# ENABLING DATA-GUIDED EVALUATION OF BIOINFORMATICS WORKFLOW QUALITY

by

Kevin Kristopher McDade

BS, University of Pittsburgh, 2000

MS, Chatham University, 2007

MS, University of Pittsburgh, 2011

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Kevin Kristopher McDade

It was defended on

April 3, 2017

and approved by

Uma Chandran, Research Associate Professor, Biomedical Informatics

Harry Hochheiser, Assistant Professor, Biomedical Informatics

Xinghua Lu, Professor, Biomedical Informatics

Daniel E. Weeks, Professor, Human Genetics

Roger Day, Associate Professor, Biomedical Informatics

Dissertation Director: Vanathi Gopalakrishnan, Associate Professor, Biomedical Informatics

**ENABLING DATA-GUIDED EVALUATION OF BIOINFORMATICS WORKFLOW QUALITY**

Kevin Kristopher McDade, PhD.

University of Pittsburgh, 2017

Bioinformatics can be divided into two phases, the first phase is conversion of raw data into processed data and the second phase is using processed data to obtain scientific results. It is important to consider the first "workflow" phase carefully, as there are many paths on the way to a final processed dataset. Some workflow paths may be different enough to influence the second phase, thereby, leading to ambiguity in the scientific literature. Workflow evaluation in bioinformatics enables the investigator to carefully plan how to process their data. A system that uses real data to determine the quality of a workflow can be based on the inherent biological relationships in the data itself. To our knowledge, a general software framework that performs real data-driven evaluation of bioinformatics workflows does not exist.

The Evaluation and Utility of workFLOW (EUFLOW) decision-theoretic framework, developed and tested on gene expression data, enables users of bioinformatics workflows to evaluate alternative workflow paths using inherent biological relationships. EUFLOW is implemented as an R package to enable users to evaluate workflow data. EUFLOW is a framework which also permits user-guided utility and loss functions, which enables the type of analysis to be considered in the workflow path decision. This framework was originally developed to address the quality of identifier mapping services between UNIPROT accessions and Affymetrix probesets to facilitate integrated analysis[1]. An extension to this framework evaluates Affymetrix probeset filtering methods on real data from endometrial cancer and TCGA ovarian serous carcinoma samples.[2] Further evaluation of RNASeq workflow paths demonstrates generalizability of the

EUFLOW framework. Three separate evaluations are performed including: 1) identifier filtering of features with biological attributes, 2) threshold selection parameter choice for low gene count features, and 3) commonly utilized RNASeq data workflow paths on The Cancer Genome Atlas data.

The EUFLOW decision-theoretic framework developed and tested in my dissertation enables users of bioinformatics workflows to evaluate alternative workflow paths guided by inherent biological relationships and user utility.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# PREFACE

This work, in part, uses figures, concepts and extends upon a mathematical framework from three published works of which I am a primary contributor and author. An early foundation for this work was presented in, "*Identifier mapping performance for integrating transcriptomics and proteomics experimental results*".[3] Our research team evaluated mapping identifiers between gene expression and protein expression platforms using the data from these platforms as a guide. A decision-theoretic framework utilized transcript to protein correlation across cancer samples. The mathematical framework was developed and published in, "*A decision theory paradigm for evaluating identifier mapping and filtering methods using data integration*[1]", As the primary author, I extended the framework to include a full evaluation of identifier filtering of Affymetrix gene expression data in, "*Improving cancer gene expression data quality through a TCGA data-driven evaluation of identifier filtering*".[2] In this dissertation, I present the application of this framework to gene expression data processing problems. I also present the software EUFLOW (Evaluation and Utility of workFLOWs) to enable users of bioinformatics data to apply the framework to their own workflow choices.

In Chapter 1, I will present the basic terminology for this dissertation, and discuss the need for general evaluation of workflows using real data. Chapter 2 will review the proliferation of bioinformatics workflows, gene expression workflow paths, and the use of correlation as a means to evaluate gene expression data. In Chapter 3, I will briefly present the general mathematical framework of EUFLOW. In Chapter 4, I will present a vignette of the EUFLOW package, which enables users to execute the framework in the R environment as a package. In Chapter 5, I will

demonstrate the use of EUFLOW to evaluate RNASeq workflows. In Chapter 6, I will discuss the relevance of workflow evaluation to bioinformatics clinical implementation, the innovation of the EUFLOW framework, the generalizability of the EUFLOW framework, and my future direction of workflow evaluation including further development of the EUFLOW package.

I thank my long-time advisor and friend, Roger Day, Sc.D., for his endless patience, devoted teaching, and vast impact on my life as a teacher and scientist. Dr. Day has coached me through writing, coding and the general organization of this work. Dr. Day has provided the foundation for the development of the EUFLOW package and has pointed me in the direction for my future work and endeavors.

I thank, posthumously, M. Michael Barmada, Ph.D., who believed in the problem that I seek to solve in this dissertation, which gave me the courage to proceed in this direction. I also thank Dr. Barmada for his service as the committee chair, and also am thankful for his tutorage on the processing of RNASeq data.

I also thank my recently appointed Chair and Major advisor, Vanathi Gopalakrishnan Ph.D., for her advice over the years and her time in the final weeks of this work for helping me with clarity, purpose, and writing of this work.

Thank you to my committee members for their advice and guidance over the years and in the recent days leading up to the defense; to Uma Chandran, Ph.D. for inspiring me to solve this problem from the very beginning and her guidance in the last decade; and to Harry Hochheiser Ph.D. for pushing me to complete this work, providing a new perspective, and helping me to assemble my committee. And for providing guidance over the years, thank you to Xinghua Lu PhD for helping state the challenges of this work. And finally, thank you to Daniel Weeks for joining my committee at a late stage of this dissertation and for providing examples of evaluations

related to this dissertation. Thank you to Toni Porterfield for the many times that she has made my life easier during my time at DBMI. Also, thank you to the countless fellow students over the years for their input and time that impacted my learning and research in a positive way. Thank you to the National Library of Medicine for supporting my training (T15 LM 007059).

