H, Cadaneanu RM, Mayle AM, Garraway IP. Epcam, CD44, and CD49f distinguish sphere-forming human prostate basal cells from a subpopulation with predominant tubule initiation capability. PloS one. 2012; 7(4):e34219. Epub 2012/04/20. doi: 10.1371/journal.pone.0034219 PMID: 22514625; PubMed Central PMCID: PMC3326009.

42. Massoner P, Thomm T, Mack B, Untergasser G, Martowicz A, Bobowski K, et al. EpCAM is overexpressed in local and metastatic prostate cancer, suppressed by chemotherapy and modulated by MET-associated miRNA-200c/205. British journal of cancer. 2014; 111(5):955–64. Epub 2014/07/06. doi: 10.1038/bjc.2014.366 PMID: 24992580; PubMed Central PMCID: PMC4150273.

43. Xu Y, Zhao H, Hou J. Correlation between overexpression of EpCAM in prostate tissues and genesis of androgen-dependent prostate cancer. Tumour biology: the journal of the International Society for Oncodevelopmental Biology and Medicine. 2014; 35(7):6695–700. Epub 2014/04/08. doi: 10.1007/s13277-014-1892-2 PMID: 24705864.

44. Fong D, Seeber A, Terracciano L, Kasal A, Mazzoleni G, Lehne F, et al. Expression of EpCAM(MF) and EpCAM(MT) variants in human carcinomas. Journal of clinical pathology. 2014; 67(5):408–14. Epub 2014/01/28. doi: 10.1136/jclinpath-2013-201932 PMID: 24465008; PubMed Central PMCID: PMC3995261.

45. Zhu B, Wu G, Robinson H, Wilganowski N, Hall MA, Ghosh SC, et al. Tumor margin detection using quantitative NIRF molecular imaging targeting EpCAM validated by far red gene reporter iRFP. Molecular imaging and biology: MIB: the official publication of the Academy of Molecular Imaging. 2013; 15 (5):560–8. Epub 2013/04/27. doi: 10.1007/s11307-013-0637-8 PMID: 23619897.

46. Gorges TM, Tinhofer I, Drosch M, Rose L, Zollner TM, Krahn T, et al. Circulating tumour cells escape from EpCAM-based detection due to epithelial-to-mesenchymal transition. BMC cancer. 2012; 12:178. Epub 2012/05/18. doi: 10.1186/1471-2407-12-178 PMID: 22591372; PubMed Central PMCID: PMC3502112.

47. Benko G, Spajic B, Kruslin B, Tomas D. Impact of the EpCAM expression on biochemical recurrence-free survival in clinically localized prostate cancer. Urologic oncology. 2013; 31(4):468–74. Epub 2011/04/26. doi: 10.1016/j.urolonc.2011.03.007 PMID: 21514185.

48. Riesenberg R, Buchner A, Pohla H, Lindhofer H. Lysis of prostate carcinoma cells by trifunctional bispecific antibodies (alpha EpCAM x alpha CD3). The journal of histochemistry and cytochemistry: official journal of the Histochemistry Society. 2001; 49(7):911–7. Epub 2001/06/19. PMID: 11410615.

49. Moscato P, Mendes A, Berretta R. Benchmarking a memetic algorithm for ordering microarray data. Biosystems. 2007; 88(1–2):56–75. doi: 10.1016/j.biosystems.2006.04.005 PMID: 16870322.

50. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protocols. 2008; 4(1):44–57. doi: 10.1038/nprot.2008.211

51. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research. 2013;

41(D1):D808–D15. doi: 10.1093/nar/gks1094 PMID: 23203871; PubMed Central PMCID: PMC3531103.

52. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Research. 2014; 42(D1):D199–D205. doi: 10.1093/nar/gkt1076 PMID: 24214961; PubMed Central PMCID: PMC3965122.

