Advancing integrated bioinformatics analyses

Benchmark for gene set enrichment methods considering an increasing number of patient samples and multiple datasets in TCGA Kidney Renal Clear Cell Carcinoma

Marshall Plan scholar:	Maciej M Kańduła*
hosting supervisor:	Eric D Kolaczyk, Department of Mathematics and Statistics, Boston University, USA
home supervisor:	David P Kreil, Chair of Bioinformatics Research Group, Department of Biotechnology, BOKU University Vienna, Austria

*contact: maciej.kandula@gmail.com

Contents

Introduction

I here describe the research project which was the main focus of my work during my research visit in Boston. I was, moreover, able to establish further collaborations leading to two additional projects under way with researchers from the Department of Mathematics and Statistics of the Boston University and the Center for Regenerative Medicine in Boston. All this work is joined by the common theme of the application of multi-layer network approach to patient stratification and on finding factors responsible for stem cell differentiation into lungs.

Big Data science

Bioinformatics seeks insight from a multitude of data collected in the biomedical sciences. The multitude of data collected in the biomedical sciences (Benton, 1996; Mushegian, 2011) increasingly comes from high-throughput experiments and data sets are of genomic scale. The identification and interpretation of biologically relevant patterns in these data, however, remains a bottleneck for both basic and applied research, and has been rate limiting in the translation of experimental advances to the clinic.

Data integration

A lot of hope is now being placed in the integrated analysis of measurements from different sources (Searls, 2005), *i.e.*, the joint analysis of different data types. Analyses incorporating multiple sources of evidence have proven to be very informative for modelling biological systems (Hartemink, Gifford, Jaakkola, & Young, 2002; Hecker, Lambeck, Toepfer, van Someren, & Guthke, 2009; Nariai, Tamada, Imoto, & Miyano, 2005). With the addition of more measured variables, also more independent measurement samples (*e.g.*, patients) are required for meaningful analysis. It is partly due to the high cost of such large-scale experiments that cancer research has been at the forefront of collecting sufficiently many matched profiles, including systematic studies of gene expression and the activities of novel regulators like microRNAs, the accumulation of somatic mutations, the prevalence of DNA methylation, as well as copy number variation (CNV), all of which are known to play key roles in this disease. An efficient integrated analysis needs to address: (1) the technical challenge of linking heterogeneous data sources and third party analysis tools, and (2) the inference problem of identifying biomedically relevant patterns in extremely high-dimensional spaces (tens of thousands of variables) *vis-a-vis* moderately small sample sizes (hundreds of patients).

Reproducibility through automation of analyses

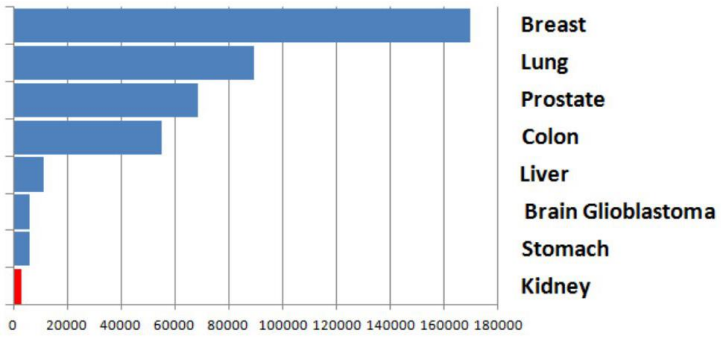
Many published research findings are false or exaggerated, and an estimated 85% of research resources are wasted (Ioannidis, 2014). It has thus been recognized that it is necessary to improve the reproducibility of research, which constitutes one of the key factors for increasing the proportion of true research findings. Reproducibility in science requires publication of not just the paper manuscript but also the original measurement data, all code of analysis software, and the exhaustive documentation typically needed to independently regenerate the results. Only that makes the analyses reusable and enables other researchers to validate if one can build on the analyses conclusions, employed algorithms, and the measurement data (Hothorn & Leisch, 2011). Usually multiple tools are applied at multiple stages in bioinformatics analyses, which can follow a sequential order, or include iterative elements and other flow control, including conditional execution. Such a multi-step analysis is commonly referred to as a workflow. Scientific Workflow Systems should aid in carrying out such procedures making the whole analysis reproducible through automated execution, documentation, and testing.

Building on my earlier work on light-weight modular workflow systems (Köster & Rahmann, 2012; Romano, 2008) for the control of the development cycle and data provenance, I have introduced a policy based specification of rules and requirements allowing in-flow enforcement of consistency constraints for audit and quality control. These, for instance, enabled the highly parallel execution of model-based optimization of assays for genome-scale transcript expression profiling experiments. Unexpected behaviour of both third party software, inconsistencies in heterogeneous external data sources, and a shared cluster environment could thus be isolated from the main analysis logic. For data integration, the workflow systems that I have developed will support quality control and systematic processing of the original data sources to unified scales and in general help in performing the computational analyses described below. In particular, I have employed complementary use-cases from sequence analysis and comparative genomics for validation of my methodological work.

Cancer research

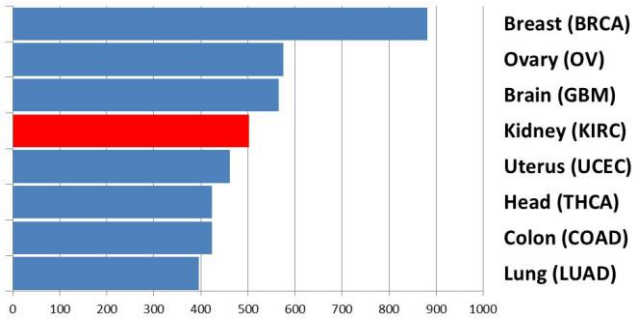
We focus on the analysis of a large collection of Kidney Renal Clear Cell Carcinoma data (KIRC). KIRC is also known as clear cell Renal Cell Carcinoma (ccRCC) and a large set of studies is available from the public ICGC / TCGA (www.icgc.org) data archives. Even though kidney cancer is the 7th (www.cancerresearchuk.org) most common cancer in males in the UK alone, and its incidence has increased by over 25% in the last decade, it has received relatively little attention in the scientific literature (Figure 1) and remains to be better characterized.

Figure 1: Number of publications (in PubMed online library) for selected cancers



The data recently compiled for this cancer covers an unusually large number of patients and data types (Figure 2). In particular, we search for pathways dysregulated in KIRC, playing a key role in cancer progression. Comparing tumour and normal tissue samples, the effect of the disease can be directly studied.

Figure 2: Sample numbers per cancer in ICGC (USA) database



Testing for gene set enrichment

Gene expression analysis has become a key tool for functional genomics, aiming to decode the blueprint of genomic DNA. Often, differential expression studies yield hundreds to thousands of affected genes. Testing for an enrichment of genesets of known functions is one of the most popular approaches to interpreting results (Khatri, Sirota, & Butte, 2012). This identifies, for instance, specific GeneOntology or KEGG pathways that are enriched in the affected genes.
The incorporation of additional molecular profiling data promises to further improve our ability to detect enriched pathways. Jointly analysing different data types could more sensitively identify canonical pathways from MSigDb that were expected from the literature than when each data type was analysed separately (Tyekucheva et al., 2011). In a recent publication Verbeke et al (Verbeke, Eynden, et al., 2015) propose a method for ranking pathways through a network-based data integration approach. Even in this most recent publication, and even though the authors show that the method yields results in agreement with a previous approach, no thorough benchmark is applied. This reflects well acknowledged challenges of benchmarking in the absence of a ‘ground truthâ€™ in the field (Alexeyenko et al., 2012; Glaab, Baudot, Krasnogor, Schneider, & Valencia, 2012) A key objective of my research work was therefore to develop and establish such an objective benchmark to fill this unaddressed need.

Latent Pathway Identification Analysis (LPIA)

It has been shown that network approaches can be helpful for understanding diseases better (Barabási, 2007; Papin et al., 2004; Silverman & Loscalzo, 2012, p. 2013; Wang et al., 2014). Latent pathway identification analysis (LPIA), introduced by Pham et al (Pham et al., 2011), combines measurements and existing knowledge by integrating structured information from several sources. Building on this, we here introduce and validate a novel network-based data integration approach for identifying metabolic pathways implicated by differential expression analysis and changes in DNA copy number variation (CNV). This way we shortlist biological functions that may be responsible for the observed patterns of complex transcriptional dysregulation in KIRC.

Originally, LPIA was designed to integrate measurements of gene expression with information about pathways from the KEGG database and functions from the GeneOntology database into a single annotated interconnected network of pathways. The method utilizes a stepwise approach where the final nodes represent pathways and the edges between nodes are weighted according to the strength of evidence for differential expression in genes relevant to the biological functions in which the corresponding biological pathways are involved. A statistical hypothesis testing framework is then used to determine pathways for which the network-wide evidence suggests significant changes in expression relevant to the phenotype of interest, and facilitates ranking of pathways. This way, the underlying cellular mechanisms of action in several studies have already been identified, including studies of prostate cancer metastasis (Pham et al., 2011).