Thank you to my parents, Deborah and Kevin, for believing in me, encouraging me to follow my dreams and instilling a strong work ethic. To my children, Rachel, Riley, and Norah, I hope that you follow your own dreams and never quit until you achieve those dreams. And to my wife, Amyjo, thank you for pushing me every step along the way, believing in me, and providing me with the support and companionship that made this work possible.

# 1.0 INTRODUCTION

The ultimate focus of bioinformatics is to provide sound and accurate representation of underlying biological mechanisms through scientific inquiry. Bioinformatics can be divided into two phases, the first phase is conversion of raw data into processed data and the second phase is performing an analysis to address the scientific inquiry. The tools employed by bioinformaticians allow for many alternative ways to process raw data based on their representation and origin. Bioinformaticians must be careful to consider how the data is processed before carrying out the goals of a scientific inquiry. Specifically there are many software tools to accomplish the data processing in gene microarray analysis, RNASeq analysis, miRNA target selection, proteomics, and other settings (Table 1). High-throughput platforms produce raw data (i.e. sequence or binding intensity) that is converted to information about the biological state (i.e. gene expression). The data processing activities form sequences called pipelines or workflow paths. Since there are often parameter settings, extra data-cleaning steps, and optional steps along the way, the multitude of workflow paths available can be quite large. Table 1 represents a very brief list of the many processing alternatives for each platform. If alternative workflow paths produce very different processed data this may confound the downstream analysis. An evaluation of the workflow path quality is a necessary to provide processed data fit for scientific inquiry. Previous evaluation of workflow paths is limited to evaluation against simulated data or evaluations of limited scope (i.e. focus on one platform). Previous workflow evaluation literature will be reviewed in Chapter 2.3.

This introduction will provide some motivating examples that different datasets are indeed produced in gene expression data processing. Furthermore, the value of using real data between platforms will be introduced as a means of an evaluation. A data-guided evaluation of workflow paths would enhance the scientific inquiry, as higher quality processed data would likely provide a more representative analysis outcome of the true biological state. For example, if a microarray based gene expression workflow path that does not consider cross-hybridization (i.e. multiple transcripts bind to the same measurement probe) gene expression values would be overestimated.

In this introductory chapter, the case will be made for a real-data driven evaluation framework of workflows in bioinformatics. In the chapters that follow I present EUFLOW (Evaluation and Utility of workFLOWs) as a software application to enable data-guided evaluation of bioinformatics workflows. In order demonstrate the need for this evaluation framework, I will define the terminology of workflows, discuss how workflow paths are selected, and present potential consequences of choosing a workflow on the processed data.

## 1.1 HIGH-THROUGHPUT BIOLOGICAL PLATFORMS

Since the advent of the central dogma of molecular biology, basic science has developed technology to measure the expression of genes and proteins. The formulation of the gene, transcript, and protein relationship continues to develop as science discovers new mechanisms of biological control. However, the current knowledge of the central dogma can help develop models for understanding the relationship between gene, transcript, protein, and functional disease states. High-throughput “-omics” platforms provide a means to measure gene expression, protein expression, as well as mechanisms of control including miRNA and methylation of DNA. The raw

data produced by these platforms must be converted from polymer sequences and/or binding intensities to data that is useful to a biologist or a clinician. Although there are many high-throughput platforms, this section will introduce five commonly utilized platforms from The Cancer Genome Atlas.[4]

**Table 1.** Brief list of bioinformatics platforms used in TCGA.[4]

Five sample platforms with data available at The Cancer Genome Atlas. Raw data is converted to final processed data using a workflow path or option. The workflow paths and options are responsible for converting the data and are discussed in detail in Section 1.1 and in Chapter 2.[5–17]

| Platforms | Raw data | Examples of workflow paths/options | Final processed data |
|---|---|---|---|
| **Gene Expression Microarray** | Signals at a probe level | Jetset<br>Plandbaffy<br>Netaffx | Expression (gene level) |
| **RNASeq** | mRNA sequence data | RNASeqV1<br>RNASeqV2<br>SALMON | Expression (gene level) |
| **Protein Expression** | Light intensity image and raw signal | MIRACLE<br>SuperCurve<br>Reno | Protein expression (gene level) |
| **miRNA** | Signals at a probe level | miRExpress<br>DeepBase | Expression (miRNA level) |
| **DNA Methylation** | Sequence data with location of the methylation functional groups | QSEA<br>DISMISS | Methylation (per CpG island) |

Measurement of gene expression is commonly performed by either probe hybridization or direct sequencing of mRNA. Prior to the development of rapid sequencing technology, oligonucleotide microarrays measured mRNA expression by printing nucleotides on a "chip" in a specific sequence. The microarray chip is then exposed to RNA samples and mRNA that is complementary to the probes would bind to the chip. Binding to the probes is measured in the form of color or light intensity. TCGA has hundreds of samples with gene expression values from the Affymetrix U133 Plus 2.0 GeneChip.[18] The Affymetrix GeneChip has 54,675 probesets, which measure about 18,000 protein coding genes. Probe binding intensity is converted to gene expression values with workflows that quantify, normalize, and filter the probesets. Raw data from the Affymetrix chip is often represented as a fold change or other transformed expression value. Table 1 lists three examples of probeset filtering, JetSet, PlandbAffy, and Netaffx. These filtering workflow options are described in greater detail in Section 2.2.

Another measurement platform for gene expression in TCGA is RNASeq, or direct sequencing of RNA from a sample. One platform is the Illumina HiSeq sequencing system.[19] After extraction of the RNA sample users of the Illumina HiSeq platform prepare the library of RNA primers, generate the clusters of RNA, and then the sequencing system can take 1.5 to 11 days to sequence the RNA.[19] The output of sequences is stored as a FASTQ file, which is the raw data for the Illumina HiSeq RNASeq platform. TCGA provides the FASTQ file to certified users as Level 1 data. This data, however, must be aligned, assembled and quantified for gene expression. Table 1 lists three methods to process the TCGA Level 1 data. TCGA provides RNASeqV1 and RNASeqV2 as processed data for public download.[9,8] The primary difference between these two platforms is that RNASeqV2 considers alignment across splice junctions[9]. However, many other

processing methods are available, such as SALMON.[10] In Section 2.2 more details on alignment, assembly, and quantification methods are presented.