53. Al-Shahrour F, Minguez P, Tarraga J, Medina I, Alloza E, Montaner D, et al. FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. Nucleic Acids Research. 2007; 35(suppl 2):W91–W6. doi: 10.1093/nar/gkm260 PMID: 17478504; PubMed Central PMCID: PMC1933151.

54. Jiang N, Zhu S, Chen J, Niu Y, Zhou L. A-Methylacyl-CoA Racemase (AMACR) and Prostate-Cancer Risk: A Meta-Analysis of 4,385 Participants. PLoS ONE. 2013; 8(10):e74386. doi: 10.1371/journal.pone.0074386 PMID: 24130666; PubMed Central PMCID: PMC3794046.

55. Ananthanarayanan V, Deaton RJ, Yang XJ, Pins MR, Gann PH. Alpha-methylacyl-CoA racemase (AMACR) expression in normal prostatic glands and high-grade prostatic intraepithelial neoplasia (HGPIN): Association with diagnosis of prostate cancer. The Prostate. 2005; 63(4):341–6. doi: 10.1002/pros.20196 PMID: 15602744.

56. Hsieh C-l, Oakley-Girvan I, Balise RR, Halpern J, Gallagher RP, Wu AH, et al. A Genome Screen of Families with Multiple Cases of Prostate Cancer: Evidence of Genetic Heterogeneity. The American Journal of Human Genetics. 2001; 69(1):148–58. doi: 10.1086/321281 PMID: 11404817; PubMed Central PMCID: PMC1226029.

57. Zheng SL, Chang B-l, Faith DA, Johnson JR, Isaacs SD, Hawkins GA, et al. Sequence Variants of α-Methylacyl-CoA Racemase Are Associated with Prostate Cancer Risk. Cancer Research. 2002; 62 (22):6485–8. PMID: 12438241

58. Luo J, Zha S, Gage WR, Dunn TA, Hicks JL, Bennett CJ, et al. α-Methylacyl-CoA Racemase: A New Molecular Marker for Prostate Cancer. Cancer Research. 2002; 62(8):2220–6. PMID: 11956072.

59. Rubin MA, Zhou M, Dhanasekaran SM, Varambally S, Barrette TR, Sanda MG, et al. A-methylacyl coenzyme a racemase as a tissue biomarker for prostate cancer. JAMA. 2002; 287(13):1662–70. doi: 10.1001/jama.287.13.1662 PMID: 11926890

60. Ganesan R, Kolumam GA, Lin SJ, Xie M-H, Santell L, Wu TD, et al. Proteolytic activation of pro-macrophage-stimulating protein by hepsin. Molecular cancer research: MCR. 2011; 9(9):1175–86. Epub 2011/08/31. doi: 10.1158/1541-7786.MCR-11-0004 PMID: 21875933.

61. Guo J, Li G, Tang J, Cao XB, Zhou QY, Fan ZJ, et al. HLA-A2-restricted cytotoxic T lymphocyte epitopes from human hepsin as novel targets for prostate cancer immunotherapy. Scandinavian journal of immunology. 2013; 78(3):248–57. Epub 2013/06/01. doi: 10.1111/sji.12083 PMID: 23721092.

62. Kim HJ, Han JH, Chang IH, Kim W, Myung SC. Variants in the HEPSIN gene are associated with susceptibility to prostate cancer. Prostate cancer and prostatic diseases. 2012; 15(4):353–8. Epub 2012/06/06. doi: 10.1038/pcan.2012.17 PMID: 22665141.

63. Wang L, Zhang J, Yang X, Chang YW, Qi M, Zhou Z, et al. SOX4 is associated with poor prognosis in prostate cancer and promotes epithelial-mesenchymal transition in vitro. Prostate cancer and prostatic diseases. 2013; 16(4):301–7. Epub 2013/08/07. doi: 10.1038/pcan.2013.25 PMID: 23917306.