We now incorporate evidence from complementary sources and thus support extended data-integration. In this work, we focus on adding CNV data. We examine two probabilistic approaches for data integration, testing for (1) all of the data sources showing a differential effect for the gene, or (2) at least one of the data sources showing a differential effect for the gene. Moreover, we implement parallel execution support for multi-core environments, which aids fast method prototyping, tests of different algorithm parameters / input sets, and tests of robustness (sub-sampling).

Methods

Data sources

The data used in the presented work was downloaded from the ICGC Data Portal v15.1, project: Kidney Renal Clear Cell Carcinoma - TCGA, US (https://dcc.icgc.org/projects/KIRC-US). We use the Sequence-based Gene Expression (EXP-S) data for, initially, 518 donors and Copy Number Somatic Mutations (CNSM) data for, initially, 522 donors. In our initial approach we only used matched Primary solid Tumor and Solid Tissue Normal samples, also matched between expression and CNV donors, resulting in 55 matched donors. We matched the samples through the column submitted_sample_id and using the codes described in Code Tables Report, under Sample Type (https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm) using the Clinical Data file.

Further we extended the dataset by the normal-unmatched tumour samples. These amounted to a total of 483 tumour and 55 normal samples - patient data that was available for both the gene expression and CNV data.

Data preparation

Sequencing based Gene Expression Profiling

We perform a TMM normalization (Robinson & Oshlack, 2010) and Voom transformation (Law, Chen, Shi, & Smyth, 2014) on the level 3 raw read counts - column raw_read_count, as described in the limma user guide (http://www.bioconductor.org/packages/release/bioc/html/limma.html).

Somatic Copy Number Variation

We match the CNV segments with a specific gene using the R package 'GenomicRanges' (Lawrence et al., 2013). We try two approaches for the mapping: (1) we use the segment lengths as they are, and (2) we assume that the genes are also influenced by the copy number variations 20.000 bases upstream and 5.000 bases downstream. As we don't observe change in the final results we decide to use the (1) straightforward approach.

Usually more fragments are matched to the same gene. In such case we calculate a mean of the log2 segment mean values for the specific gene. Ultimately, we arrive at one CNV value per gene which is used further for calculating the effect size for the gene.

Effect size calculations

We use standard linear models to compute empirical Bayes regularized t-statistics for each data type. Specifically, we employ the R limma (Smyth, 2004) package for the effect size calculations of each data type. We run the data through lmFit, contrasts.fit, and finally the eBayes functions. As it is nontrivial to decide what is the right prior to use for an analysis we first apply the eBayes function with a default prior of 0.01 to obtain raw *p*-values, for the purpose of moderating *t*-values. Then, a more precise estimate for the prior is set by convex estimation from the corresponding raw, unadjusted *p*-values. We then rerun the eBayes function and calculate the obtained log-odds ratios B for a differential signal between tumour and control samples for each data type. From the B values we calculate the posterior probabilities of a differential effect occurring for a gene.

Network construction

Gene sets

We use 186 KEGG (Kanehisa & Goto, 2000) pathways as provided in the Molecular Signatures Database (MSigDB), C2: curated gene sets, Gene sets derived from the KEGG pathway database. We reformat the gene sets file for use in LPIA as described by Pham *et al* (Pham et al., 2011).

GO terms

We download the gene_association.goa_human.gz (16-Apr-2014 07:47, 5.3M) from http://biomirror.aarnet.edu.au/biomirror/geneontology/gene-associations/ and use it to create our GO terms collection. We again reformat the file for LPIA after Pham *et al* (Pham et al., 2011): We map the genes to their Entrez names with a custom file downloaded from http://www.genenames.org/cgi-bin/ and select only biological processes with the sizes between 15 and 350.

Compilation of Reference Pathway Lists

Positive reference

We performed a literature search for KEGG (Kanehisa & Goto, 2000) pathways found to be dysregulated in clear cell Renal Cell Carcinoma (ccRCC). We found 62 pathways (Chen et al., 2013; Huang et al., 2014; The Cancer Genome Atlas Research Network, 2013; Tun et al., 2010; Zaravinos et al., 2014; ZENG et al., 2014) out of the total of 186 KEGG pathways.

Negative reference

From the set of 186 KEGG pathways we selected pathways that should theoretically not be dysregulated in ccRCC, resulting in 26 pathways out of the total of 186 KEGG pathways. We based the selection on manual literature search and are aware that it does not necessarily have to be fully correct.

Clinical annotation

Initially, we divided the matched samples in two groups based on their clinical annotation: (a) 34 samples classified as ‘remission, alive’ and (b) 21 samples showing ‘progression’, including both ‘alive’ and ‘deceased’ patients. We did not consider the 5 ‘remission, deceased’ patients because of potential confounding co-morbidities or misclassifications.

However, after multiple analyses it became apparent that due to the little number of matched patients available our sampling approach would not yield informative results and decided on using all the 55 matched, and further the 483 unmatched, samples as one group.

Setup of the tools used for evaluation

In case a default setting of a tool is changed we specify it below.

KEGGprofile (Zhao et al., 2015) is a hypergeometric test for pathway enrichment, with no evidence considered but only the set membership, where the input is a list of selected genes of interest. The genes in the list are selected as 0.95 percentile of the posterior probabilities of each gene being differentially expressed. We also tested other approaches for gene list selection, choosing all the genes that had the posterior probabilities above 0.5, 0.75, 0.9, 0.95 and 0.99. KEGGprofile with genes selected with these alternative approaches performed significantly worse in all the cases (data not shown).

RTopper (Tyekucheva et al., 2011) accepts expression levels and/or CNV estimates as input. The genes are then ranked based on the model fit measured by logit-link regression and finally a one-sided Wilcoxon rank-sum test is performed for gene set enrichment.

GSEA (Subramanian et al., 2005) is used in the 'Preranked' mode, as advised in the GSEA online manual for RNA-Seq data, where a preranked list of genes (ranking inclusive) is used as input. Such an approach simply enables both the single- and multi-track data analysis if the integrated data is used for preparing the ranked list of genes. This way it is possible to use the gene expression data alone or the previously integrated data as a list of genes ranked by evidence - posterior probabilities. Finally, as advised in the online manual, the "classic" Kolmogorov-Smirnov--like statistic is used for the gene set enrichment.

Results

Clinical data lets us relate molecular signatures to cancer remission or progression, or survival. We, therefore, considered splitting the matched samples in two groups based on their clinical annotation: (a) 34 samples classified as 'remission, alive' and (b) 21 samples showing 'progression'. We concluded, however, that the sample sizes for each group were too small to arrive at meaningful, unbiased results and therefore decided to first analyse all 55 patients as one group.

The integrative approaches we test are probabilistic but with different assumptions. As each data type needs special consideration, we first focus on gene expression and CNV, as these are expected to be most directly related but our framework allows for simple integration of any number of data types which enable calculating a size effect per gene.

We examine two questions: (i) Whether CNV data can be exploited meaningful within the tested frameworks analysing posterior probabilities that all the data sources together support a differential effect for the gene. We also tested a setup with posterior probabilities that at least one data source supports a differential effect for the gene but the results proved to be poor and/or random (analysis not shown here). (ii) Whether extending the analysed data by adding more samples can improve the results of a gene set enrichment test.

In a standard gene set enrichment experiment only gene expression is analysed. On the other hand, CNV data alone is known to be noisy. (Kuijjer et al., 2012; Louhimo et al., 2012) It has been suggested that the combined gene expression and CNV analysis should have a better predictive power (Lu et al., 2011). We therefore combined evidence from gene expression profiling with copy number somatic mutation to test if a better performance can be achieved. We compute a joint posterior probability of a gene being differentially affected using our suggested integrative approach.

To combine evidence of *all* the data sources indicating a differential effect for the gene we compute an estimate of the posterior probability of both gene expression and copy number somatic mutation of showing a differential effect:

$P_{all} = \prod_{i=1}^{N} \quad P_i$, (1)