TCGA also provides raw and processed data for protein expression in the form of Reverse Phase Protein Assay (RPPA).[20] The RPPA process uses serial diluted protein lysate from frozen cell pellets and prints a nitrocellulose slide. Detection of the protein expression uses validated antibodies, which enables the measurement of protein expression from the lysate with an indicator solution of avidin-biotinylated peroxidase.[20] The serial dilution intensity curve is converted into a fold-change of the protein expression from a known spike-in protein sample. Alternative methods have been developed to determine the measurement of protein expression. These methods interpret the dilution curve using different algorithms and include SuperCurve, Modified SuperCurve and NormaCurve.[11–13]

Another platform offered by TCGA is measurement of microRNA with the Illumina HiSeq 2000 platform. The processing of miRNA is similar to RNASeq with some differences in library preparation.[19] miRNA is a molecule which can bind to mRNA and can either prevent the target mRNA from binding to the ribosome during translation or destroy the mRNA by recruitment of a nuclease.[21] One processing step that is necessary prior to the analysis of miRNA is determination of a miRNA target. There are many miRNA targeting algorithms available including mirExpress and DeepBase.[14,15]

Measurement of DNA methylation enables the analysis of epigenetic data of DNA regulation. TCGA provides Level 3 data of DNA methylation with the Illumina Infinium Assay platform.[22] Methylated residues of DNA are detected in a bead assay as nucleotides are added, where the unmethylated and methylated residues have beads which are detected at a residue sequence level. This results sequence file which has the location of methylation residues for a

sample. There are alternative processing methods to determine the methylation states from the sequence file including QSEA and DISMISS.[16,17]

Each of these platforms produce data that is in a raw form, either as a sequence file or some binding intensity. This data must be processed to the expression or quantifiable value at the level of biological interpretation. In the next section the terminology of workflows will be explicitly stated.

## 1.2 TERMINOLOGY

The terminology of workflows are not consistently defined across the literature. Inconsistent terms such as pipeline, workflow, procedure, and methods have all been utilized in the literature with ambiguity.[23,24]

One workflow is a sequence of workflow components (WC). A workflow component is an individual data processing task, which is either required to obtain final processed data or will further modify the processed data. The platforms in Table 1 each have well-established examples of workflow components (discussed in detail in Chapter 2). The RNASeq platform, for example, starts as a massive sequence file (FASTQ) that must be aligned to the reference genome, assembled into exon or gene level information, and then quantified for the gene expression value. These three workflow components are required to obtain final processed data, however further workflow components may be applied to clean and normalize the gene expression values. For each of these workflow components, workflow options represent decisions on how to execute this workflow. Specific workflow options are presented in Chapter 2. Sometimes a WC is actually not necessary for obtaining final processed data, so choosing not to execute a WC is yet

another WO. The utilization of data cleaning and quality control workflow components is highly inconsistent in the literature.[25–28] Data cleaning, or filtering, is defined as a process where some of the analysis features are deleted from the data set, often guided by a quality assessment criterion. Details on filtering are provided in the publication provided in Appendix A.[2] Some workflow components may require users to select a parameter by which a threshold is selected, such as deciding the minimal number of reads in an RNASeq analysis.[29,30] The sequence of workflow options that are selected for the workflow will be defined as a workflow path. There are often many workflow paths possible for converting the same raw data to the final processed data.

In summary, a *workflow* is a sequence of *workflow components* for constructing a *workflow path*, which itself consists of *workflow options* in a sequence. To illustrate the complexity of a workflow, consider Table 2A. Table 2A demonstrates the use of this terminology in four workflow paths (WP1,WP2,WP3,WP4) with 4 workflow components (WC1,WC2,WC3,WC4) resulting in 4 different final processed datasets (FPD1,FPD2,FPD3,FPD4). For example, WP1 uses two Workflow Steps (WO1a, WO2a) to obtain the final processed data. As a specific example Table 2B describes the RNASeq workflow as a sequence of four components: 1) Alignment, 2) Assembly, 3) Quantification, and 4) Threshold. These components have many options, but for brevity Table 2B list the workflow options utilized for RNASeqV1 and RNASeqV2. The two workflow paths are presented as a sequence of these workflow options that they utilized to obtain the final processed data. Next, the challenges of choosing between workflow paths are presented.

**Table 2.** General and RNASeq specific definition of a workflow.

Table 2A presents four different workflow paths (*WP*) constructed from the available workflow options. Each *WP* results in a distinct final processed dataset (*FPD*). Each *WP* consists of choices made for each workflow components (*WC*). The *WC* have multiple workflow options (*WO*). In Table 2B two separate RNASeq workflow paths are constructed for the workflow components Alignment, Assembly, Quantification, and Threshold. The available workflow options are discussed in detail in Chapter 2. The two workflow paths have been available on TCGA as Level 2 gene expression data on thousands of cancer patient samples.[4] The workflow options in Table 2B are discussed in Section 2.2.

# A

| **One Workflow = A Sequence of Workflow Components** | | ***Workflow Paths = A Sequence of Workflow Options*** | | | |
|---|---|---|---|---|---|
| ***Workflow Component*** | ***Workflow Options*** | ***WP1*** | ***WP2*** | ***WP3*** | ***WP4*** |
| WC1 (exactly 1) | WO1a, WO1b | WO1a | WO1a | WO1b | WO1b |
| WC2 (0 or 1) | WO2a, WO2b | WO2a | WO2b | WO2a | (none) |
| WC3 (0 or more) | WO3a, WO3b | (none) | WO3a | WO3a, WO3b | WO3a, WO3b |
| WC4 (threshold) | $\theta$(WO4) | (none) | (none) | $\theta$(WO4)=0.2 | $\theta$(WO4)=0.5 |
| FINAL PROCESSED DATASETS ➔ | | FPD1 | FPD2 | FPD3 | FPD4 |

