Complementary Domain Prioritization: A Method to Improve Biologically Relevant Detection in Multi-Omic Data Sets

Benjamin A. Neely[1] and Paul E. Anderson[2]
[1]*Marine Biochemical Sciences, National Institute of Standards and Technology, Charleston, SC 29412, U.S.A.*
[2]*Department of Computer Science, College of Charleston, Charleston, SC 29424, U.S.A.*

Keywords: Genomics, Trancriptomics, Proteomics, Proteogenomics, Proteotranscriptomics.

Abstract: As the speed and quality of different analytical platforms increase, it is more common to collect data across multiple biological domains in parallel (*i.e.*, genomics, transcriptomics, proteomics, and metabolomics). There is a growing interest in algorithms and tools that leverage heterogeneous data streams in a meaningful way. Since these domains are typically non-linearly related, we evaluated whether results from one domain could be used to prioritize another domain to increase the power of detection, maintain type 1 error, and highlight biologically relevant changes in the secondary domain. To perform this feature prioritization, we developed a methodology called Complementary Domain Prioritization that utilizes the underpinning biology to relate complementary domains. Herein, we evaluate how proteomic data can guide transcriptomic differential expression analysis by analyzing two published colorectal cancer proteotranscriptomic data sets. The proposed strategy improved detection of cancer-related genes compared to standard permutation invariant filtering approaches and did not increase type I error. Moreover, this approach detected differentially expressed genes that would not have been detected using filtering alone while also highlighted pathways that might have otherwise been overlooked. These results demonstrate how this strategy can effectively prioritize transcriptomic data and drive new hypotheses, though subsequent validation studies are still required.

1 INTRODUCTION

Individually, the fields of genomics, transcriptomics, and proteomics continue to receive significant research attention as their utility as novel discovery platforms increases (Larance and Lamond, 2015; de Klerk and a.C. t Hoen, 2015); however, there is growing interest in algorithms and tools that leverage these heterogeneous data streams (Boja et al., 2014). This is for two main reasons: (i) it is becoming reasonable both from an experimental and cost perspective to run two or more analytical methods simultaneously (*e.g.*, proteomics and transcriptomics), and (ii) it is believed that integrating data sources will give rise to a deeper understanding of the system being interrogated. Thus using multidomain data in a non-trivial manner supports improved systems biology approaches to high-throughout biological analysis.

The most straightforward approach to harmonizing heterogeneous data streams is to combine *p*-values (Alves and Yu, 2014) or simply to identify statistical agreement following parallel analyses. For example, Zhang *et al.* utilized proteomic analysis to complement a seminal genomic/transcriptomic analysis of colorectal cancer (Cancer and Atlas, 2012). By analyzing the same samples, the analyses could be directly compared to identify shared changes at the gene, transcript and protein level such that hypotheses from the genomic study were confirmed at the protein level. In other words, identified differentially abundant proteins increased the confidence of differentially expressed genes. In addition to differential analyses of multiple domains, topological network approaches and gene set enrichment analysis can be used to predict activated/inhibited transcription factors. Agreement between this type of analysis of transcriptomic and proteomic data has been used in studies of renal cell carcinoma and psoriasis to identify disease-relevant transcription targets(Neely et al., 2016; Piruzian et al., 2010). These standard approaches make comparisons in parallel, as opposed to directly incorporating the data sets into a unified computational analysis. It has yet to be demonstrated how to capitalize fully on the relationship between genomic/transcriptomic and proteomic data sets.

Another approach to analyzing mutlidomain data

Neely B. and Anderson P.
Complementary Domain Prioritization: A Method to Improve Biologically Relevant Detection in Multi-Omic Data Sets.
DOI: 10.5220/0006151500680080
In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017)*, pages 68-80
ISBN: 978-989-758-214-1

sets is to combine information from features in multiple biological domains into a singular analytical space. Numerous tools that combine different data modalities are being developed (see (Kumar et al., 2016; Haider and Pal, 2013; Kuo et al., 2013) for extensive reviews). For example, tools such as 3Omics offer a web-based one-click tool to combine data from different domains by constructing correlation networks and co-expression profiles to highlight transcript/protein/metabolites with strong agreement between the domains (Kuo et al., 2013). There is also an effort to develop techniques that integrate networks between domains, such as metabolomic and proteomic data (Pirhaji et al., 2016), to identify relationships that are driving biological changes. It is evident that merging data from different domains into a shared analytical approach is a powerful approach.

A third approach is to use a unified computational method that relies on data from complementary domains to prioritize traditional methods of differential expression to highlight biologically relevant changes. Prioritization can be used to reduce the search space of an analytical domain by using data from a secondary domain. Such prioritization methods are becoming more important as the number of variables (*e.g.*, transcripts, proteins, metabolites) that are studied using next-generation technology may range from a few hundreds to tens of thousands. Transcriptomics can include gene arrays which measure tens of thousands of probes or RNA-seq analysis which can sequence and measure tens of thousands of transcripts, while proteomics (specifically mass spectrometry based shotgun proteomics) can provide a proteome sampling of around ten thousand proteins (Geiger et al., 2012). These high-dimensional search spaces, along with relatively small sample sizes, reduce the power to detect potentially biologically relevant changes while controlling the FDR. One of the most common forms of analysis in high-dimensional data such as these, is variable-by-variable statistical testing. This is used to test the null hypothesis that behavior for a given variable does not change between conditions. In the case of gene microarrays, this can be accomplished on a gene-by-gene level using a *t*-test, and extended to more complex experimental designs using ANOVA. For next-generation sequencing projects where read counts are available, *p*-values may be computed using read count statistics (Robinson and Smyth, 2007). For high-dimensional data this results in a large number of hypotheses that need to be evaluated, and thus, many genes may be detected as significant even though the null hypothesis is true (*i.e.*, there is really no change). This is known as type I error and is controlled with various false positive measures, including family-wise error rate (FWER) or false discovery rate (FDR), that control for type I error or the extent to which false positives occur but also reduces the power to detect true positives. A review of these methods can be found in (Dudoit et al., 2003).

One general methodology to reduce the transcriptomic search space is gene prioritization, which utilizes *a priori* phenotype relationships in published databases (reviewed extensively here (Börnigen et al., 2012)). Since this method requires manually curated and compiled information to prioritize experimental data, it can overestimate relationships once the data being queried becomes part of the *a priori* relationship, and it is limited to the search space of the database. Another method is to apply a filtering scheme prior to statistical testing. A comparison of these methods can be found in Bourgon *et al*. In these filtering schemes, a set of variables (*e.g.*, genes) are identified that generate uninformative signal. Then the formal statistical test is applied to the remaining features. The goal of this approach is to select a filter that makes multiple test correction less severe and thus enhance true positive detection.