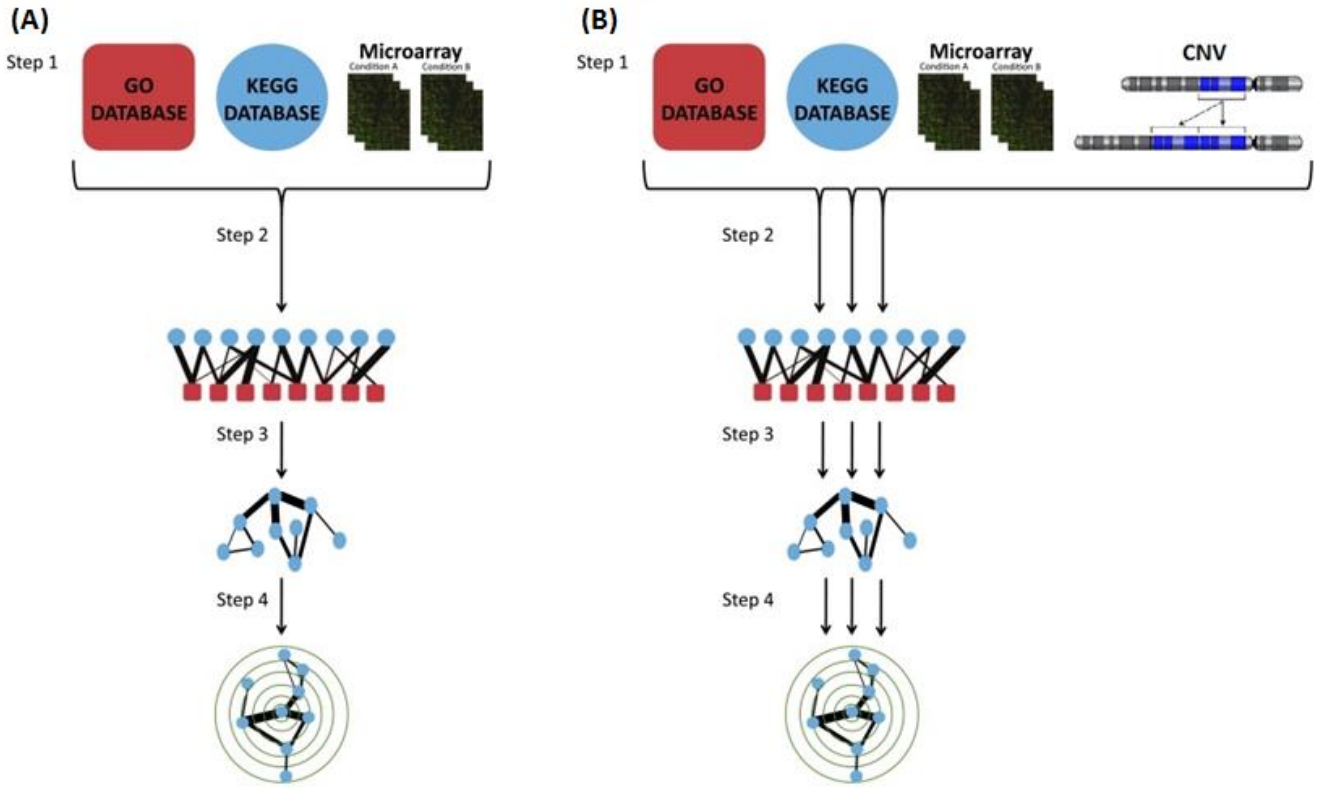
Second, we test whether at least one (*any*) of the data sources shows a differential effect for the gene, where we compute an estimate for the posterior probability of either gene expression or copy number somatic mutation showing an effect:

$P_{any} = 1 - \prod_{i=1}^{N} \quad (1 - P_i)$, (2)

Extending LPIA

We developed, and evaluated eLPIA (Figure 1) - extension of the weighted network construction in LPIA for incorporating multiple additional complementary measurement types. Moreover, we accelerate analysis by optimizing the original LPIA algorithm to support parallel execution for multi-core environments. The integrative approaches we test are probabilistic but with different assumptions. As each data type needs special consideration, we first focus on gene expression and CNV, as these are expected to be most directly related but our framework allows for simple integration of any number of data types which provide a size effect per gene.

Figure 3: (A) LPIA original design; (B) eLPIA - with additional data type and parallelized execution (figures reproduced from and modified after the original LPIA paper by Pham *et al* (Pham et al., 2011))



Adding the support for parallel execution for multi-core environments involved rewriting the software and incorporating the Perl Parallel::ForkManager. The main iteration of the centrality scores calculations is ran first and afterwards all the, *e.g.*, 1000, iterations needed for performing the bootstrap resampling for ranking the pathways can be run simultaneously, depending on the number of threads available. This version will be made available upon publication of the full-length manuscript reporting our analyses.

Dysregulated pathways detection

Now we focus on the question whether CNV data can be exploited meaningful within the LPIA framework. We test: (a) LPIA analysing the Posterior probabilities of differential gene expression and CNV alone, (b) eLPIA analysing Posterior probabilities that all the data sources together support a differential effect for the gene, and (c) eLPIA analysing the Posterior probabilities that at least one data source supports a differential effect for the gene.

It was shown that alternative preprocessing of RNA-Seq data in TCGA improves the results of the

analysis as compared to data normalized within the TCGA consortium (Rahman et al, 2015). We thus applied the state-of-the-art Voom+TMM normalization in order to yield possibly more accurate results. We TMM-normalize (Robinson & Oshlack, 2010) and Voom-transform (Law et al., 2014) the gene expression raw read counts and use standard linear models to compute empirical Bayes regularized t-statistics for each data type. An estimate for the prior is set by convex estimation from the unadjusted *p*-values. The obtained log-odds ratios (B values) then represent the differential signal between tumour and control for each data type. From the B values we calculate the posterior probabilities of a differential effect occurring for a gene. We then use these posterior probabilities as input for LPIA.

Therefore, we perform all the experiments twice - first on the original TCGA-normalized and later on the TMM/Voom-normalized data.

Analysing the original TCGA-normalzied data

(a) LPIA analysing the Posterior probabilities of differential gene expression and CNV alone

Initially, we performed the established LPIA analysis on KIRC data, separately for gene expression data, and separately for copy number somatic mutation data. The resulting pathways (Table 1) largely comply with previous studies already at this analysis step, with 73% (80% in top 10) matching known KIRC-dysregulated pathways for the gene expression analysis and 40% (40%) for the CNV analysis in the top 15 mostly dysregulated pathways.

Table 1: Top 15 KEGG pathways found by the standard LPIA approach; pathways in blue are found in literature as related to clear cell Renal Cell Carcinoma (Chen et al., 2013; Huang et al., 2014; The Cancer Genome Atlas Research Network, 2013; Tun et al., 2010; Zaravinos et al., 2014; ZENG et al., 2014).

Gene expression		
Pathway name	***FDR***	***found in literature?***
Oxidative Phosphorylation	0.00	+
Ribosome	0.00	*new*
Alzheimers Disease	0.00	+
Parkinsons Disease	0.00	+
Huntingtons Disease	0.00	*new*
Cell Adhesion Molecules Cams	0.06	+
Viral Myocarditis	0.19	+
Glycolysis Gluconeogenesis	0.21	+
Primary Immunodeficiency	0.21	+
Citrate Cycle TCA Cycle	0.29	+
Peroxisome	0.29	*new*
Complement and Coagulation Cascades	0.29	+
Systemic Lupus Erythematosus	0.29	+
Vibrio Cholerae Infection	0.29	+
ABC Transporters	0.29	*new*

Copy number variation		
Pathway name	***FDR***	***found in literature?***
Non Small Cell Lung Cancer	0.00	*new*
Melanoma	0.00	*new*
Endometrial Cancer	0.00	*new*
Aldosterone Regulated Sodium Reabsorption	0.00	*new*
Erbb Signaling Pathway	0.00	+
Glioma	0.00	*new*
Bladder Cancer	0.00	+
Renal Cell Carcinoma	0.00	+
Regulation of Actin Cytoskeleton	0.02	+
Tgf Beta Signaling Pathway	0.02	*new*
Gap Junction	0.02	*new*
Dorso Ventral Axis Formation	0.03	*new*
Chronic Myeloid Leukemia	0.03	*new*
Fc Gamma R Mediated Phagocytosis	0.04	+
Pathways in Cancer	0.06	+

It has been suggested that the combined gene expression and CNV analysis should have even better predictive power (Lu et al., 2011). We therefore extended the LPIA framework by integrating data sources in the step of computing gene weights for the construction of pathways, combining evidence from gene expression profiling with copy number somatic mutation data. We compute a joint weight per gene using our suggested integrative approaches.

(b) eLPIA analysing Posterior probabilities that *all* the data sources together support a differential effect for the gene.

First, we tested whether *all* the data sources indicated a differential effect for the gene. We compute an estimate of the posterior probability of both gene expression and copy number somatic mutation of showing a differential effect (equation 1), resulting in eLPIA yielding pathways shown in Table 2.

Table 2: Top 15 KEGG pathways found by approach (ii); pathways in blue are found in literature as related to clear cell Renal Cell Carcinoma (Chen et al., 2013; Huang et al., 2014; The Cancer Genome Atlas Research Network, 2013; Tun et al., 2010; Zaravinos et al., 2014; ZENG et al., 2014).

Gene expression and CNV		
Pathway name	***FDR***	***found in literature?***
Glycolysis Gluconeogenesis	0.00	+
Oxidative Phosphorylation	0.00	+
Cell Adhesion Molecules Cams	0.00	+
Ribosome	0.00	*new*
Parkinsons Disease	0.00	+
Citrate Cycle TCA Cycle	0.03	+
Valine Leucine and Isoleucine Degradation	0.19	+
Fructose and Mannose Metabolism	0.19	+
Pyruvate Metabolism	0.19	+
Viral Myocarditis	0.19	+
Primary Immunodeficiency	0.22	+
Huntingtons Disease	0.22	*new*
Antigen Processing and Presentation	0.23	+
Vibrio Cholerae Infection	0.41	+
Lysosome	0.43	*new*

(c) eLPIA analysing the Posterior probabilities that at least one data source supports a differential effect for the gene.