# B

| **One Workflow = A Sequence of Workflow Components** | | ***Workflow Paths = A Sequence of Workflow Options*** | |
|---|---|---|---|
| ***Workflow Component*** | ***Workflow Options*** | ***WP1*** | ***WP2*** |
| Alignment | Bowtie,BWA | BWA | Bowtie |
| Assembly | Mapslice, Samtools | Samtools | Mapsplice |
| Quantification | RSEM, RPKM | RPKM | RSEM |
| Threshold | Prune low counts | Prune lowest values | Prune lowest values |
| FINAL PROCESSED DATASETS ➔ | | RNASeqV1 FPD | RNASeqV2 FPD |

## 1.3 CHALLENGES OF CHOOSING A WORKFLOW PATH

### 1.3.1 How biologists currently choose workflow paths

Most workflows have multiple paths available to the user. This abundance of the workflow paths may be beneficial to the analyst as more choices provide flexibility. However, it is unsettling if the choice of the workflow path is arbitrary, rather than based on evidence. Given the many workflow paths available to choose from, users can approach the choice quite differently. These decisions can be related to one or more of the following categories: 1) *availability* of fully pre-processed data, 2) *quality* of fully processed data, 3) the investigator's *familiarity* of the workflow path, 4) *novelty* of the workflow path, and 5) a decision based on a *comparative* analysis in the literature.

Data that is already pre-processed is attractive to new investigators simply due to *availability*. If the data is from a team of high reputation, it is assumed that the data has been verified, curated, and has been demonstrated to be useful in other publications. Therefore, the data is assumed to be of high *quality*. In this example, the workflow path has been performed without guidance or participation of the user. Repositories of information such, as the Gene Expression Omnibus, cBio, and The Cancer Genome Atlas, contain hundreds of platforms and thousands of patient samples which describe gene expression, copy number variation, protein expression and post-translational modification.[31–33] Most of the repositories only provide *Final Processed Data (FPD)* to uncertified users, due to information security concerns. This "Ready to Go" (RTG) data is commonly downloaded and used for publication of cancer biomarker studies and other scientific inquiry due to *availability* and the *quality* of the data. It is hard to dispute the attractiveness of *availability* of data, as the investigator can proceed directly to the primary research question. We must be careful not to assume that available data is of high quality. In

order to determine quality of data, users should continuously evaluate workflow paths in the light of new evidence. For example, TCGA RNASeq data reprocessing was recently shown to alter scientific inquiry conclusions on lung histology and HER2 activation pathway status.[34]

Alternatively, many analysts prefer to start with raw data (i.e. sequence data, light intensity) and apply a *familiar* workflow path consisting of their preferred sequence of processing steps. Familiar workflows are selected based upon habit or more specifically the set of skills employed by the analysis. For example, some workflow options are only executed from command line in Unix/Linux. Often web-based platforms are easier to use, but there are many more workflow options available on R Bioconductor, Python, and Linux/UNIX based tools. R Bioconductor, is an open source environment which continues to grow in the number and quality of analysis tools to permit workflow path analysis on raw biological data.[35] However, the use of R Bioconductor, requires a moderate knowledge of the R programming language. Finally, Linux/Unix tools require the ability to execute at command line or through shell scripts. Many bioinformaticians utilize command line tools due to the ability to automate the process and decrease computational time. In addition, many bioinformatics tools are only available as a command line tool. Linux/Unix tools can also provide the ability to string together workflow path using workflow steps from different programming languages in the same environment. Given all of these factors, it is easy to see how a workflow path is often selected by the user's *familiarity* with particular methods. Users must be careful not to assume the quality of a "default" or familiar workflow path when higher quality paths are available.

Another factor to consider is the *novelty* of a workflow path or choice. Recently published methods have salience and a presumption of superiority over previous methods simply by virtue of novelty. Contrarily, users may be reluctant to explore or utilize workflow path steps

that are not reviewed by recent literature. Search engines also may contribute to choosing a novel workflow path as the default literature search parameter is “sort by date”. The *novelty* of a workflow choice is not a compelling factor when deciding on how to process the data.

Using the literature, in the form of a published *comparative* analysis, to inform a user on workflow path quality is another factor in workflow path choice. It is standard for any new workflow path or bioinformatics method to compare itself against the current workflow paths and methods. The comparison is often based upon speed, prediction of “spiked-in” data, or some other biological parameter.[36–43] The greatest disadvantage in these studies is an arbitrary selection of which workflow paths are included in the comparison. Examples of published comparative studies are presented in Chapter 2 of this dissertation.

Lastly, the ideal decision is when the user performs data-driven evaluative workflow path choice. Evaluation of workflow paths is defined as making a judgment of the quality of the workflow path. Some evaluation on workflow quality is present in the literature.[44–49] When these evaluations are performed they are often not generalizable, utilize a small sample of workflow paths, focus on a single workflow step, or they use simulated data to evaluate a workflow path. These examples will be reviewed in Chapter 2 of this dissertation.

### 1.3.2 Consequences of workflow path choice

Given the variety of choices in workflow paths, the most important question is: Does a workflow path decision make a difference in the final analysis? A robust conclusion may be provided by only a small portion of the workflow paths. Any change in a workflow option or change of a parameter setting constitutes a new workflow path. Table 2 illustrates this concept between four general workflow paths and the final processed data. Workflow paths performed on the same

data may produce equivalent final processed datasets. Equivalent data is defined as datasets which have identical features and identical values for each feature. If two (or more) sets of *final processed data* (FPD) contain equivalent representations of data, the data is *completely consistent FPD*. Very often, however, workflow paths produce non-equivalent final processed data, but the analysis outcome does not change, in other words the conclusions of the analysis would be the same regardless of workflow path choice. If two (or more) workflow paths produce final processed datasets which are not identical, but the outcome of the analysis is not altered, then they are defined as *practically consistent FPD*. Finally, the most concerning relationship is *inconsistent FPD*. Inconsistent FPD provide not only data that are not equivalent representations of the data, but also result in datasets that lead to different analysis outcome. In this section, examples of inconsistency in the FPD results directly from 1) identifier mapping, 2) identifier filtering, 3) threshold selection, 4) use of different RNASeq workflow paths, or other types of workflow path selection.