64. Lai Y-H, Cheng J, Cheng D, Feasel ME, Beste KD, Peng J, et al. SOX4 interacts with plakoglobin in a Wnt3a-dependent manner in prostate cancer cells. BMC cell biology. 2011; 12:50. Epub 2011/11/22. doi: 10.1186/1471-2121-12-50 PMID: 22098624; PubMed Central PMCID: PMC3227594.

65. Moreno CS. The Sex-determining region Y-box 4 and homeobox C6 transcriptional networks in prostate cancer progression: crosstalk with the Wnt, Notch, and PI3K pathways. The American journal of pathology. 2010; 176(2):518–27. Epub 2009/12/19. doi: 10.2353/ajpath.2010.090657 PMID: 20019190; PubMed Central PMCID: PMC2808058.

66. Scharer CD, McCabe CD, Ali-Seyed M, Berger MF, Bulyk ML, Moreno CS. Genome-wide promoter analysis of the SOX4 transcriptional network in prostate cancer cells. Cancer research. 2009; 69 (2):709–17. Epub 2009/01/17. doi: 10.1158/0008-5472.CAN-08-3415 PMID: 19147588; PubMed Central PMCID: PMC2629396.

67. Haram KM, Peltier HJ, Lu B, Bhasin M, Otu HH, Choy B, et al. Gene expression profile of mouse prostate tumors reveals dysregulations in major biological processes and identifies potential murine targets for preclinical development of human prostate cancer therapy. The Prostate. 2008; 68(14):1517–30. Epub 2008/08/01. doi: 10.1002/pros.20803 PMID: 18668517.

68. Kwan PS, Lau CC, Chiu YT, Man C, Liu J, Tang KD, et al. Daxx regulates mitotic progression and prostate cancer predisposition. Carcinogenesis. 2013; 34(4):750–9. doi: 10.1093/carcin/bgs391 PMID: 23239745.

69. Tsourlakis MC, Schoop M, Plass C, Huland H, Graefen M, Steuber T, et al. Overexpression of the chromatin remodeler death-domain–associated protein in prostate cancer is an independent predictor

of early prostate-specific antigen recurrence. Human Pathology. 2013; 44(9):1789–96. doi: 10.1016/j.humpath.2013.01.022 PMID: 23642739

70. Bernkopf DB, Williams ED. Potential role of EPB41L3 (Protein 4.1B/Dal-1) as a target for treatment of advanced prostate cancer. Expert Opinion on Therapeutic Targets. 2008; 12(7):845–53. doi: 10.1517/14728222.12.7.845 PMID: 18554153.

71. Schulz W, Alexa A, Jung V, Hader C, Hoffmann M, Yamanaka M, et al. Factor interaction analysis for chromosome 8 and DNA methylation alterations highlights innate immune response suppression and cytoskeletal changes in prostate cancer. Molecular Cancer. 2007; 6(1):14. doi: 10.1186/1476-4598-6-14

72. Schulz W, Ingenwerth M, Djuidje C, Hader C, Rahnenführer J, Engers R. Changes in cortical cytoskeletal and extracellular matrix gene expression in prostate cancer are related to oncogenic ERG deregulation. BMC Cancer. 2010; 10(1):1–9. doi: 10.1186/1471-2407-10-505

73. Engl T, Relja B, Blumenberg C, Müller I, Ringel EM, Beecken W-D, et al. Prostate tumor CXC-chemokine profile correlates with cell adhesion to endothelium and extracellular matrix. Life Sciences. 2006; 78(16):1784–93. doi: http://dx.doi.org/10.1016/j.lfs.2005.08.019. PMID: 16263140.

74. König JE, Senge T, Allhoff EP, König W. Analysis of the inflammatory network in benign prostate hyperplasia and prostate cancer. The Prostate. 2004; 58(2):121–9. doi: 10.1002/pros.10317 PMID: 14716737.