Second, we test whether at least one (*any*) of the data sources shows a differential effect for the gene. We compute an estimate for the posterior probability of either gene expression or copy number somatic mutation showing an effect (assuming independence, in first approximation, *see* equation 2). Among the top ranked pathways we find many which are not currently annotated in the literature (Chen et al., 2013; Huang et al., 2014; The Cancer Genome Atlas Research Network, 2013; Tun et al., 2010; Zaravinos et al., 2014; ZENG et al., 2014) (Table 3).

Table 3: Top 15 KEGG pathways found by approach (iii); pathways in blue are found in literature as related to clear cell Renal Cell Carcinoma (Chen et al., 2013; Huang et al., 2014; The Cancer Genome Atlas Research Network, 2013; Tun et al., 2010; Zaravinos et al., 2014; ZENG et al., 2014).

Gene expression and CNV		
Pathway name	***FDR***	***found in literature?***
Non Small Cell Lung Cancer	0.00	*new*
Chronic Myeloid Leukemia	0.00	*new*
Renal Cell Carcinoma	0.06	+
Glioma	0.06	*new*
Bladder Cancer	0.06	+
Fc Gamma R Mediated Phagocytosis	0.06	+
Pancreatic Cancer	0.08	+
Erbb Signaling Pathway	0.09	+
Insulin Signaling Pathway	0.17	*new*
Melanoma	0.22	*new*
Acute Myeloid Leukemia	0.25	*new*
Regulation of Actin Cytoskeleton	0.26	+
Endocytosis	0.29	*new*
Small Cell Lung Cancer	0.29	+
Colorectal Cancer	0.32	+
Cardiac Muscle Contraction	0.32	+
Pathways In Cancer	0.32	+

The results returned by this approach seem very far from the reference literature findings, returning many pathways unlikely to be dysregulated in KIRC.

Analysing the TMM/Voom-normalzied data

(a) LPIA analysing the Posterior probabilities of differential gene expression and CNV alone

Again we perform the established LPIA analysis on KIRC data, separately for gene expression data, and separately for copy number somatic mutation data. The resulting pathways (Table 4) largely comply with previous studies already at this analysis step, with 60% matching known KIRC-dysregulated pathways for the gene expression analysis and 40% for the CNV analysis in the top 15 mostly dysregulated pathways.

Table 4: Top 15 KEGG pathways found by the standard LPIA approach; pathways in blue are found in literature as related to clear cell Renal Cell Carcinoma (Chen et al., 2013; Huang et al., 2014; The Cancer Genome Atlas Research Network, 2013; Tun et al., 2010; Zaravinos et al., 2014; ZENG et al., 2014).

Gene expression			*Copy number variation*		
Pathway name	***FDR***	***found in literature?***	***Pathway name***	***FDR***	***found in literature?***
Primary Immunodeficiency	0.87	+	Non Small Cell Lung Cancer	0.00	*new*
Propanoate Metabolism	0.87	+	Melanoma	0.00	*new*
Axon Guidance	0.87	*new*	Endometrial Cancer	0.00	*new*
Proximal Tubule Bicarbonate Reclamation	0.87	*new*	Aldosterone Regulated Sodium Reabsorption	0.00	*new*
Cell Adhesion Molecules Cams	0.87	+	Erbb Signaling Pathway	0.00	+
Viral Myocarditis	0.87	+	Glioma	0.00	*new*
Endocytosis	0.87	*new*	Bladder Cancer	0.00	+
ABC Transporters	0.87	*new*	Renal Cell Carcinoma	0.00	+
Phenylalanine Metabolism	0.87	*new*	Regulation of Actin Cytoskeleton	0.02	+
Pathways In Cancer	0.87	+	Tgf Beta Signaling Pathway	0.02	*new*
Complement And Coagulation Cascades	0.87	+	Gap Junction	0.02	*new*
Glycine Serine And Threonine Metabolism	0.87	+	Dorso Ventral Axis Formation	0.03	*new*
Neuroactive Ligand Receptor Interaction	0.87	*new*	Chronic Myeloid Leukemia	0.03	*new*
Ppar Signaling Pathway	0.87	+	Fc Gamma R Mediated Phagocytosis	0.04	+
Pancreatic Cancer	0.87	+	Pathways in Cancer	0.06	+

(ii) eLPIA analysing Posterior probabilities that all the data sources together support a differential effect for the gene.

First, we tested whether *all* the data sources indicated a differential effect for the gene. We compute an estimate of the posterior probability of both gene expression and copy number somatic mutation of showing a differential effect (equation 1), resulting in eLPIA yielding pathways shown in Table 5, with 67% matching known KIRC-dysregulated pathways.

Table 5: Top 15 KEGG pathways found by approach (ii); pathways in blue are found in literature as related to clear cell Renal Cell Carcinoma.{Chen et al., 2013(Huang et al., 2014)The Cancer Genome Atlas Research Network, 2013(Tun et al., 2010(Zaravinos et al., 2014)ZENG et al., 2014)

Gene expression and CNV		
Pathway name	***FDR***	***found in literature?***
Ribosome	0.77	*new*
Alzheimer's Disease	0.77	+
Glycolysis Gluconeogenesis	0.77	+
Cell Adhesion Molecules Cams	0.77	+
Cell Cycle	0.77	+
Base Excision Repair	0.77	*new*
Huntington's Disease	0.77	*new*
Parkinson's Disease	0.77	+
Oxidative Phosphorylation	0.77	+
Ecm Receptor Interaction	0.77	+
Bladder Cancer	0.77	+
Calcium Signaling Pathway	0.77	*new*
Propanoate Metabolism	0.77	+
Proximal Tubule Bicarbonate Reclamation	0.77	*new*
Pyruvate Metabolism	0.77	+

(iii) eLPIA analysing the Posterior probabilities that at least one data source supports a differential effect for the gene.

Second, we test whether at least one (*any*) of the data sources shows a differential effect for the gene. We compute an estimate for the posterior probability of either gene expression or copy number somatic

mutation showing an effect (assuming independence, in first approximation, *see* equation 2).

As in the previous TCGA-normalized data results among the top ranked pathways we find many pathways which are not currently annotated in the literature (Chen et al., 2013; Huang et al., 2014; The Cancer Genome Atlas Research Network, 2013; Tun et al., 2010; Zaravinos et al., 2014; ZENG et al., 2014) (Table 3), with only 33% matching known KIRC-dysregulated pathways.

Table 3: Top 15 KEGG pathways found by approach (iii); pathways in blue are found in literature as related to clear cell renal cell carcinoma (Chen et al., 2013; Huang et al., 2014; The Cancer Genome Atlas Research Network, 2013; Tun et al., 2010; Zaravinos et al., 2014; ZENG et al., 2014).

Gene expression or CNV Pathway name	***FDR***	***found in literature?***
Non Small Cell Lung Cancer	0.00	*new*
MTOR Signaling Pathway	0.00	+
Chronic Myeloid Leukemia	0.00	*new*
Glioma	0.00	*new*
Prion Diseases	0.00	*new*
Pancreatic Cancer	0.00	+
Bladder Cancer	0.00	+
Progesterone Mediated Oocyte Maturation	0.00	*new*
Long Term Depression	0.02	*new*
Erbb Signaling Pathway	0.03	+
Fc Gamma R Mediated Phagocytosis	0.03	+
Vegf Signaling Pathway	0.05	*new*
Insulin Signaling Pathway	0.07	*new*
Dorso Ventral Axis Formation	0.08	*new*
Gap Junction	0.09	*new*

The results returned by this approach are again very far from the reference literature findings, returning additionally many pathways unlikely to be dysregulated in KIRC.

Assessing performance

We performed evaluation of the eLPIA algorithm, comparing its performance to three established gene set enrichment analysis tools, each using a different approach for the enrichment calculations and/or integration method: KEGGprofile (Zhao, Guo, & Shyr, 2015), RTopper (Tyekucheva et al., 2011) and GSEA (Subramanian et al., 2005). We used the same tools for both the single- (gene expression) and multi-track (gene expression + CNV) data analysis.

We randomize the data by subsampling, where we randomly sample 36 patients from the group of 55 matched normal-tumour samples 21 times (not more, due to computational cost). We use two complementary subsampling approaches, where (1) compute the Positive Predictive Value against a positive reference list of pathways, subsampling normal paired with shuffled normal samples, and (2) we compute the Positive Predictive Value (PPV) against a negative reference list of pathways, subsampling the matched normal-tumour samples. The PPV is calculated according to formula (3) and always for the top 30 pathways reported by a tool to be dysregulated.

$$PPV = \frac{true\ positives}{true\ positives + false\ positives}, \quad (3)$$

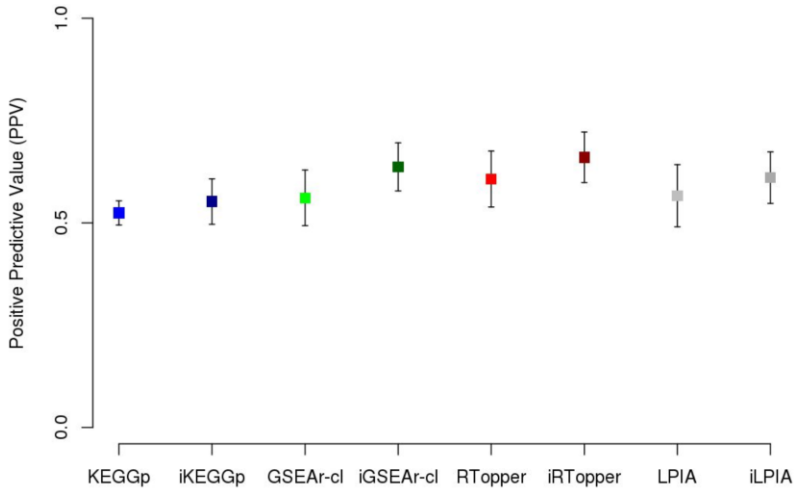
where *true positives* stands for the number of pathways that are correctly found to be dysregulated in KIRC according to literature and *false positives* is the number of pathways (3.1) from the positive reference set that are still found in the data even after the data set is made meaningless by shuffling the data, or (3.2) that are are found to be dysregulated but should not according to a negative reference

list.