#### 1.3.2.1 Consequences of identifier mapping inconsistency

As a motivating example I will first consider workflow components in Affymetrix microarray chips. The Affymetrix U133 Plus 2.0 microarray chip has 54,675 probesets, which represent oligonucleotides on the chip designed to be complementary to the mRNA from human cells. The probeset identifiers, however, may represent one, many, or no actual mRNA in a sample. As these probesets are redundant or of variable quality in matching to the mRNA, investigators must decide how to interpret the microarray data. Workflow components to obtain the final processed data include conversion of the binding intensity to a measure of gene expression, normalization of the

expression values, quality control feature selection, and mapping of the probeset values to other platforms such as protein expression.

Identifier mapping is a workflow component, which maps between "–omics" identifiers. Mapping between an Affymetrix microarray probeset and a UNIPROT protein accession identifier for integrative bioinformatics applications is an identifier mapping workflow component. Affymetrix offers the Netaffx database to lookup what Uniprot accession matches a particular probeset[7]. However, the Netaffx database has been found to be inconsistent and redundant when mapping the identifiers[50]. Other tools to map between these identifiers includes the DAVID ontology, ENFIN Envision database, PICR application, and the BridgeDB application.[51–54]

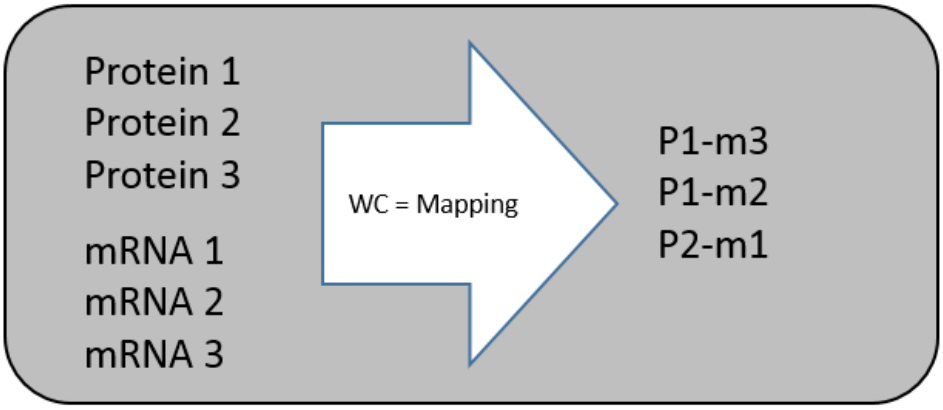


**Figure 1.** Identifier mapping illustration.

An identifier mapping tool is a workflow component (WC) that maps identifiers that are biologically connected. The output of the identifier mapping tool is a set of pairs that correspond across the platforms. In this example, three protein identifiers are mapped to the three transcript identifiers.

These identifier mapping tools resemble a library search tool for bioinformatics identifiers. Figure 1 illustrates a small identifier mapping example where a user has three protein identifiers and three mRNA identifiers from the same hypothesized gene product. Identifier mapping tools will provide a matched pair for what mRNA corresponds to a particular protein. Users may expect

that the tools agree most of the time for the same identifiers, but this is not what is observed when comparing these tools. Previously, Day et al. (2011) compared multiple tools on the same set of identifiers.[3] Figure 2 illustrates disagreement of Netaffx_Q and DAVID_Q mapping between 11,879 Uniprot accession identifiers to an Affymetrix probeset.[7,51] The horizontal axis is the number of probesets retrieved by each service and the vertical axis is a protein identifier that is entered in the query. Since the Affymetrix probeset represents a transcript that will be biologically translated to a protein (Uniprot) a high level of agreement is expected. Red and blue represent probesets uniquely mapped to a protein by NetAffx_Q and DAVID_Q, respectively. These services agree when the section is gray (see Figure 2). In fact, entirely different retrieval sets are mapped for 497 protein identifiers (Netaffx) and 809 protein identifiers (DAVID). Figure 2 also shows that for 186 Uniprot ACC the two platforms identified extra probesets that the other service did not map. Furthermore, the same exact map between transcript identifier and protein identifier occurred with only 52.5 % agreement. Further details of identifier mapping disagreement are available in Day et al (2011).[3]