In order to overcome *a priori* biases, several authors have recommended nonspecific or unsupervised filters that do not make use of sample class labels, and thus, have little influence on the formal statistical testing (Talloen et al., 2007). One criticism of supervised filtering techniques is that they are in fact a statistical test and may result in optimistic adjusted *p*-values and a true false positive rate that is larger than reported. The goal therefore of any filtering technique is to enrich true differential expression/abundance, if not then these techniques have the potential to generate overly optimistic conclusions. Common permutation-invariant filters include variance or abundance thresholds since it is assumed that this removes genes that are unlikely to be detected as significantly different and emphasize biologically relevant signal. While these filters are a common first step used by tools commonly used to analyze transcriptomic data (*e.g.*, limma, DESeq2, EdgeR, *etc*.), they do not incorporate data from multiple domains.

We propose a general method for feature prioritization and filtering that combines data from different biological domains and is built upon the biological relationship between independent data domains: Complementary Domain Prioritization. Unlike methods that attempt to merge domains into a singular analytical space, this method prioritizes features in the target domain by integrating empirical data from complementary domains through evidence-based relationships in curated databases. Herein, we demonstrate

how this general methodology can use proteomic data to guide transcriptomic differential expression analysis. The method utilizes heterogeneous data streams by leveraging available pathway data in Kyoto Encyclopedia of Genes and Genomes [KEGG; (Kanehisa et al., 2014)], WikiPathways (Kelder et al., 2009), and MSigDB:Transcription Factor Target databases (Matys et al., 2006) as a bridge between proteomic and transcriptomic data to increase the power to detect biologically relevant differential expression patterns in transcriptomic data. We apply this prioritization strategy to two independent studies with publicly available data. Our results indicate that this approach to prioritization followed by independent filtering provides an experimentally informed prioritization strategy leading to an increase in the power to detect meaningful changes while still controlling type I error.

2 METHODS

Complementary Domain Prioritization (CDP) is a two-stage prioritization and filtering approach that combines data from two biological domains with the goal of increasing the power of discovery while still controlling type I error (Figure 1). In the first stage, enrichment analysis is performed on one domain (in this example, proteomic data) using pathway or gene set databases. Gene lists from pathways or gene sets are then extracted and provide the input to the second stage, which applies these gene lists as prioritization criteria to the second domain (in this case, transcriptomic data). This is followed by a permutation-invariant gene list filtering approach. This general methodology, which can be extended to other domains, is demonstrated herein by using proteomic data to prioritize transcriptomic data. An R package has been developed during the course of this study and is available for general use by the research community. This repository contains all of the required scripts and documentation for their use https://github.com/Anderson-Lab/ComplementaryDomainPrioritization. This package provides a programmatic method to obtain the gene lists derived from the protein enrichment data.

2.1 Gene List Generation

The first stage of CDP is gene list generation, which in this example is driven by proteomic data analysis. Quantitative proteomic data was used along with class labels to detect differentially abundant

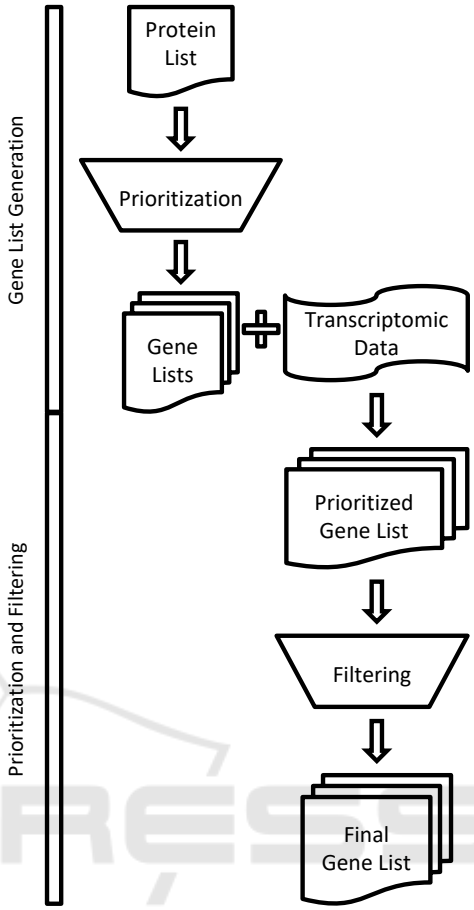


Figure 1: **Complementary Domain Prioritization flowchart.** This figure shows the overall workflow of prioritization and filtering.

proteins at a given threshold. This protein list was uploaded to WebGestalt (WEB-based GEne SeT AnaLysis Toolkit), a web-based tool that can perform pathway and gene set enrichment against different databases for eight different species (Wang et al., 2013), though *Homo sapiens* is the focus of this example. For each enrichment test the reference set was specified as protein-coding genes (*e.g.*, hsapiens_entrezgene_protein-coding) when using proteomic data. The statistical parameters used for the hypergeometric statistical method were BH p-value adjustment, requiring a significance level < 0.05 and minimum of two proteins per pathway/gene set. For clarification, transcription factor (TF) target analysis enrichment is gene set enrichment analysis against the transcription factor targets in the C3:motif gene set of MSigDB (Matys et al., 2006). Following KEGG, WikiPathways (WikiP) or TF enrichment analysis, the resulting pathways and gene sets were downloaded from WebGestalt as .tsv files.

2.2 Gene Prioritization and Filtering

The second stage of CDP first uses the enriched pathways or gene sets to prioritize the genes in the second domain belonging to each pathway or gene set of interest. Pathway information for KEGG was extracted using the KEGGREST R package. Pathway information for WikiPathways was retrieved using the official web service provided by WikiPathways. Gene set information from the transcription factor database (via MSigDB) was downloaded and queried locally. The prioritized gene list G is defined as:

$$P = \{P_i\}, \quad (1)$$

$$genes(P_i) = \{g_j\}, \quad (2)$$

$$G = \bigcup_{P_i \in P} genes(P_i), \quad (3)$$

where P is the set of all pathways or gene sets identified in stage 1, $genes(P_i)$ is the list of genes in pathway P_i, and G is the resulting prioritized gene list.