(1) Normal *vs* shuffled normal

With this approach we are testing if the method can be 'cheated' into finding positive reference pathways even if the data is meaningless - after subsampling normal paired with shuffled normal samples. If a tool finds pathways from the positive reference in the normal vs shuffled normal data it is assumed to be incorrect. We perform a gene set enrichment analysis with each of the tools, including eLPIA and repeat this 21 times. With the results we re-calculate the PPV according to equation 3.1.

Figure 4: PPVs for *Normal vs shuffled normal* approach; 'i' in front of a tool's name indicates that it shows the results of the integrated data run, otherwise the results are shown for the gene expression alone

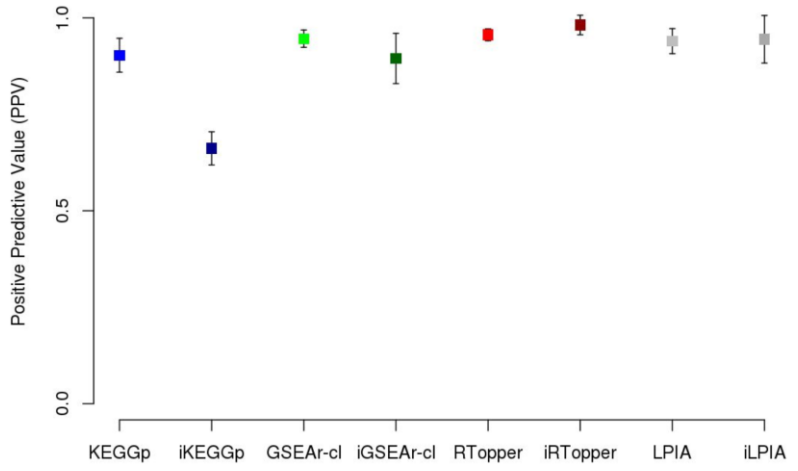


There seems to be a visual trend of data integration increasing the PPV. None of the results are, however, significant. Tools seem to be performing similarly well and it is impossible to find the superior algorithm. Therefore, we tested another approach.

(2) Matched normal-tumour

We now employ a negative reference set. With this approach we are testing if the method finds truly wrong pathways - negative reference pathways, in meaningful data. Here the PPV is calculated according to equation 3.2.

Figure 2: PPVs for *Matched normal-tumour* approach; 'i' in front of a tool's name indicates that it shows the results of the integrated data run, otherwise the results are shown for the gene expression alone



Apart from KEGGprofile, all the tools again seem to be performing similarly well and it is impossible to find a significantly superior algorithm.

We performed extensive benchmarking using data set extended by unmatched samples which can be accessed in the Supplementary file.

Extended benchmarking

After our two benchmarking approaches with the matched data resulted in almost no differences in all of the methods performance we now test if adding more data - unmatched samples, will show a difference in the performance. We sample 101 patients from the groups of 483 tumour and 55 unmatched normal samples. For sake of computational cost we leave eLPIA method out of the further analyses, assuming that only if there is any space left for performance improvement does it make sense to actually compare our novel method.

We run each of the gene set enrichment tools 101 times. Initially, with the results we again calculated the PPV on the top 30 pathways. This way the calculations were still not yielding any more useful benchmarking results. Therefore, we decided to employ another measure of performance - sensitivity, and not use the 30 top pathways for calculations but rather base the pathway selection on an eFDR threshold set for each method individually to the same level.

We thus performed extensive performance evaluation of three established gene set enrichment analysis tools now employing two measures of performance: (i) sensitivity

$$sensitivity = \frac{true\ positives}{true\ positives + false\ negatives}, \quad (4)$$

where *true positives* stands for the number of pathways that are correctly found to be dysregulated in KIRC according to literature and the *false negatives* is the number of pathways from the positive reference list that are not found by a method to be dysregulated.

And (ii) Positive Predictive Value (PPV) (3.2), where *false positives* is the number of pathways that are found to be dysregulated but should not according to a negative reference list.

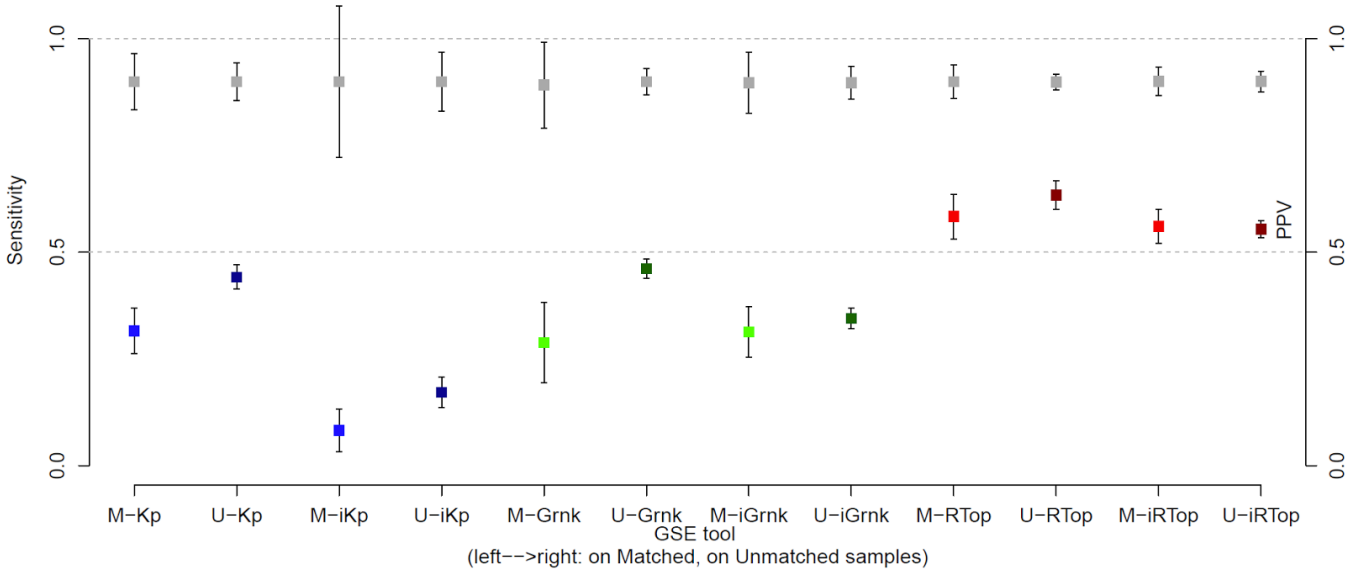
For each method each of the sensitivities and PPV values are calculated on results returned at a 10% eFDR level. To make the results between methods comparable the *q*-value threshold for achieving the 10% eFDR is adjusted for each method and each approach - single- and multi-track data runs, separately.

Comparing runs with matched to runs with unmatched samples

We examine how adding (i) an additional data type and (ii) more data - unmatched samples, changes the performance of a tool.

We randomize the data by sub-sampling. For the matched samples runs we randomly sample 36 patients from the group of 55 matched normal-tumour samples. For the dataset extended with the addition of unmatched samples we take all the 483 tumour patients and sample 483 times from the 55 unmatched normal samples. We perform a gene set enrichment analysis with each of the tools and repeat this 101 times. With the results we calculate the sensitivity and PPV.