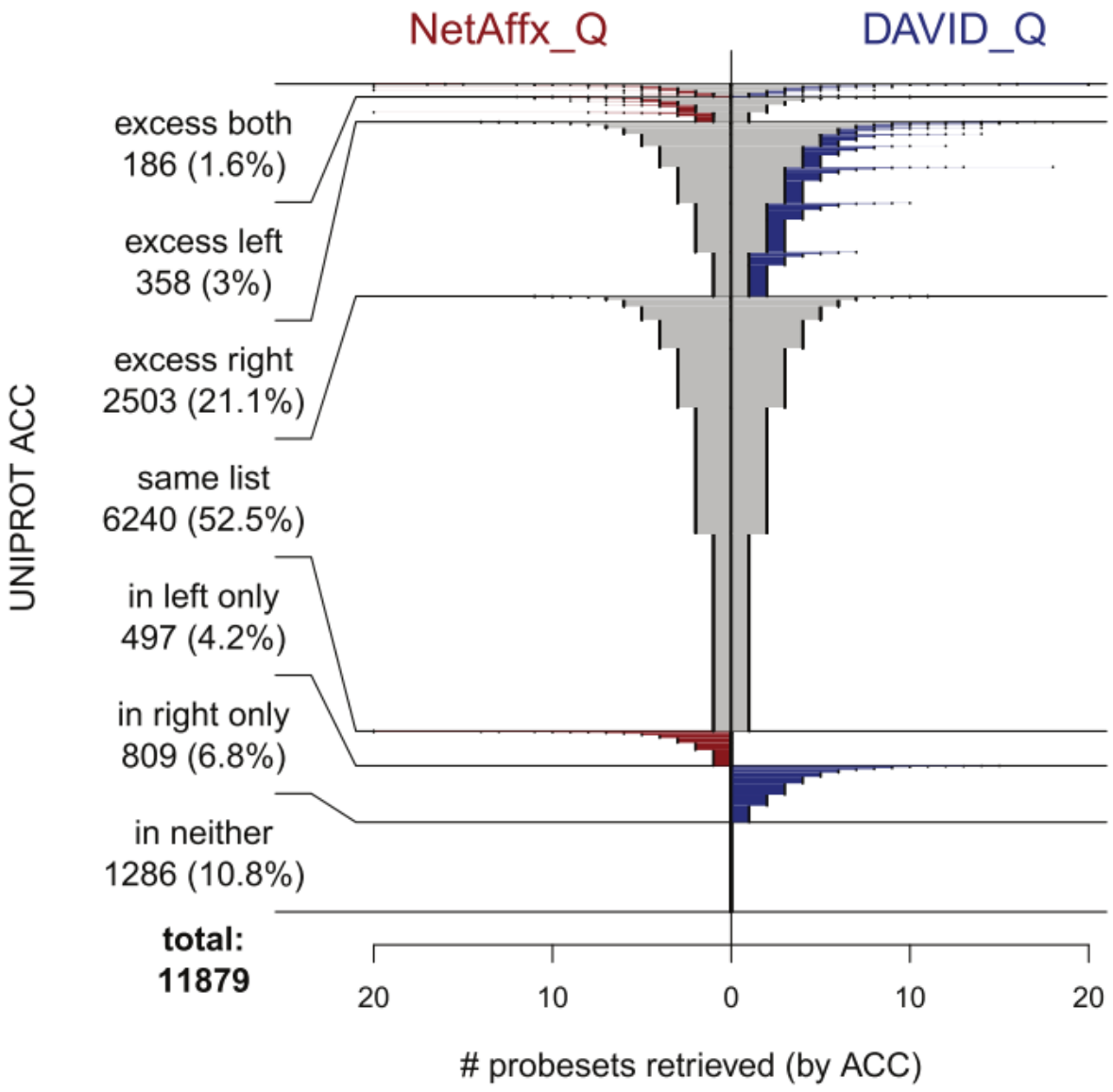


**Figure 2.** Identifier mapping disagreement.

ACC = one uniprot accession identifier, NetAffx_Q = NetAffx query probesets mapped, DAVID_Q = DAVID query probesets mapped. In Day et al. (2011), identifier mapping tools retrievals were compared for mapping between a list of 11,879 Uniprot Accessions by DAVID query and Netaffx query. Red represents probesets identified by Netaffx only, blue represents probesets identified by DAVID only, and gray represents probesets identified by both mapping tools.[3] (This figure is reproduced from Day et al. 2011 from which the authors retain copyright, including Kevin McDade)

The consequence of choosing the wrong identifier mapping tool can be meaningful. As a demonstration, Table 3 shows the three workflow options (identifier mapping tools) utilized to map 6 Affymetrix probeset identifiers to the protein ANXA2. Choosing one of these workflow options represents a simple workflow path. The workflow paths each result in a different set of mapped pairs. Which path does the user perform for the analysis? Quality of the identifier mapping tool is ideally the deciding factor. The choice could impact the results of a scientific inquiry.

If a user has, for example, protein expression data and mRNA data on the same samples, then it would be a fair expectation that coorelation between pairs would be a good guide for the quality of the identifier tool. Identifer mapping workflow paths that produce high coorelation between the pairs that they map can be considered of higher quality then workflow paths that produce less highly coorelated pairs. A complete argument for using coorelation as a quality metric is presented in Chapter 2.

For the example, in Table 3 coorelation can be utilized in deciding which path to select. In Day et al. (2011), 98 endometrial cancer samples were used to determine MS-MS spectral count and Affymetrix probe intensity.[3] Spearman correlation was determined between ANXA2 protein 6 probesets mapped by the DAVID identifier mapping tool to ANXA2. For this example, Figure 3 shows the scatterplot of unlogged mRNA expression vs spectral count for two probesets (213503_x_at and 1568126_at).[3] The criterion (Spearman rho) for the two probesets is a model quality for the workflow path. 1568126_at is identified by Netaffx and DAVID, but not by Envision. In fact, Envision does a good job of identifying the best 3 identifier pairs based on the model quality. A full analysis in Day et al. (2011) revealed that Envision was the optimal

identifier mapping tool for this dataset based on correlation as a model quality criterion.[3] Users that choose DAVID in this example may have included multiple irrelevant data points. This is an example based on one protein, but the task becomes more difficult when considering much larger datasets.

**Table 3.** Identifier mapping results for one mRNA/protein query.

Results of identifier mapping tools on six probeset features. For the mapping workflow component (WC) there are three workflow options (WO), DAVID (D), ENVISON (E), and Netaffx (N). Choosing DAVID will result in an identifier map of 6 pairs. The criterion is the Spearman rho from Day et al (2011)[3] based on endometrial cancer data. "+" = means that the service (D,E or N) reports a mapping between the feature and ANXA2, "-" means that the feature is not reported by the service.

| **WC** | **WO** | **WP1** | **WP2** | **WP3** | **Total** | | |
|---|---|---|---|---|---|---|---|
| **Mapping** | **{D,E,N}** | **D** | **E** | **N** | **Identifiers mapped** | **Features** | **Criterion** |
| | | + | - | + | 2 | 1568126_at | 0.176 |
| | | + | + | + | 3 | 201590_x_at | 0.532 |
| | | + | + | + | 3 | 210427_x_at | 0.531 |
| | | + | + | + | 3 | 213503_x_at | 0.557 |
| | | + | - | - | 1 | 210876_at | 0.321 |
| | | + | - | - | 1 | 211241_at | 0.305 |
| Final Processed Data mapped for ANXA2 | | 6 | 3 | 4 | | | |