CDP prioritizes genes in the transcriptome involved in pathways or gene sets showing enrichment at the protein level and removing genes not present in these pathways or gene sets. Next, permutation-invariant filtering is applied to further enhance the power of detection by applying variance or mean abundance filtering. Variance filtering is defined as ranking the genes according to variance across samples (ignoring class labels). This has been shown to be similar to fold-change filtering, but in an unsupervised manner (see Bourgon *et al.* for a detailed discussion). Mean abundance filtering ranks the genes by the mean abundance of each gene. In the methodology described herein, these permutation-invariant filtering methods are applied after prioritization as follows:

$$Var(G_i) = \frac{1}{n-1} \sum_{j=1}^{n} (E_{ij} - \bar{E}_i)^2 \quad (4)$$

$$Mean(G_i) = \frac{1}{n} \sum_{j=1}^{n} E_{ij} \quad (5)$$

$$R = order(G) \quad (6)$$

$$F = \bigcup_{i=1 \ldots \theta} R_i \quad (7)$$

$$(8)$$

where E_{ij} is the expression estimate of gene i and sample j, R is an ordered list of genes that have been sorted by their variance ($Var(G_i)$) or mean abundance ($Mean(G_i)$), θ is the desired number of genes, and F is the final set of genes which have been prioritized and filtered.

2.3 Experimental Data Sets

Two data sets were selected for evaluation, referred to herein as TCGA CRCa and Marra. The first data set, TCGA CRCa, is comprised of transcriptomic and proteomic analyses of 87 tumor samples from individuals with colorectal carcinoma, for which transcript levels were measured by RNA sequencing yielding FPKM measurements (Cancer and Atlas, 2012), and protein levels were measured by label-free shotgun proteomic analysis yielding spectral counts that were quantiled and then log-transformed (Zhang et al., 2014). Transcriptomic data for the TCGA CRCa data set were retrieved from supplemental tables of the published paper (Cancer and Atlas, 2012). Using the key provided by both the authors and the key in the follow-up proteomics paper (Zhang et al., 2014), we determined the 87 samples which overlapped between the two studies. The transcriptomic data supplied in supplemental were utilized directly (as opposed to reprocessing raw data), which was given as FPKM values for 20,531 genes. These gene symbols were first confirmed as being current accepted HGNC symbols, followed by converting them to Entrez gene IDs. Of the original 20,531 genes, 18,995 had Entrez gene ID annotation and were used for analysis. A side note is that we understand that more robust statistics can be performed using direct count data with tools such as EdgeR or baySeq, but for our purposes comparing FPKM values with a t-test was acceptable.

Proteomic data for the TCGA CRCa data set were retrieved from supplemental tables of the published paper (Zhang et al., 2014). The supplemental data was used directly and consisted of already processed data that was spectral count data for 7,211 proteins, which had been quantiled then log transformed. Of these 7,211 proteins (with gene symbol identifiers), 7,147 had Entrez gene ID annotation and were used for analysis. To generate a p-value for these proteins to be used in CDP, a moderated t-test was applied to the data using the *limma* package (Smyth, 2004) followed by a Benjamini-Hochberg procedure to correct for multiple hypothesis testing. For this analysis, the samples are dichotomized based on whether lymphatic invasion was present in an attempt to classify tumor aggressiveness similar to the original analysis (Cancer and Atlas, 2012). The original paper utilized a model consisting of tumour stage, lymph node status, distant metastasis and vascular invasion at the time of surgery. Of the 87 samples analyzed, eight did not have a value specified, 44 had lymphatic invasion and 35 did not.

The second data set, Marra, is comprised of transcriptomic and proteomic analyses of paired adenoma

and normal mucosa samples from individuals with pre-cancerous colorectal tumors or lesions. Transcript levels of 35 adenomas with matched normal mucosa are measured using Affymetrix GeneChip Human Exon ST Arrays (Cattaneo et al., 2011). Transcriptomic data for the Marra data set (Cattaneo et al., 2011) were retrieved from the NCBI Gene Expression Omnibus database identifier GSE21962, which contained raw files from Affymetrix GeneChip Human Exon ST Arrays (HuEx-1_0-st-v2). Array data that was specific to pre-cancerous tumors and lesions was utilized, without duplicate samples resulting in 35 adenomas with matched normal mucosa. Files were processed using the *oligo* package (Carvalho and Irizarry, 2010) and pd.huex.1.0.st.v2 library. Robust multichip average (RMA) normalization was applied at the transcript level (*i.e.*, core gene ST probeset) generating expression values on a log-scale. This resulted in data for 20,011 probes, which contained 16,132 genes with Entrez gene ID annotation and were used for analysis.

Proteomic analysis of 30 similar but not identically matched samples was performed later using isobaric labeling (iTRAQ) to quantify protein abundance (Uzozie et al., 2014). Even though the Marra data set is comprised of data for similar but not identical samples, this is still relevant since often researchers use publicly available transcriptomic data to augment original analysis, such as Shimwell *et al.* (Shimwell et al., 2013). Proteomic data for the Marra data set (Uzozie et al., 2014) were retrieved from the published paper. This supplemental data was quantified at the peptide level using 8-plex iTRAQ across 10 experiments, with iTRAQ labels 113 and 114 being a pooled reference sample and the remaining six labels being three adenoma/normal mucosa pairs from the same patient per experiment (*i.e.*, 30 pairs total). These data were processed using InfernoRDN(Polpitiya et al., 2008) similar to the steps described by the authors(Uzozie et al., 2014). Briefly, for each experimental data set only unique peptides were utilized and label intensity measurements were first log2 transformed followed by mean centering (central tendency adjustment). These mean centered log transformed peptide level data were rolled up to the protein level using reference peptide based scaling (default parameters except one-hit wonders were allowed). This resulted in protein level quantification that was mean normalized log2 transformed by experiment. The distribution of 113 and 114 labels was found to be nearly identical between the 10 experiments, and so inter-experimental normalization was not required. Therefore we utilized the normalized log-intensity values for each adenoma/normal mucosa pair to generate a log-fold change value. Next we found proteins that were measured in all experiments, and used these 820 proteins for analysis. Of these 820 proteins with UniProtKB Identifiers, 768 had Entrez gene IDs, which were used to perform a moderated *t*-test with the *limma* package (Smyth, 2004) followed by a Benjamini-Hochberg procedure to correct for multiple hypothesis testing.