Figure 1: Sensitivity (in colour) and PPV (in grey) at 10% eFDR, for M-: *Matched* data results compared with U-: *Unmatched* data; “i” in front of a tool’s name indicates that it shows the results of the integrated data run, otherwise the results are shown for the gene expression alone; GSE tools: Kp = KEGGprofile, Grnk = GSEA-Preranked, RTop = RTopper



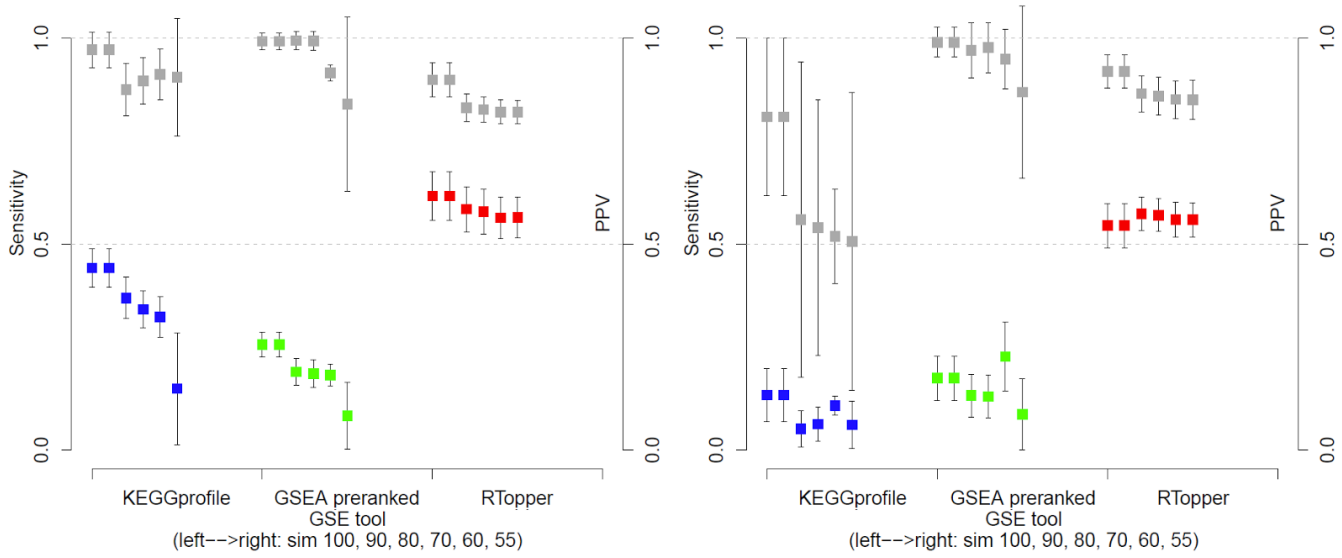
We observe that the addition of *more data* (here a 10-fold increase in sample size) is generally beneficial and performance of a specific tool for a specific setup increases with increased number of samples. This is not apparent when considering the PPV of a method but becomes evident when sensitivity is examined. On the other hand, integrating additional data type - CNV, seems to mostly, apart from KEGGprofile, not change a method’s performance when matched samples are considered. In case of using more (unmatched) samples the decrease of performance with integrated CNV is clearly visible.

We further focus only on the extended data set, using the unmatched samples.

Signal reduction in runs on extended data

We next investigated the relative performance of different methods in the detection of more subtle signals. Here, for the sake of computing power, we sample 101 patients from the total of 483 tumour samples and sample 101 times from the 55 unmatched normal samples. We perform a gene set enrichment analysis with each of the tools and repeat this 101 times. With the results we calculate the sensitivity and PPV. The tumour samples signals were down-mixed by admixture of normals and vice versa. *E.g.*, 100% means that we used "pure" tumour signal for the run - no down-mixing, whereas 80% means that we mixed 80% of tumour sample signal with 20% of normal sample signal for the “tumour” sample and 20% of tumour sample signal with 80% of normal sample signal for the “normal” sample.

Figure 2: Sensitivity (in colour) and PPV (in grey) at 10% eFDR, for *Unmatched* data results; left panel: analysis of expression data alone, right panel: analysis of integrated data; “sim 100” stands for 100% tumour sample (no down-mixing), “sim 80” stands for 80% down-mixing, *etc.*



A diluted signal seems to influence performance of the methods that are based on unweighted lists of genes. Interestingly, for the most sophisticated method - RTopper, both the sensitivity and PPV decrease only slightly and mostly insignificantly with diluted signal.

Discussion

Pathways found to be dysregulated by LPIA

As shown the results returned by the second integrative approach, where we tested whether at least one (*any*) of the data sources (gene expression or CNV) shows a differential effect for the gene, were so far from the reference literature findings, returning additionally many pathways unlikely to be dysregulated in KIRC, that here we only focus on evaluating the pathways returned by our first proposed integrative approach - where we test whether *all* the data sources indicated a differential effect for the gene.

...in the TCGA-normalized data

Individual results from gene expression alone match pathways previously reported in the literature well (73% of the top 15), and this is an independent validation of LPIA as an effective method for the inference of pathways. We find four new pathways: The ‘Ribosome’ pathway unspecifically reflects cell protein production activity. More interestingly, it is proven that the mTOR signaling pathway is active in ccRCC (Robb et al., 2007) and its inhibition may have disease-modifying effects in both ccRCC (Battelli & Cho, 2011) and neurodegenerative disorders (Wong, 2013) linking the 'Huntingtons Disease' pathway with KIRC. The ‘Peroxisome’ novel specific pathway may be of direct interest, considering that it is well known that lipid biosynthesis is largely dysregulated in ccRCC (Drabkin & Gemmill, 2012) and peroxisomes play a key role in lipid metabolism (del Río, 2013). ABC transporters are widely expressed in cancer cells and are a known to cause resistance to chemoterapeutic drugs, and are thus a suggested target for cancer therapy (Szakács et al., 2006), making the ‘ABC Transporters’ pathway a plausible candidate for investigation in kidney cancer.

We expected that the interpretation of the joint analyses would be more challenging but it in fact it

yielded results in even a higher concordance with the reference. Requiring a signal from both CNV and gene expression yields a slightly better concordance with pathways earlier reported to be associated with KIRC (80% of the top 15 pathways). Two of the novel dysregulated pathways identified as of potential interest are the ‘Ribosome’ and 'Huntingtons Disease' pathways similar to the findings by the gene expression profiling alone. However, the third pathway identified by this joint analysis - ‘Lysosome’, is different. Lysosomes, known for being the waste disposal system of a cell, are believed to be both pro- and anti-oncogenic (Kirkegaard & Jäättelä, 2009) with an important function of facilitating cell death even in cancer cells where the classical apoptosis pathway becomes dysfunctional (Jäättelä, 2004) thus becoming of increased interest in oncology.

In summary, the combined analysis seems to partially confirm the findings by the gene expression profiling analysis alone and also yield additional interesting outcome.

Results from CNV alone recover a smaller set of pathways reported to be involved (40% of the top 15). This can be understood as a result of two effects: (1) most of the pathways currently reported in the literature have been derived from gene expression analyses, and (2) much fewer genes (54-77% fewer) show a significant differential signal in the CNV data. Interestingly, some of the pathways found to be dysregulated in the KIRC CNV data suggest a less specific connection to KIRC, yielding pathways linking to more, seemingly unrelated, cancer types, like melanoma, glioma or lung cancer.

Requiring a signal from CNV or gene expression data recovers a much smaller number of pathways earlier reported (59% of the top 15). Notably, while the number of pathways that have not been identified before is close to an analysis of CNV alone, the pathways differ but we again see pathways related to other types of cancers, with glioma, melanoma or leukemia. The question arises whether these are false positives or new discoveries of biological relevance.

...in the TMM/Voom-normalized data

Individual results from gene expression alone match pathways previously reported in the literature well (60% of the top 15), and this is an independent validation of LPIA as an effective method for the inference of pathways. We find six new pathways: The ‘Proximal tubule bicarbonate reclamation’ seems a plausible candidate as it is involved in maintaining the right pH of the tubule lumen. It is known that acidic environment is important to cancer progression because it protects cancer cells from immune system (Pinthus, 2011; Kanehisa & Goto, 2000). ‘Endocytosis’ pathways is long known to be derailed in cancer cells, as an effect of multitude of oncogenic alterations (Mosesson et al., 2008), leading to functional dysregulation of multiple receptors thus enabling cancer to grow (Mellman & Yarden, 2013). ABC transporters are widely expressed in cancer cells and are known to cause resistance to chemotherapeutic drugs, and are thus a suggested target for cancer therapy (Szakács et al., 2006), making the ‘ABC Transporters’ pathway a plausible candidate for investigation in kidney cancer. It is well known that restriction of amino acids like phenylalanine can inhibit growth and metastasis of cancer (Y.-M. Fu et al., 1999). Furthermore, death of, *e.g.*, prostate cancer cells is closely related to changes in glucose metabolism, which can be influenced by the aforementioned amino acid restriction (Y.-M. Fu et al., 2010). Thus the ‘Phenylalanine Metabolism’ pathway might in fact be worth closer investigation. ‘Axon Guidance’ and ‘Neuroactive Ligand Receptor Interaction’ pathways are, however, unexpected. On the other hand, pathways of neurodegenerative diseases, like Alzheimer's and Parkinson's, have been previously reported to be dysreguated in ccRCC (Huang et al., 2014). The neural degeneration of kidney cancer patients remains thus to be further examined.