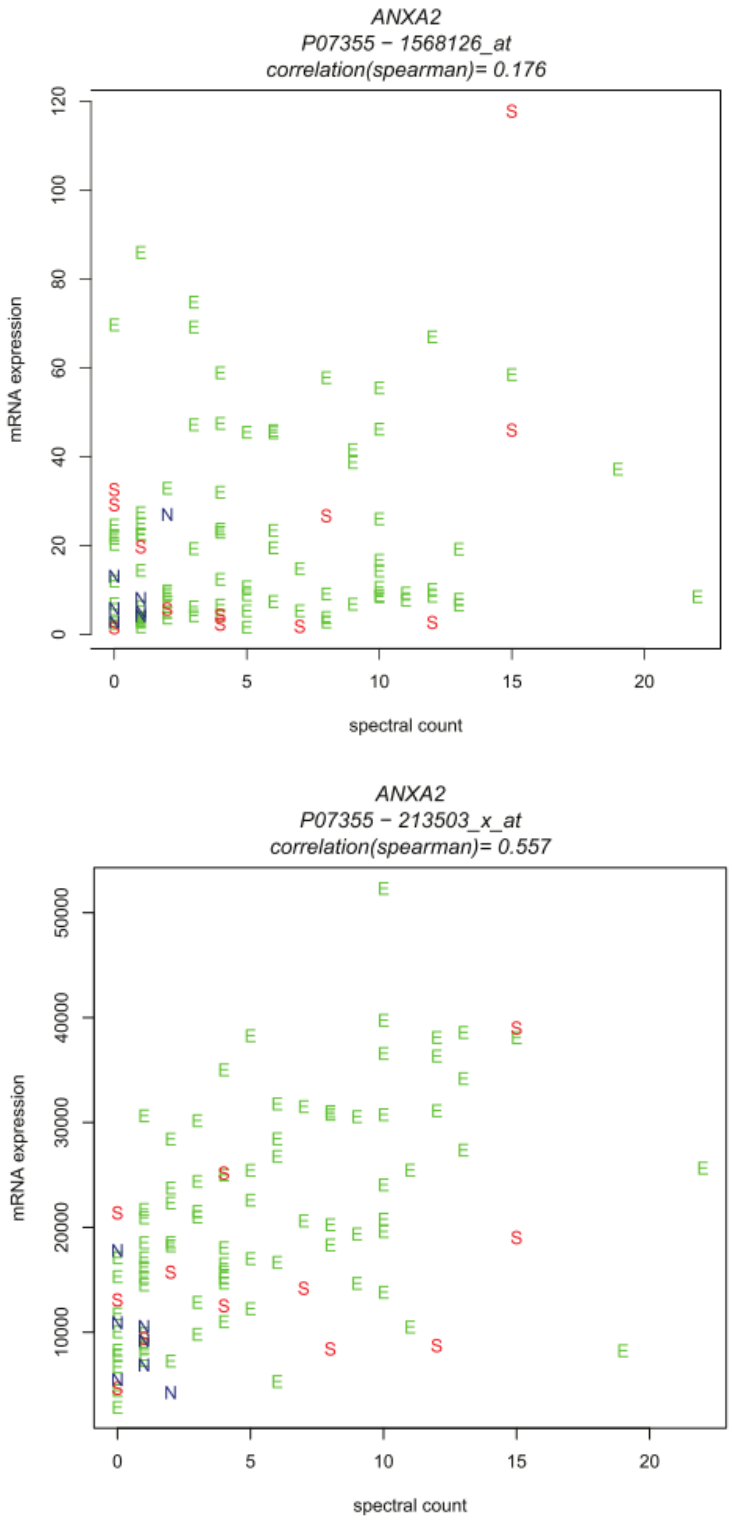


**Figure 3.** Scatter plot for ANXA2 protein and 2 probesets.

Figure from Day et al. (2011), Transcript signal (mRNA) versus Annexin 2 spectral counts (protein).[3] E= endometriod cancer, S= serous cancer, N=Normal. (Kevin McDade included in copyright)

#### 1.3.2.2 Consequences of identifier filtering inconsistency

Inconsistency also has been observed between methods to "filter" poor quality Affymetrix probesets. The Affymetrix chip has 54,000 plus "probesets", which are thought to evaluate ~18,000 human transcripts. Many methods have been developed to convert these probesets to reliable transcript abundances through normalization and probeset filtering. Some of these methods are more accepted than others, but no "best practice" has been determined. Table 4 is a short list of some identifier filtering methods used to remove redundant or inaccurate probesets from Affymetrix microarray data.[2]

Since these Affymetrix chips have been available for more than a decade, there are many approaches that have been applied to similar data, some of which show drastic differences in relative transcript abundance. In Yu et al. (2007), investigators remapped the probesets in the mouse and human U133 microarray chips through transcript alignment.[26] In remapping the probesets, they demonstrated how different methods result in different conclusions on a subset of data.

In addition, our laboratory has demonstrated the differences among probeset filtering methods and how the selection of filtering methods can have a drastic effect on the set of transcripts included in an analysis.[1] The workflow option differences were compared by arranging the number of probesets filtered and not filtered between each pair of workflow options in a set of 2X2 tables. Some notable odds ratios include: odds ratios of correlation (i.e. Jetset to Plandbaffy; OR=3.54), independence (i.e. Masker to Affytag; OR=1.08), and even inverse correlation (i.e. Plandbaffy to Affytag; OR=0.59). The odds ratio between different filtering tools imply that workflow options were inconsistent. Figure 4 shows a simple example where three mRNA probesets represent the same gene product are filtered. Filtering is not unique

to Affymetrix chips. Identifier filtering is defined as any workflow component where a raw set of identifiers are pared down to a smaller and more reliable or accurate set of identifiers. Identifier filtering can also be performed in workflows, such as proteomics, microRNA targeting and RNASeq processing.[29,55,56]