Since we use pathway and gene set databases to create prioritization filters, it is important to report relevant percent overlap related to the data sets and these databases. The TCGA CRCa data set is comprised of 7,147 unique proteins and 18,787 unique transcripts with Entrez gene IDs. The KEGG database contains 3,185 (45%) and 6,307 (34%) of these proteins and transcripts, respectively. The WikiP database contains 3,719 (52%) and 7,215 (38%) of these proteins and transcripts, respectively. The TF database contains 5,208 (73%) and 12,035 (64%) of these proteins and transcripts, respectively. The Marra data set is comprised of 768 unique proteins and 15,807 unique transcripts with Entrez gene IDs. The KEGG database contains 474 (62%) and 6,073 (38%) of these proteins and transcripts, respectively. The WikiP database contains 533 (69%) and 6,841 (43%) of these proteins and transcripts, respectively. The TF database contains 628 (82%) and 11,556 (73%) of these proteins and transcripts, respectively. The relative overlap is higher for protein than transcript data sets against the pathway databases (KEGG and WikiP) since pathways are comprised of proteins. Also for reference, there are 27,228 genes with Entrez gene IDs in the *Homo sapiens* genome, and 21,061 of these are protein coding (Ensembl 81, *H. sapiens* GRCh38.p3).

3 RESULTS

3.1 Prioritization and Novel Discovery

The goal of Complementary Domain Prioritization (CDP) is to guide differential analysis in one domain using differential analysis of a parallel domain. The proposed method was evaluated using two published proteotranscriptomic data sets, TCGA CRCa and Marra, with diverse gene expression and protein abundance patterns. The TCGA CRCa data set presents an example of a difficult differential expression discovery task when studying differences in tumor aggressiveness using lymphatic invasion as a phenotype. The original TCGA analysis identified 40 genes related to tumor aggressiveness, which was determined by more factors than just lymphatic invasion (Cancer and Atlas, 2012). In the current anal-

ysis, using lymphatic invasion as a phenotype there were minor changes with respect to differential expression at the transcript level with 49 differentially expressed genes (two-sided equal t-test, Benjamini-Hochberg corrected $p < 0.2$; BH < 0.2) and 26 differentially abundant proteins (BH < 0.2). Of these 49 differentially expressed genes, two were identified in the original analysis. In contrast, using CDP detected on average 10 of the 40 genes identified in the original TCGA study by filtering alone or prioritization then filtering (data not shown). This provided a modest improvement in analysis to detect genes related to tumor aggressiveness.

In order to demonstrate whether CDP was prioritizing otherwise disregarded genes that are interesting candidates correlated to cancer phenotypes, we report the number of rejected null hypotheses (*e.g.*, number of candidate discoveries) as a function of the FDR for only cancer-related genes. To accomplish this, a list of 1663 cancer-related genes, retrieved via the Human Protein Atlas (Uhlén et al., 2015), were tracked and evaluated for each data set (Figure 2 and Supplementary Figure S1 online). In the TCGA CRCa data set prioritization followed by filtering consistently detected approximately two to three-fold more cancer-related transcripts with BH < 0.3 than using filtering only, and higher θ resulted in better detection (Figure 2). Also, cancer-related transcripts were detected at BH < 0.1 only when using WikiPathways (WikiP) and transcription factor (TF) enrichment based prioritization. The effect of CDP on the Marra data set, which had a strong differential expression profile with 40% of genes differentially expressed without CDP, was less obvious (Figure 3). For the Marra data set, filter-only performed as well as CDP in some cases with CDP outperforming filter-only when used with WikiP (Figure 3b and Figure 3e).

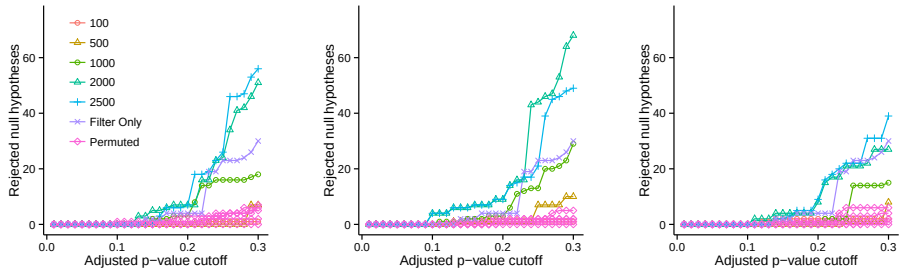
In addition to detecting previously reported genes of interest, successful prioritization should also result in detecting genes that would otherwise not be detected. The Marra data set has 6,300 differentially expressed genes (BH < 0.5) and 437 differentially abundant proteins (BH < 0.5) and provides an interesting example of the effect of CDP on prioritizing gene detection. Using CDP drastically decreased the number of gene candidates by prioritizing 6 to 43% of the original 15,807 genes prior to permutation invariant filtering. Additional filtering targeted 3 to 22% of the original genes (488, 813 and 3,427 using WikiP, KEGG, and TF, respectively). Prioritization followed by variance filtering detected 6 to 164 genes that would not have been detected with variance filtering alone, while prioritization followed by mean abundance filtering detected less than 10 genes that would not have been detected with mean abundance filtering alone. Although these changes in detection may seem minor, they lead to large differences at the pathway level, which is investigated later in this paper.

3.2 Controlling False Positive Rate

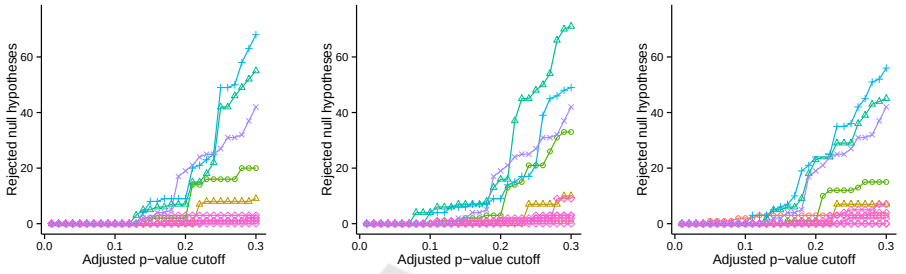
Using filters to improve detection power can lead to a loss of false positive control depending on the choice of filter. Permutation-invariant filters, such as variance and abundance, have been shown to be appropriate filters (Bourgon et al., 2010). An example of a permutation-variant filter is a fold-change filter, where the fold-change between two classes is dependent on the labels of the samples, and thus, the ordering of the samples is important in the calculation. In contrast, a variance filter, abundance filter, or complementary domain based prioritization is independent on the ordering of the samples. To evaluate false positive control, the conditional and unconditional marginal distributions of test statistics after using permutation-invariant filtering were compared. It has been shown that the conditional marginal distributions of test statistics after using permutation-invariant filtering are the same as the unconditional distributions before filtering, where the conditional distributions are the same as the distributions after applying the filter (Bourgon et al., 2010). This behavior is a necessary criteria for multiple testing adjustments that attempt to control experiment-wide type I error rate. We show the distribution of t-statistics for the conditioned and unconditioned TCGA CRCa data set in order to empirically demonstrate the effect of these filters (Figure 4). The findings on the Marra data set were also consistent with these results (data not shown). We deliberately eschewed fold-change from other comparisons since this requires incorporating class labels, and thus, will affect type I error. A variance based filter is similar in practice to a fold-change filter, despite being independent of class levels. This is discussed in detail by Bourgon *et al.*, but briefly, for small sample sizes, the bound is essentially a constant multiple of the cut-off on the variance.