75. Nagpal ML, Chen Y, Lin T. Effects of overexpression of CXCL10 (cytokine-responsive gene-2) on MA-10 mouse Leydig tumor cell steroidogenesis and proliferation. Journal of Endocrinology. 2004; 183(3):585–94. doi: 10.1677/joe.1.05795 PMID: 15590984.

76. Nagpal ML, Davis J, Lin T. Overexpression of CXCL10 in human prostate LNCaP cells activates its receptor (CXCR3) expression and inhibits cell proliferation. Biochimica et Biophysica Acta (BBA)—Molecular Basis of Disease. 2006; 1762(9):811–8. doi: 10.1016/j.bbadis.2006.06.017 PMID: 16934957.

77. Shen H, Schuster R, Lu B, Waltz SE, Lentsch AB. Critical and opposing roles of the chemokine receptors CXCR2 and CXCR3 in prostate tumor growth. The Prostate. 2006; 66(16):1721–8. doi: 10.1002/pros.20476 PMID: 16941672

78. Wedel S, Raditchev I, Jones J, Juengel E, Engl T, Jonas D, et al. CXC chemokine mRNA expression as a potential diagnostic tool in prostate cancer. Molecular Medicine Reports. 2008; 1(2):257–62. PMID: 21479406

79. Wu Q, Dhir R, Wells A. Altered CXCR3 isoform expression regulates prostate cancer cell migration and invasion. Molecular Cancer. 2012; 11(1):3. doi: 10.1186/1476-4598-11-3 PMID: 22234808

80. Reinertsen T, Halgunset J, Viset T, Flatberg A, Haugsmoen LL, Skogseth H. Gene expressional changes in prostate fibroblasts from cancerous tissue. APMIS. 2012; 120(7):558–71. doi: 10.1111/j.1600-0463.2011.02865.x PMID: 22716211

81. Caggia S, Libra M, Malaponte G, Cardile V. Modulation of YY1 and p53 expression by transforming growth factor-β3 in prostate cell lines. Cytokine. 2011; 56:403–10. doi: 10.1016/j.cyto.2011.06.024 PMID: 21807531

82. Oji Y, Tatsumi N, Fukuda M, Nakatsuka S-I, Aoyagi S, Hirata E, et al. The translation elongation factor eEF2 is a novel tumorassociated antigen overexpressed in various types of cancers. International journal of oncology. 2014; 44(5):1461–9. Epub 2014/03/05. doi: 10.3892/ijo.2014.2318 PMID: 24589652; PubMed Central PMCID: PMC4027928.

83. Wullner U, Neef I, Eller A, Kleines M, Tur MK, Barth S. Cell-specific induction of apoptosis by rationally designed bivalent aptamer-siRNA transcripts silencing eukaryotic elongation factor 2. Current cancer drug targets. 2008; 8(7):554–65. doi: 10.2174/156800908786241078 PMID: 18991566.

84. Hamilton G, Olszewski-Hamilton U, Theyer G. Type I Collagen Synthesis Marker Procollagen I N-Terminal Peptide (PINP) in Prostate Cancer Patients Undergoing Intermittent Androgen Suppression. Cancers. 2011; 3(3):3601–9. doi: 10.3390/cancers3033601 PubMed Central PMCID: PMC3759212. PMID: 24212969

85. Gorlov I, Byun J, Gorlova O, Aparicio A, Efstathiou E, Logothetis C. Candidate pathways and genes for prostate cancer: a meta-analysis of gene expression data. BMC Medical Genomics. 2009; 2(1):48. doi: 10.1186/1755-8794-2-48 PubMed Central PMCID: PMC2731785.