We expected that the interpretation of the joint analyses would be more challenging but in fact it yielded results even in a higher concordance with the reference. Requiring a signal from both CNV and gene

expression yields a slightly better concordance with pathways earlier reported to be associated with KIRC (67% of the top 15 pathways). One of the novel dysregulated pathways identified as of potential interest is ‘Proximal tubule bicarbonate reclamation’ pathway similar to the findings by the gene expression profiling alone. However, the four additional pathways identified by this joint analysis are different. The ‘Ribosome’ pathway unspecifically reflects cell protein production activity. 'Base Excision Repair' is essential for the cell to remain healthy, repairing the damaged DNA, removing damaged nucleotides able to cause mutations. Interestingly the defects in the base excision repair system have been associated with both neurological disorders and cancer (Wallace et al., 2012). Furthermore, it has been proven that the mTOR signaling pathway is active in ccRCC (Robb et al., 2007) and its inhibition may have disease-modifying effects in both ccRCC (Battelli & Cho, 2011) and neurodegenerative disorders (Wong, 2013) linking the 'Huntington’s Disease' pathway with KIRC. We also found ‘Calcium Signaling Pathway’ to be dysregulated in ccRCC. Dysregulation of the intracellular Ca^{2+} signalling, being essential in modulating diverse cellular functions, has been suggested as a driving factor for malignant phenotypes emergence (Chen et al., 2013b). Serum calcium is also used as one of the prognostic risk factors for categorizing metastatic RCC patients into risk groups (Motzer et al., 1999). The original LPIA algorithm was designed to be very specific and therefore the FDR is very high in the ranked pathways. Interestingly, together with the reduction of the number of newly found pathways after the integration also the FDR drops by 10% showing that also the certainty in the results is increased.

In summary, the combined analysis seems to partially confirm the findings by the gene expression profiling analysis alone and also yield additional interesting outcomes. These are interesting candidates for further investigation and we hope our findings will extend our understanding of the mechanisms of cancer and improve diagnosis.

Results from CNV alone recover a smaller set of pathways reported to be involved (40% of the top 15). This can be understood as a result of two effects: (1) most of the pathways currently reported in the literature have been derived from gene expression analyses, and (2) much fewer genes (54-77% fewer) show a significant differential signal in the CNV data. Interestingly, some of the pathways found to be dysregulated in the KIRC CNV data suggest a less specific connection to KIRC, yielding pathways linking to more, seemingly unrelated, cancer types, like melanoma, glioma or lung cancer.

Requiring a signal from CNV or gene expression data recovers a much smaller number of pathways earlier reported (33% of the top 15). Notably, while the number of pathways that have not been identified before is close to an analysis of CNV alone, the pathways differ but we again see pathways related to other diseases, *e.g.*, glioma, leukemia and prion disease. The question arises whether these are false positives or new discoveries of biological relevance.

Summary

Our initial evaluation of LPIA has proven that using a positive and negative pathway reference enables the use of an evaluation metric - PPV, which might be helpful in assessing performance of the gene set analysis tools but alone is unable to show a method’s superiority. We were able to confirm that on the small data set investigated in this paper all the compared tools perform similarly well. We confirmed that LPIA / eLPIA is in fact as reliable as other established methods. The inevident superiority of eLPIA might have been expected as the algorithm was designed to detect very small dysregulation in data, whereas cancer samples tend to be highly differentiated from normal samples and the high dysregulaton can be spotted by other tools. In fact, the reliability of LPIA was already shown in the original publication (Pham et al., 2011). Here we were able to confirm that the extended framework is still as reliable. Benchmarking our method has proven, however, to be enormously difficult and this challenge is in line with earlier observations (Alexeyenko et al., 2012; Glaab et al., 2012; Verbeke et al.,

2015). Even after extending the data set in order to give better power in assessing tool performance / improvements our extensive analyses have shown that on this particular data set showing a superiority of a novel method is impossible as there is simply no space left for improvement. Thus identification of more sensitive benchmarks or more powerful data sets remains an area of active research. Other advantages that our novel framework presents are independence of any artificially thresholded set of genes of interest - as needed in hypergeometric tests; ease of integrating any additional differential data type - which is the foundation of our framework; and insensitivity to ties in the gene list, which is an evident issue when running methods that take a ranked list of genes as input. Notably, the parallelization of the LPIA algorithm was essential in performing the evaluation as the subsampling involved multiple re-runs of the analysis.

Moreover, we have shown that the choice of normalization method for the data analysed leads to a method yielding different results. This was not the focus of our experiment but should be considered closely when performing bioinformatics analyses.

Benchmarking gene set enrichment methods

Our benchmarking approach has further shown that using a positive and negative pathway reference enables the use of an additional evaluation metric - sensitivity, together with PPV, which help in assessing performance of the gene set analysis tools.

We have compiled a reference set of pathways expected to be involved in clear cell Renal Cell Carcinoma (ccRCC) from the literature and proposed a systematic test of robustness, reproducibility, sensitivity, and specificity for benchmarking gene set enrichment approaches. We could show that cohort size is a strong driver of performance in general. In the data set examined, cohort size was the strongest determinant of performance. In-silico mixtures of real samples that attenuated cancer-relevant signals showed that method choice became crucially important for weaker signals. From the most attenuated gene expression signal alone, RTopper already successfully identified the reference pathways with a sensitivity of ~50%, maintaining this rate when CNV data was added. This suggests that the reliable performance of a modern algorithm for pathway enrichment testing is robust but advantages from integrating different molecular profiling data types are not apparent on this dataset.

Conclusions

It has been shown that the integration of multiple data types alone (Tyekucheva et al., 2011) or network-based analysis alone (Pham et al., 2011) can improve the power of pathway enrichment analyses. Most recent analyses report that a combination of data integration and networks could further improve survival rate prediction (Wang et al., 2014). To date, such an approach has not yet been systematically investigated in the context of pathway enrichment analysis. In a recent publication Verbeke *et al* (Verbeke et al., 2015) propose a method for ranking pathways through network-based data integration. While the authors discuss that their method yields results in agreement with a previous approach, no thorough benchmark is applied. This reflects well acknowledged challenges of benchmarking in the absence of a ‘ground truth’ in the field (Alexeyenko et al., 2012; Glaab et al., 2012; Verbeke et al., 2015). To address this open need, I have compiled manually curated pathway lists to serve as positive and negative reference sets for advanced analysis methods applied to an extensive multi-track data collection on KIRC of the TCGA/ICGC repositories of cancer profiling experiments. We could then apply these to validate our novel implementation of modern pathway enrichment analysis employing network-based data integration. Control-treatment samples were first analyzed with regard to differential effect. The result was then directly applied to weight the edges of the network of pathways. This way we could avoid using thresholds or binarization of datasets in the process of

network creation, making the most use of the information encoded in the measurements. Notably, on the compiled benchmark set, the new method performs at least as well as methods only integrating multiple -omics data types or approaches to integration only employing network-based structures. This suggests that pathway enrichment analysis is more robust and less sensitive to inputs and method choice than other analysis tasks like survival prediction. The compilation of more sensitive benchmark frameworks for assessing more subtle advances in pathway enrichment performance of course remains an area of active research.

Acknowledgements

Maciej M Kańduła is a Marshall Plan Scholar, and the performed research was supported by the Austrian Marshall Plan Foundation.

Owing to the funding provided by Austrian Marshall Plan Foundation I was able to participate in scientific exchange in Boston, US in the time period of 10. April - 16. November 2016. Not only could I develop my collaboration with Prof. Eric Kolaczyk and work on the research project outlined in this report, but I was also able to extend my collaborative network, get involved in new exciting joint research venues, and improve my statistical/mathematical skills by participating in various workshops, lectures and seminars. The exchange visit also further developed my interpersonal skills, allowing me to meet students and researchers in Boston from all over the world. I’m thankful to the Austrian Marshall Plan Foundation for this opportunity.

References

Alexeyenko, A., Lee, W., Pernemalm, M., Guegan, J., Dessen, P., Lazar, V., … Pawitan, Y. (2012). Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*, *13*, 226. http://doi.org/10.1186/1471-2105-13-226

Battelli, C., & Cho, D. C. (2011). mTOR inhibitors in renal cell carcinoma. *Therapy*, *8*(4), 359–367. http://doi.org/10.2217/thy.11.32

Barabási, A.-L. (2007). Network Medicine — From Obesity to the "Diseasome." New England Journal of Medicine, 357(4), 404–407. http://doi.org/10.1056/NEJMe078114

Benton, D. (1996). Bioinformatics--principles and potential of a new multidisciplinary tool. *Trends in Biotechnology*, *14*(8), 261–272. http://doi.org/10.1016/0167-7799(96)10037-8

Chen, J., Zhang, D., Zhang, W., Tang, Y., Yan, W., Guo, L., & Shen, B. (2013). Clear cell renal cell carcinoma associated microRNA expression signatures identified by an integrated bioinformatics analysis. *Journal of Translational Medicine*, *11*, 169. http://doi.org/10.1186/1479-5876-11-169

Chen, Y.-F., Chen, Y.-T., Chiu, W.-T., & Shen, M.-R. (2013b). Remodeling of calcium signaling in tumor progression. *Journal of Biomedical Science*, *20*, 23. http://doi.org/10.1186/1423-0127-20-23

Del Río, L. A. (Ed.). (2013). *Peroxisomes and their Key Role in Cellular Signaling and Metabolism* (Vol. 69). Dordrecht: Springer Netherlands. Retrieved from http://link.springer.com/10.1007/978-94-007-6889-5

Drabkin, H. A., & Gemmill, R. M. (2012). Cholesterol and the development of clear-cell renal carcinoma. *Current Opinion in Pharmacology*, *12*(6), 742–750. http://doi.org/10.1016/j.coph.2012.08.002

Fu, Y.-M., Lin, H., Liu, X., Fang, W., & Meadows, G. G. (2010). Cell death of prostate cancer cells by specific amino acid restriction depends on alterations of glucose metabolism. *Journal of Cellular Physiology*, *224*(2), 491–500. http://doi.org/10.1002/jcp.22148

Fu, Y. M., Yu, Z. X., Pelayo, B. A., Ferrans, V. J., & Meadows, G. G. (1999). Focal adhesion kinase-dependent apoptosis of melanoma induced by tyrosine and phenylalanine deficiency. *Cancer Research*, *59*(3), 758–765.