3.3 Effect on Pathway Analysis

To more clearly define the similarities in the prioritization strategies, we compared the overlap of prioritized and differentially expressed genes using different databases for prioritization. Since there are only 26 differentially abundant proteins in the TCGA CRCa data set (at BH < 0.2), we did not pursue using this data set as a proof of principle. Differentially

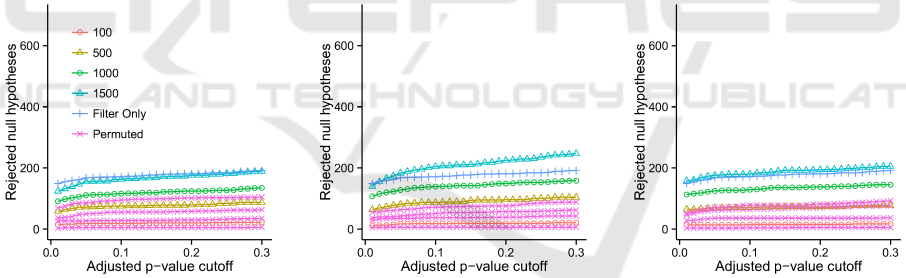


(a) KEGG Variance Filter (b) WikiP Variance Filter (c) TF Variance Filter

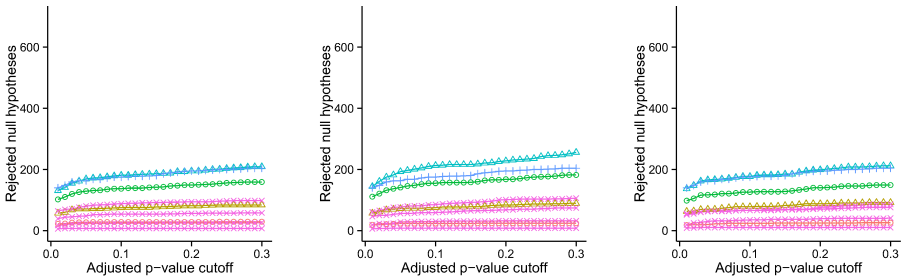


(d) KEGG Mean Abundance Filter (e) WikiP Mean Abundance Filter (f) TF Mean Abundance Filter

Figure 2: **Power analysis and prioritization comparison on the TCGA CRCa data set.** Figures a - f show the number of rejected null hypotheses as a function of the BH adjusted *p*-value cutoff using three different prioritization strategies (KEGG, WikiP and TF) and two different permutation invariant filtering strategies (variance and mean abundance). All methods were evaluated using thresholds θ=100, 500, 1000, 2000, and 2500, where θ is the desired number of genes. For presentation clarity, only a single θ is shown for methods using only a permutation invariant strategy (filter only, no prioritization), which was the threshold resulting in the maximum number of significant genes for a FDR of 0.2, and is labeled as Filter Only.



(a) KEGG Variance Filter (b) WikiP Variance Filter (c) TF Variance Filter



(d) KEGG Mean Abundance Filter (e) WikiP Mean Abundance Filter (f) TF Mean Abundance Filter

Figure 3: **Power analysis and prioritization comparison on the Marra data set.** Figures a - f show the number of rejected null hypotheses as a function of the BH adjusted *p*-value cutoff using three different prioritization strategies (KEGG, WikiP and TF) and two different permutation invariant filtering strategies (variance and mean abundance). All methods were evaluated using thresholds θ=100, 500, 1000, 2000, and 2500, where θ is the desired number of genes. For presentation clarity, only a single θ is shown for methods using only a permutation invariant strategy (filter only, no prioritization), which was the threshold resulting in the maximum number of significant genes for a FDR of 0.2, and is labeled as Filter Only.

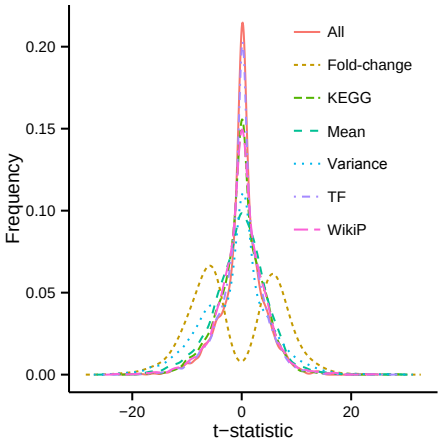


Figure 4: **Distribution of t-statistic values before and after conditioning.** The proposed prioritization and filtering approaches are permutation invariant, in contrast to a fold-change filter that is not permutation invariant.

abundant proteins in the Marra data set (437 proteins at BH < 0.05) were used to perform pathway enrichment or gene set analysis against KEGG, WikiP or TF databases (Figure 5, and Figure 6). Regardless of the permutation-invariant filter used following prioritization, the trends were the same between the databases: TF based CDP resulted in the highest number of genes, of which 80% were unique to TF based prioritization, while KEGG and WikiP prioritized far fewer genes. Regardless of the prioritization approach used, approximately half of the resulting genes were detected as differentially expressed, which was the same as in the untreated data set. Lastly, the KEGG and WikiP approaches were more similar than TF, likely since these are pathway databases as opposed to gene set databases.

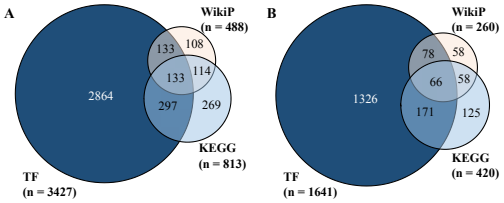


Figure 5: **Comparison of the gene overlap following prioritization treatments after variance filtering.** Different prioritization approaches were used (KEGG, WikiP or TF) followed by an independent filter (50% variance) for the Marra data set. **(A)** Overlap of prioritized genes following prioritization with different databases. **(B)** Overlap of differentially expressed genes (BH adjusted $p < 0.05$) following prioritization with different databases.