86. Schilit S, Benzeroual KE. Silodosin: A selective α1A-adrenergic receptor antagonist for the treatment of benign prostatic hyperplasia. Clinical Therapeutics. 2009; 31(11):2489–502. doi: 10.1016/j.clinthera.2009.11.024 PMID: 20109995

87. Strittmatter F, Walther S, Gratzke C, Göttinger J, Beckmann C, Roosen A, et al. Inhibition of adrenergic human prostate smooth muscle contraction by the inhibitors of c-Jun N-terminal kinase, SP600125 and BI-78D3. British Journal of Pharmacology. 2012; 166(6):1926–35. doi: 10.1111/j.1476-5381.2012.01919.x PMID: 22364229

88. Burns-Cox N, Avery NC, Gingell JC, Bailey AJ. Changes in collagen metabolism in prostate cancer: a host response that may alter progression. The Journal of Urology. 2001; 166(5):1698–701. doi: http://dx.doi.org/10.1016/S0022-5347(05)65656-X. PMID: WOS:000171547200021.

89. Qi M, Yang X, Zhang F, Lin T, Sun X, Li Y, et al. ERG rearrangement is associated with prostate cancer-related death in Chinese prostate cancer patients. PloS one. 2014; 9(2):e84959. Epub 2014/02/12. doi: 10.1371/journal.pone.0084959 PMID: 24516518; PubMed Central PMCID: PMC3917829.

90. Wang L, Li Y, Yang X, Yuan H, Li X, Qi M, et al. ERG-SOX4 interaction promotes epithelial-mesenchymal transition in prostate cancer cells. The Prostate. 2014; 74(6):647–58. Epub 2014/01/18. doi: 10.1002/pros.22783 PMID: 24435928.

91. Andrews C, Humphrey PA. Utility of ERG versus AMACR expression in diagnosis of minimal adenocarcinoma of the prostate in needle biopsy tissue. The American journal of surgical pathology. 2014; 38(7):1007–12. Epub 2014/04/08. doi: 10.1097/PAS.0000000000000205 PMID: 24705308.

92. Klee EW, Bondar OP, Goodmanson MK, Dyer RB, Erdogan S, Bergstralh EJ, et al. Candidate Serum Biomarkers for Prostate Adenocarcinoma Identified by mRNA Differences in Prostate Tissue and Verified with Protein Measurements in Tissue and Blood. Clinical Chemistry. 2012; 58(3):599–609. doi: 10.1373/clinchem.2011.171637 PMID: 22247499; PubMed Central PMCID: PMC3951013.

93. Yamamoto-Ishikawa K, Suzuki H, Nezu M, Kamiya N, Imamoto T, Komiya A, et al. The isolation and identification of apolipoprotein C-I in hormone-refractory prostate cancer using surface-enhanced laser desorption/ionization time-of-flight mass spectrometry. Asian journal of andrology. 2009; 11 (3):299–307. doi: 10.1038/aja.2008.38 PMID: 19182819.

94. Dong Y, Zhang H, Gao AC, Marshall JR, Ip C. Androgen receptor signaling intensity is a key factor in determining the sensitivity of prostate cancer cells to selenium inhibition of growth and cancer-specific biomarkers. Molecular Cancer Therapeutics. 2005; 4(7):1047–55. doi: 10.1158/1535-7163.Mct-05-0124 PMID: WOS:000230537500003.

95. Mendes A, Scott R, Moscato P. Microarrays—Identifying Molecular Portraits for Prostate Tumors with Different Gleason Patterns. In: Trent RA, editor. Clinical Bioinformatics. Methods in Molecular Medicine. 141: Humana Press; 2008. p. 131–51. PMID: 18453088

96. Belldegrun A, Oliver PT, Kirby RS, Newling DWW. New Perspectives in Prostate Cancer: CRC Press; 1998. 350 p.

97. Weber GF. Molecular Mechanisms of Cancer: Springer; 2007.

98. Zhong M, Boseman ML, Millena AC, Khan SA. Oxytocin Induces the Migration of Prostate Cancer Cells: Involvement of the Gi-Coupled Signaling Pathway. Molecular Cancer Research. 2010; 8 (8):1164–72. doi: 10.1158/1541-7786.mcr-09-0329 PMID: 20663860