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., & Valencia, A. (2012). EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, *28*(18), i451–i457. http://doi.org/10.1093/bioinformatics/bts389

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., & Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models—A review. *Biosystems*, *96*(1), 86–103. http://doi.org/10.1016/j.biosystems.2008.12.004

Hothorn, T., & Leisch, F. (2011). Case studies in reproducibility. *Briefings in Bioinformatics*, bbq084. http://doi.org/10.1093/bib/bbq084

Huang, H., Tang, Y., He, W., Huang, Q., Zhong, J., & Yang, Z. (2014). Key pathways and genes controlling the development and progression of clear cell renal cell carcinoma (ccRCC) based on gene set enrichment analysis. *International Urology and Nephrology*, *46*(3), 539–553. http://doi.org/10.1007/s11255-013-0511-2

Ioannidis, J. P. A. (2014). How to Make More Published Research True. *PLOS Med*, *11*(10), e1001747. http://doi.org/10.1371/journal.pmed.1001747

Jäättelä, M. (2004). Multiple cell death pathways as regulators of tumour initiation and progression. *Oncogene*, *23*(16), 2746–2756. http://doi.org/10.1038/sj.onc.1207513

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30.

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology*, *8*(2). http://doi.org/10.1371/journal.pcbi.1002375

Kirkegaard, T., & Jäättelä, M. (2009). Lysosomal involvement in cell death and cancer. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1793*(4), 746–754. http://doi.org/10.1016/j.bbamcr.2008.09.008

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. http://doi.org/10.1093/bioinformatics/bts480

Kuijjer, M. L., Rydbeck, H., Kresse, S. H., Buddingh, E. P., Lid, A. B., Roelofs, H., … Cleton-Jansen, A.-M. (2012). Identification of osteosarcoma driver genes by integrative analysis of copy number and gene expression data. Genes, Chromosomes & Cancer, 51(7), 696–706. http://doi.org/10.1002/gcc.21956

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29. http://doi.org/10.1186/gb-2014-15-2-r29

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., … Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol*, *9*(8), e1003118. http://doi.org/10.1371/journal.pcbi.1003118

Louhimo, R., Lepikhova, T., Monni, O., & Hautaniemi, S. (2012). Comparative analysis of algorithms for integration of copy number and expression data. *Nature Methods*, *9*(4), 351–355. http://doi.org/10.1038/nmeth.1893

Lu, T.-P., Lai, L.-C., Tsai, M.-H., Chen, P.-C., Hsu, C.-P., Lee, J.-M., … Chuang, E. Y. (2011). Integrated analyses of copy number variations and gene expression in lung adenocarcinoma. *PloS One*, *6*(9), e24829. http://doi.org/10.1371/journal.pone.0024829

Mellman, I., & Yarden, Y. (2013). Endocytosis and Cancer. *Cold Spring Harbor Perspectives in Biology*, *5*(12), a016949. http://doi.org/10.1101/cshperspect.a016949

Mosesson, Y., Mills, G. B., & Yarden, Y. (2008). Derailed endocytosis: an emerging feature of cancer. Nature Reviews Cancer, 8(11), 835–850. http://doi.org/10.1038/nrc2521

Motzer, R. J., Mazumdar, M., Bacik, J., Berg, W., Amsterdam, A., & Ferrara, J. (1999). Survival and prognostic stratification of 670 patients with advanced renal cell carcinoma. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *17*(8), 2530–2540.

Mushegian, A. (2011). Grand Challenges in Bioinformatics and Computational Biology. *Frontiers in Genetics*, *2*. http://doi.org/10.3389/fgene.2011.00060

Nariai, N., Tamada, Y., Imoto, S., & Miyano, S. (2005). Estimating gene regulatory networks and protein–protein interactions of Saccharomyces cerevisiae from multiple genome-wide data. *Bioinformatics*, *21*(suppl 2), ii206–ii212. http://doi.org/10.1093/bioinformatics/bti1133

Papin, J. A., Reed, J. L., & Palsson, B. O. (2004). Hierarchical thinking in network biology: the unbiased modularization of biochemical

networks. *Trends in Biochemical Sciences*, *29*(12), 641–647. http://doi.org/10.1016/j.tibs.2004.10.001

Pinthus, J. H. (2011). ADT and the metabolic syndrome: no good deed goes unpunished. *Canadian Urological Association Journal*, *5*(1), 33. http://doi.org/10.5489/cuaj.11017

Rahman, M., Jackson, L. K., Johnson, W. E., Li, D. Y., Bild, A. H., & Piccolo, S. R. (2015). Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. Bioinformatics, btv377. http://doi.org/10.1093/bioinformatics/btv377

Robb, V. A., Karbowniczek, M., Klein-Szanto, A. J., & Henske, E. P. (2007). Activation of the mTOR signaling pathway in renal clear cell carcinoma. *The Journal of Urology*, *177*(1), 346–352. http://doi.org/10.1016/j.juro.2006.08.076

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. http://doi.org/10.1186/gb-2010-11-3-r25

Romano, P. (2008). Automation of in-silico data analysis processes through workflow management systems. *Briefings in Bioinformatics*, *9*(1), 57–68.

Searls, D. B. (2005). Data integration: challenges for drug discovery. *Nature Reviews Drug Discovery*, *4*(1), 45–58. http://doi.org/10.1038/nrd1608

Silverman, E. K., & Loscalzo, J. (2012). Network Medicine Approaches to the Genetics of Complex Diseases. *Discovery Medicine*, *14*(75), 143–152.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*, Article3. http://doi.org/10.2202/1544-6115.1027

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. http://doi.org/10.1073/pnas.0506580102

Szakács, G., Paterson, J. K., Ludwig, J. A., Booth-Genthe, C., & Gottesman, M. M. (2006). Targeting multidrug resistance in cancer. *Nature Reviews Drug Discovery*, *5*(3), 219–234. http://doi.org/10.1038/nrd1984

The Cancer Genome Atlas Research Network. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature, 499(7456), 43–49. http://doi.org/10.1038/nature12222

Tun, H. W., Marlow, L. A., von Roemeling, C. A., Cooper, S. J., Kreinest, P., Wu, K., … Copland, J. A. (2010). Pathway Signature and Cellular Differentiation in Clear Cell Renal Cell Carcinoma. PLoS ONE, 5(5), e10696. http://doi.org/10.1371/journal.pone.0010696

Tyekucheva, S., Marchionni, L., Karchin, R., & Parmigiani, G. (2011). Integrating diverse genomic data using gene sets. *Genome Biology*, *12*(10), R105. http://doi.org/10.1186/gb-2011-12-10-r105

Verbeke, L. P. C., Van den Eynden, J., Fierro, A. C., Demeester, P., Fostier, J., & Marchal, K. (2015). Pathway Relevance Ranking for Tumor Samples through Network-Based Data Integration. *PLoS ONE*, *10*(7), e0133503. http://doi.org/10.1371/journal.pone.0133503

Wallace, S. S., Murphy, D. L., & Sweasy, J. B. (2012). Base excision repair and cancer. *Cancer Letters*, *327*(1-2), 73–89. http://doi.org/10.1016/j.canlet.2011.12.038

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., … Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, *11*(3), 333–337. http://doi.org/10.1038/nmeth.2810

Wong, M. (2013). Mammalian Target of Rapamycin (mTOR) Pathways in Neurological Diseases. *Biomedical Journal*, *36*(2). http://doi.org/10.4103/2319-4170.110365

Zaravinos, A., Pieri, M., Mourmouras, N., Anastasiadou, N., Zouvani, I., Delakas, D., & Deltas, C. (2014). Altered metabolic pathways in clear cell renal cell carcinoma: A meta-analysis and validation study focused on the deregulated genes and their associated networks. *Oncoscience*, *1*(2), 117–131.

ZENG, Z., QUE, T., ZHANG, J., & HU, Y. (2014). A study exploring critical pathways in clear cell renal cell carcinoma. *Experimental and Therapeutic Medicine*, *7*(1), 121–130. http://doi.org/10.3892/etm.2013.1392

Zhao, S., Guo, Y., & Shyr, Y. (2015). KEGGprofile: An annotation and visualization package for multi-types and multi-groups expression data in KEGG pathway. *R Package Version 1.10.0*.