The candidate discoveries identified in a differential expression studies are often the starting point for pathway and gene set enrichment analyses; therefore,

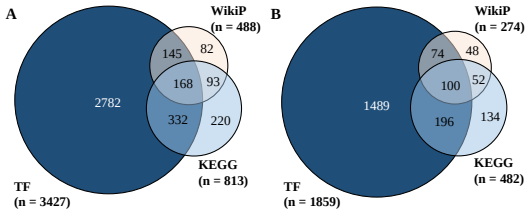


Figure 6: **Comparison of the gene overlap following prioritization treatments after mean abundance filtering.** Different prioritization approaches were used (KEGG, WikiP or TF) followed by an independent filter (50% mean abundance) for the Marra data set. **(A)** Overlap of prioritized genes following prioritization with different databases. **(B)** Overlap of differentially expressed genes (BH adjusted $p < 0.05$) following prioritization with different databases.

it is critical to explore how the proposed prioritization strategy affects downstream analysis. For this we compared how filtering alone and prioritization followed by filtering affected results when analyzing the Marra data set (Table 1) and report the identified pathways. We used the differentially expressed genes (BH < 0.05) from these prioritized and/or filtered data to perform pathway enrichment analysis against the KEGG database. The unfiltered data resulted in 206 pathways, variance filtering alone identified 189 pathways and prioritizing with KEGG, WikiP and TF followed by variance filtering identified 151, 87 and 154 pathways, respectively. Since often the top enriched pathways are used for followup experimental studies, the top 10 enriched pathways were compared (Table 1). Each approach yielded slightly different top 10 pathways and different rankings of specific pathways, with only 'Metabolic pathways' being shared across approaches. Interestingly, the 'Complement and coagulation cascades' pathway was ranked highest after protein prioritization (using WikiP) with 26 genes being differentially expressed, while it was fourth highest using KEGG (albeit with 29 genes being differentially expressed). For comparison, this pathway was ranked 17th using a variance filter alone (with 28 differentially expressed genes). Also, the 'Protein processing in ER' pathway was unique to KEGG based protein prioritization, with 33 genes in this pathway being differentially expressed. When using a variance filter alone, this pathway was the 99th highest ranked pathway with 30 differentially expressed genes, while in the unfiltered treatment this pathway was ranked 20th with 72 differentially expressed genes (data not shown). For comparison, 'Protein processing in the ER' was ranked 15th with 13 differentially abundant proteins following pathway enrichment analysis using the proteomic data. Overall, Using CDP on the Marra data set changed the ranking of identified path-

ways, possibly indicating pathways that are more biologically relevant than those identified using just the transcriptomic results.

4 DISCUSSION

As the speed, quality and complexity of analysis used to quantify the complex biochemical interactions within biological systems continues to improve, studies are often confronted with the issue of over-fitting high-dimensional data with a relatively small sample set. These data sets can also span multiple biological domains (genes, transcripts, proteins, metabolites), presenting the opportunity to utilize the data across domains in a meaningful way. It was the aim of this study to develop and evaluate a method that leveraged proteomic data to prioritize transcriptomic data while avoiding increased type I error. Similar to methods that use permutation-invariant filters such as variance or mean abundance to improve power (Bourgon et al., 2010), our approach improved the power to detect cancer-related genes while controlling type I error in two different experimental data sets. These results demonstrate that prioritization of transcriptomic data using proteomic data provides similar power improvements as other permutation-invariant filters, while utilizing the underpinning biological hierarchy to create an empirical prioritization filter.

One issue when applying filters to data is that using class level differences to filter or prioritize data can inflate the number of significant variables. This problem is mitigated during Complementary Domain Prioritization because the domains are independent and class level differences across different biological domains are largely non-linearly related. Even though the majority of transcript and protein levels are positively correlated, the average correlation from a subset of published data is 0.26 (Gygi et al., 1999; Foss et al., 2007; Zhang et al., 2014). It is possible that genes detected as differentially expressed do not manifest changes at the protein level due to post-transcriptional and post-translational regulation. This has been observed when proteomic analysis has been used in tandem with transcriptomic studies (Zhang et al., 2014; Gygi et al., 1999; Foss et al., 2007), where changes observed at the transcript level were not propagated to the protein level. Alternatively, significant protein changes have been observed in the absence of transcript changes (Zhang et al., 2014; Gygi et al., 1999), further emphasizing the dynamic nature of post-transcriptional regulation, which can attenuate protein abundance, as well as the independence of the different domains. By utilizing protein levels, which are arguably more informative of phenotypic changes while also being non-linearly related to transcript changes, to prioritize pathways and gene sets in the transcript domain, Complementary Domain Prioritization does not artificially inflate detection of significant changes across domains.

During Complementary Domain Prioritization a filter is created using enrichment analysis against pathway and gene set databases, thereby incorporating secondary experimental data with evidence-based relationships. In other words, CDP is able to leverage evidence-based relationships (databases) with secondary domain data (proteomics) to prioritize signal at the transcriptomic level. Since Complementary Domain Prioritization relies heavily on databases and secondary domain data, both of these offer room for improvement. Databases are inherently ambiguous association lists that aren't exhaustive or completely accurate. This means that the choice of database can affect downstream results since they rely on different assumptions and curation. We show that different prioritization databases affect detection of differentially expressed genes, with Complementary Domain Prioritization targeting 6 to 43% of the genes detected by filtering alone. Also, prioritization using pathway based databases (KEGG and WikiPathways) generated results that were more similar than prioritization based on a gene set database (MSigDB). Moreover, if a database omits a gene in a given pathway or gene set, then this gene will not be present following the prioritization procedure. In both data sets, the pathway based databases include approximately 40% of transcripts in the transcriptomic data, whereas the gene set database includes approximately 75% of both proteomic or transcriptomic data. Also, though the WikiPathways database consistently included more proteins and transcripts from the data sets than KEGG, fewer WikiPathways pathways were identified by pathway enrichment analysis. This may explain why although both KEGG and WikiPathways prioritization strategies identified the 'Complement and Coagulation Cascades' pathway in the Marra data set, though the number of differential expressed genes was different between the two prioritization strategies. We have shown performance using KEGG, WikiPathways, and MSigDB (*i.e.*, TF) databases, but in the future this could be expanded to include other databases (*e.g.*, Reactome). As more heterogeneous domain data sets become available, isolating database specific effects from data set effects will help establish best practices for Complementary Domain Prioritization. In general, with continued development of more accurate databases, the prioritization quality of this approach will improve.

Table 1: **Comparison of top 10 enriched pathways.** Results for the Marra data set are shown using four different strategies: permutation invariant filter only (50% variance) and prioritization (KEGG, WikiP or TF) followed by filter (50% variance). After ranking enriched pathways using the BH adjusted *p*-value, the top 10 are shown in order. Unique pathways are in bold.