99. Thackare H, Nicholson HD, Whittington K. Oxytocin—its role in male reproduction and new potential therapeutic uses. Human Reproduction Update. 2006; 12(4):437–48. doi: 10.1093/humupd/dmk002 PMID: 16436468

100. Reversi A, Rimoldi V, Marrocco T, Cassoni P, Bussolati G, Parenti M, et al. The Oxytocin Receptor Antagonist Atosiban Inhibits Cell Growth via a "Biased Agonist" Mechanism. Journal of Biological Chemistry. 2005; 280(16):16311–8. doi: 10.1074/jbc.M409945200 PMID: 15705593

101. Acton QA. Prostate Cancer: New Insights for the Healthcare Professional: 2013 Edition: ScholarlyEditions; 2013.

102. Savli H, Szendroi A, Romics I, Nagy B. Gene network and canonical pathway analysis in prostate cancer: a microarray study. Exp Mol Med. 2008; 40:176–85. doi: 10.3858/emm.2008.40.2.176 PMID: 18446056

103. Govindarajan R, Zhao S, Song X-H, Guo R-J, Wheelock M, Johnson KR, et al. Impaired Trafficking of Connexins in Androgen-independent Human Prostate Cancer Cell Lines and Its Mitigation by α-Catenin. Journal of Biological Chemistry. 2002; 277(51):50087–97. doi: 10.1074/jbc.M202652200 PMID: 12205082.

104. Chakraborty S. Regulatory Aspects of Gap Junction Assembly: University of Nebraska Medical Center; 2008.

105. Min J, Zaslavsky A, Fedele G, McLaughlin SK, Reczek EE, De Raedt T, et al. An oncogene-tumor suppressor cascade drives metastatic prostate cancer by coordinately activating Ras and nuclear factor-[kappa]B. Nat Med. 2010; 16(3):286–94. doi: 10.1038/nm.2100 PMID: WOS:000275289500033.

106. Adjei AA. Blocking Oncogenic Ras Signaling for Cancer Therapy. Journal of the National Cancer Institute. 2001; 93(14):1062–74. PMID: 11459867.

107. Tennakoon JB, Shi Y, Han JJ, Tsouko E, White MA, Burns AR, et al. Androgens regulate prostate cancer cell growth via an AMPK-PGC-1[alpha]-mediated metabolic switch. Oncogene. 2013. doi: 10.1038/onc.2013.463

108. Zunich S, Valdovinos M, Douglas T, Walterhouse D, Iannaccone P, Lamm M. Osteoblast-secreted collagen upregulates paracrine Sonic hedgehog signaling by prostate cancer cells and enhances osteoblast differentiation. Molecular Cancer. 2012; 11(1):1–13. doi: 10.1186/1476-4598-11-30

109. Chung LWK, Isaacs WB, Simons JW. Prostate Cancer: Biology, Genetics, and the New Therapeutics: Humana Press; 2007.

110. Hu Y, Sun H, Owens RT, Gu Z, Wu J, Chen YQ, et al. Syndecan-1-Dependent Suppression of PDK1/Akt/Bad Signaling by Docosahexaenoic Acid Induces Apoptosis in Prostate Cancer. Neoplasia. 2010; 12(10):826–36. doi: http://dx.doi.org/10.1593/neo.10586. PMID: 20927321; PubMed Central PMCID: PMC2950332.

111. Li R, Wheeler T, Dai H, Ayala G. Neural cell adhesion molecule is upregulated in nerves with prostate cancer invasion. Human Pathology. 2003; 34(5):457–61. doi: 10.1016/S0046-8177(03)00084-4 PMID: WOS:000183190200008.

112. Yuen HF, Chiu YT, Chan KK, Chan YP, Chua CW, McCrudden CM, et al. Prostate cancer cells modulate osteoblast mineralisation and osteoclast differentiation through Id-1. Br J Cancer. 2009; 102 (2):332–41. doi: 10.1038/sj.bjc.6605480 PMID: 20010941