Filter Only	**KEGG + Filter**	**WikiP + Filter**	**TF + Filter**
Metabolic pathways	Metabolic pathways	Complement and coagulation cascades	Metabolic pathways
Cell cycle	Purine metabolism	Metabolic pathways	Cell cycle
Cell adhesion molecules (CAMs)	RNA transport	Ribosome	Pathways in cancer
Pathways in cancer	Complement and coagulation cascades	**Parkinson's disease**	Purine metabolism
Cytokine-cytokine receptor interaction	**Protein processing in endoplasmic reticulum**	**Oxidative phosphorylation**	p53 signaling pathway
Purine metabolism	**Pyrimidine metabolism**	Alzheimer's disease	**DNA replication**
Chemokine signaling pathway	Huntington's disease	Huntington's disease	**MAPK signaling pathway**
p53 signaling pathway	Alzheimer's disease	**Starch and sucrose metabolism**	RNA transport
Rheumatoid arthritis	Ribosome	**Staphylococcus aureus infection**	Cell adhesion molecules (CAMs)
Regulation of actin cytoskeleton	**Glycolysis / Gluconeogenesis**	**Systemic lupus erythematosus**	**Prostate cancer**

In addition to reliance on databases, Complementary Domain Prioritization relies on the data quality of the secondary domain. Current shotgun proteomics data sets are prone to false negatives by the very nature of mass spectrometric analysis. As newer techniques and technologies are developed and perfected, such as data independent analysis, false negative rates should decrease though this still poses a problem when integrating across domains. For example, the type of sample preparation and separation may not perform well at isolating and identifying membrane-bound proteins, which could handicap the Complementary Domain Prioritization procedure away from signaling cascades. Proteomic analysis also varies greatly in coverage (or sampling depth) of the proteome based on separation techniques and speed/resolution of the mass spectrometer. Our results demonstrate that a set of 800 proteins (Marra data set) or 7,000 proteins (TCGA CRCa data set) can successfully be used to prioritize companion transcriptomic data. It is likely that there are more false negatives in the smaller protein data set and using this set is likely prioritizing pathways with higher abundance proteins. If the goal is to discover a biomarker with clinical utility, then this limitation is actually an advantage at prioritizing transcript changes that are related to measurable protein changes. Regardless, as both proteomic analysis and database quality improves, Complementary Domain Prioritization will also improve.

It was not the goal of this study to compare results to the original transcriptomic analysis. Regardless, using Complementary Domain Prioritization followed by unsupervised filtering performed better at detecting known cancer-related genes in both data sets relative to filtering alone or using no filter at all. In the first data set, the TCGA authors investigated gene expression related to tumor aggressiveness, though tumor aggressiveness was not evaluated in the follow-up proteomic analysis by Zhang *et al*. Using lymphatic invasion as a proxy of tumor aggressiveness, our method identified more of the 40 genes related to tumor aggressiveness than filtering alone (six versus four). If the pathways and gene sets prioritized by the protein data did not include these 40 genes, then they were omitted from further analysis. On average only 10 of these genes were selected following filtering alone or prioritization followed by filtering. This finding demonstrates how Complementary Domain Prioritization can be used to improve previous results based on a single domain. In the second data set evaluated, Cattaneo *et al.* evaluated polyploid pre-cancerous colorectal lesions whereas the follow-up proteomics analysis by Uzozie *et al.*, and therefore the prioritization analysis described herein, focused on normal versus pre-cancerous lesions. The published proteomic analysis identified alterations in sorbitol dehydrogenase (SORD) levels in pre-cancerous lesions, specifically in the sorbitol-aldose reductase pathway. Though this pathway was only ranked 163

in the unfiltered transcriptomic data, following Complementary Domain Prioritization (KEGG-based with a variance filter of 50%), the rank improved to 29. This clearly demonstrates that proteomics data can prioritize transcriptomic data in a meaningful way similar to published results.

In addition to having similar findings as the original studies, Complementary Domain Prioritization was able to highlight pathways that might have been otherwise overlooked. Using Complementary Domain Prioritization and filtering of the Marra data set identified 'Protein Processing in the ER' as being different in pre-cancerous colorectal lesions. Although only 13 differentially abundant proteins were identified in this pathway, the transcriptomic data was prioritized such that this pathway was ranked 5th (versus 20th in the unfiltered data). Protein folding is known to be crucial in many oncogenic processes, especially ER chaperones (Luo and Lee, 2012), and may be an area of research to pursue further. These targets could be confirmed at the protein level by immunoblot analysis similar to the SORD confirmation in the original study. Lastly, by using both the KEGG and WikiP prioritization approaches we identified similar pathways emphasizing the importance of the complement cascade in pre-cancerous lesions. By applying Complementary Domain Prioritization to the Marra data set, we have shown how complementary proteomic data can drive new hypotheses, though only future experiments will demonstrate the relevance of these findings.

5 CONCLUSIONS

Utilizing data across biological domains is inherently difficult because it is not fundamentally understood how multiple changes at each domain result in phenotypic changes. Significant changes at the transcript level are not always present at the protein level (Zhang et al., 2014; Gygi et al., 1999; Foss et al., 2007), while studies of the low-abundance transcriptome of some cancer have confirmed that changes in low-abundance genes are responsible for deleterious biological changes (Bizama et al., 2014). This may be extrapolated to other domain relationships meaning that moving from gene to transcript to protein to metabolite, changes at each step are not dependent on single changes at previous steps. Likewise, this means that signal from each domain is representative of signal from a larger number of features in the previous domain. For this reason, we did not focus on directly combining data between domains but instead present an approach that utilizes results from one domain to prioritize data from the underlying domain. Our results demonstrate that proteomic data can be used to prioritize transcriptomic data, though this approach is not limited to these complementary domains. Lipidomic data could be used to prioritize genomic/transcriptomic/proteomic data via LIPID MAPS pathway database, and metabolomics data could be likewise used to prioritize data following analysis with XCMS Online. Stepping further away from genes in this hierarchy of biological domains creates smaller and smaller prioritization lists, but ones that are more biologically relevant to phenotypic changes. Utilizing data from complementary domains as a prioritization tool can be a powerful approach to integrating complex high-dimensional biological data sets.

ACKNOWLEDGEMENTS

The authors wish to thank Benilton Carvalho for his help utilizing the *oligo* R package as well as Giancarlo Marra and Anuli Uzozie for invaluable assistance in harmonizing published data sets. This research was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM103542, the South Carolina SmartState Center of Economic Excellence in Proteomics, and the Medical University of South Carolina Proteomics Center. Identification of certain commercial equipment, instruments, software or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.

REFERENCES

Alves, G. and Yu, Y. K. (2014). Accuracy evaluation of the unified P-value from combining correlated P-values. *PLoS ONE*, 9(3).

Bizama, C., Benavente, F., Salvatierra, E., Gutiérrez-Moraga, A., Espinoza, J. a., Fernández, E. a., Roa, I., Mazzolini, G., Sagredo, E. a., Gidekel, M., and Podhajcer, O. L. (2014). The low-abundance transcriptome reveals novel biomarkers, specific intracellular pathways and targetable genes associated with advanced gastric cancer. *International Journal of Cancer*, 134:755–764.

Boja, E. S., Kinsinger, C. R., Rodriguez, H., and Srinivas, P. (2014). Integration of omics sciences to advance biology and medicine. 11(1):1–12.

Börnigen, D., Tranchevent, L. C., Bonachela-Capdevila, F., Devriendt, K., De Moor, B., De Causmaecker, P., and

Moreau, Y. (2012). An unbiased evaluation of gene prioritization tools. *Bioinformatics*, 28(23):3081–3088.

Bourgon, R., Gentleman, R., and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 107:9546–9551.

Cancer, T. and Atlas, G. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–7.

Carvalho, B. S. and Irizarry, R. a. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19):2363–2367.

Cattaneo, E., Laczko, E., Buffoli, F., Zorzi, F., Bianco, M. A., Menigatti, M., Bartosova, Z., Haider, R., Helmchen, B., Sabates-Bellver, J., Tiwari, A., Jiricny, J., and Marra, G. (2011). Preinvasive colorectal lesion transcriptomes correlate with endoscopic morphology (polypoid vs. nonpolypoid). *EMBO molecular medicine*, 3(6):334–47.

de Klerk, E. and a.C. t Hoen, P. (2015). Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends in Genetics*, 31(3):128–139.

Dudoit, S., Dudoit, S., Shaffer, J. P., Shaffer, J. P., Boldrick, J. C., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103.

Foss, E. J., Radulovic, D., Shaffer, S. a., Ruderfer, D. M., Bedalov, A., Goodlett, D. R., and Kruglyak, L. (2007). Genetic basis of proteome variation in yeast. *Nature genetics*, 39(11):1369–1375.

Geiger, T., Wehner, a., Schaab, C., Cox, J., and Mann, M. (2012). Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Molecular & Cellular Proteomics*, 11:M111.014050–M111.014050.

Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology*, 19(3):1720–1730.

Haider, S. and Pal, R. (2013). Integrated analysis of transcriptomic and proteomic data. *Current genomics*, 14(2):91–110.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Research*, 42(D1):199–205.

Kelder, T., Pico, A. R., Hanspers, K., Van Iersel, M. P., Evelo, C., and Conklin, B. R. (2009). Mining biological pathways using WikiPathways web services. *PLoS ONE*, 4(7):2–5.

Kumar, D., Bansal, G., Narang, A., Basak, T., Abbas, T., and Dash, D. (2016). Integrating transcriptome and proteome profiling: Strategies and applications. *Proteomics*, pages 1–12.

Kuo, T.-C., Tian, T.-F., and Tseng, Y. J. (2013). 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC systems biology*, 7(1):64–78.

Larance, M. and Lamond, A. I. (2015). Multidimensional proteomics for cell biology. *Nature Reviews Molecular Cell Biology*, 16(5):268–80.

Luo, B. and Lee, a. S. (2012). The critical roles of endoplasmic reticulum chaperones and unfolded protein response in tumorigenesis and anticancer therapies. *Oncogene*, 32(7):805–818.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, a. E., and Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(Database issue):D108–D110.

Neely, B. A., Wilkins, C. E., Marlow, L. A., Malyarenko, D., Kim, Y., Ignatchenko, A., Sasinowska, H., Sasinowski, M., Nyalwidhe, J. O., Kislinger, T., Copland, J. A., and Drake, R. R. (2016). Proteotranscriptomic Analysis Reveals Stage Specific Changes in the Molecular Landscape of Clear-Cell Renal Cell Carcinoma. *PloS one*, 11(4):e0154074.

Pirhaji, L., Milani, P., Leidl, M., Curran, T., Avila-Pacheco, J., Clish, C. B., White, F. M., Saghatelian, A., and Fraenkel, E. (2016). Revealing disease-associated pathways by network integration of untargeted metabolomics. *Nature methods*, 13(9):770–776.

Piruzian, E., Bruskin, S., Ishkin, A., Abdeev, R., Moshkovskii, S., Melnik, S., Nikolsky, Y., and Nikolskaya, T. (2010). Integrated network analysis of transcriptomic and proteomic data in psoriasis. *BMC systems biology*, 4(41).

Polpitiya, A. D., Qian, W. J., Jaitly, N., Petyuk, V. a., Adkins, J. N., Camp, D. G., Anderson, G. a., and Smith, R. D. (2008). DAnTE: A statistical tool for quantitative analysis of -omics data. *Bioinformatics*, 24(13):1556–1558.

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics (Oxford, England)*, 23(21):2881–7.

Shimwell, N. J., Bryan, R. T., Wei, W., James, N. D., Cheng, K. K., Zeegers, M. P., Johnson, P. J., Martin, a., and Ward, D. G. (2013). Combined proteome and transcriptome analyses for the discovery of urinary biomarkers for urothelial carcinoma. *British journal of cancer*, 108(9):1854–61.

Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical applications in genetics and molecular biology*, 3(1):1–26.

Talloen, W., Clevert, D.-A., Hochreiter, S., Amaratunga, D., Bijnens, L., Kass, S., and Göhlmann, H. W. H. (2007). I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics (Oxford, England)*, 23(21):2897–902.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-k., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-h., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., Feilitzen, K. V., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., Heijne, G. V., Nielsen, J., and Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419.

Uzozie, A., Nanni, P., Staiano, T., Grossmann, J., Barkow-Oesterreicher, S., Shay, J. W., Tiwari, A., Buffoli, F., Laczko, E., and Marra, G. (2014). Sorbitol dehydrogenase overexpression and other aspects of dysregulated protein expression in human precancerous colorectal neoplasms: a quantitative proteomics study. *Molecular & Cellular Proteomics*, 13(5):1198–1218.

Wang, J., Duncan, D., Shi, Z., and Zhang, B. (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic acids research*, 41(Web Server issue):77–83.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J. C., Carr, S. a., Tabb, D. L., Coffey, R. J., Slebos, R. J. C., and Liebler, D. C. (2014). Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382–387.