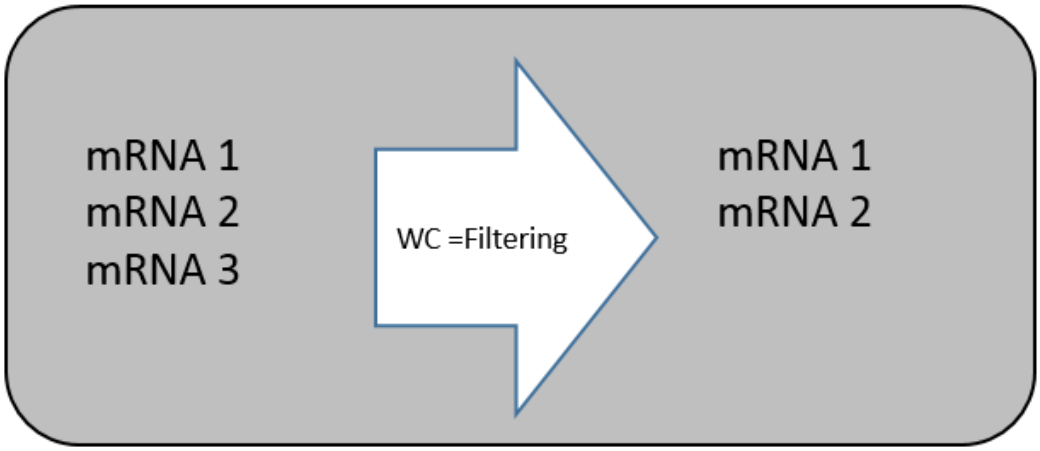


**Figure 4.** Identifier filtering illustration.

Identifier filtering is a workflow component (WC) that removes data points that are poor representations of the biological signal. In this example, mRNA 3 is removed from the final processed data.

**Table 4.** Subset of Affymetrix filtering workflow options[2].

Filter symbols, descriptions, developer criteria, and selected filtering condition for Affymetrix probeset platform.

| Filter Symbol | Description | Developer Criteria | Identifier Filtering |
|---|---|---|---|
| **AT**[57,58] | Affytag - Pre-2004 Affymetrix annotation for the Affymetrix HGU133 Plus 2.0 array | Original annotation determined by mapping to UniGene and Locus Link. "_at is considered unique. | Filter al annotation tags that begin with "_[agirxsf]_at" |
| **AG**[57,58] | Affy Grade - Netaffx Transcript Assignment Pipeline | "A" grade is the highest grade where ≥ 9 probes match transcript sequence. | Filter grades not equal to A. |
| **M**[59] | Masker - National Cancer Institute alternative chip definition file (CDF) masking out probesets with poor target location | A CDF file which eliminates a probe when more than 2 nucleotides to not match the target as well as nonspecific probes | Filter any probeset that has no remaining probes on the mask |
| **GSEN**[60] | GeneAnnot Sensitivity | The fraction of the probes in a probeset that match Watson-Crick nucleotide base pairs in the nominal gene | Filter probesets with Geneannot Sensitivity < 90% |
| **GSPE**[60] | GeneAnnot Specificity | Sum over the number of matching probes with lower weight to non-specific probes | Filter probesets with Geneannot Specificity ≤ 50% |
| **GQ**[60] | Geneannot Quality Score | A pipeline which confirms the probeset annotation with GeneCard data. | GQ= 1 is confirmed entirely with GeneCard data; Filter probesets with a GQ = [2-6] |
| **E**[61] | Encode - Encyclopedia of DNA elements | Protein coding genes are determined by human curation, RNA sequence and comparative genomics | Filter all probesets that map to a non-"Protein coding" target |
| **PD**[62] | PlandbAffy database | BLAT of target to the probe and evaluation of nucleotide mismatch or exon location | Filter all probesets with a proportion of "good" probes <30% |
| **J**[5] | Jetset Bioconductor package | Determines features such as robustness of the probe, coverage, as well as nucleotide alignment with the reference genome | Filter all except the highest-scoring probeset among those annotated for target gene. |

Using the same data as described in the previous section, the use of Spearman correlation as a criterion is also helpful.[1] In Table 5, three identifier filtering tools are utilized to remove probesets from the dataset. Identifier filters can utilize sequence information, hybridization location or probe complementarity to determine the value of the probeset.[5,6] The three workflow paths represented in this example are; Encode (E), AffyGrade(AG), and PlandbAffy(PD).[6,63,64] Each of the three identifier filters act as an interrogation on the probeset. Ideally, a probeset should bind to a complementary mRNA molecule in the exonic region and represent a protein coding gene. Encode determines the status of "protein coding" for 4 probesets. AffyGrade finds 5 probesets to have excellent sequence complementarity (see Table 4 for details). The PlandbAffy filter determines if the probes cross hybridize with an exonic region in 5 probesets.

**Table 5.** Brief identifier filtering results.

Results of identifier filtering tools on six probeset features. For the filtering workflow component (WC) there are three workflow options (WO), ENCODE (E), AffyGrade (AG), and PlandbAffy (PD). Choosing the ENCODE WP will result in a reduced final processed data of 4 probesets. The criterion is the Spearman rho from Day et al (2011).[3] "+" = the filter allows the feature, "-" = the filter disallows the feature.

| **WC** | **WO** | **WP1** | **WP2** | **WP3** | **Total** | | |
|---|---|---|---|---|---|---|---|
| **Filtering** | **{E,AG,PD}** | **E** | **AG** | **PD** | **# Not removed** | **Features** | **Criterion** |
| | | + | - | + | 2 | 1568126_at | 0.176 |
| | | + | + | + | 3 | 201590_x_at | 0.532 |
| | | + | + | + | 3 | 210427_x_at | 0.531 |
| | | + | + | + | 3 | 213503_x_at | 0.557 |
| | | - | + | + | 2 | 210876_at | 0.321 |
| | | - | + | - | 1 | 211241_at | 0.305 |
| Final Processed Data | | 4 | 5 | 5 | | | |

In the case of identifier filtering, workflow paths can be applied in Boolean combinations (i.e. an intersection of two or more resources). Figure 5 illustrates the successive application of filters E and AG upon the 6 ANXA2 probesets. Three probesets (201590_x_at, 210427_x_at, and 213503_x_at) are not removed by any of the three filters. They also have the highest Spearman correlations, suggesting that the intersection of the three filters can improve data quality. In the previous section, we saw that these are the same three probesets that Envision mapped to ANXA2 in the ID mapping example. The confluence of optimal filtering method, optimal ID mapping, and high correlations strongly reinforces the validity of our evaluation approach. Presumably, applying rigorous feature selection will raise the average quality of the features, but at the cost of reducing the dataset.

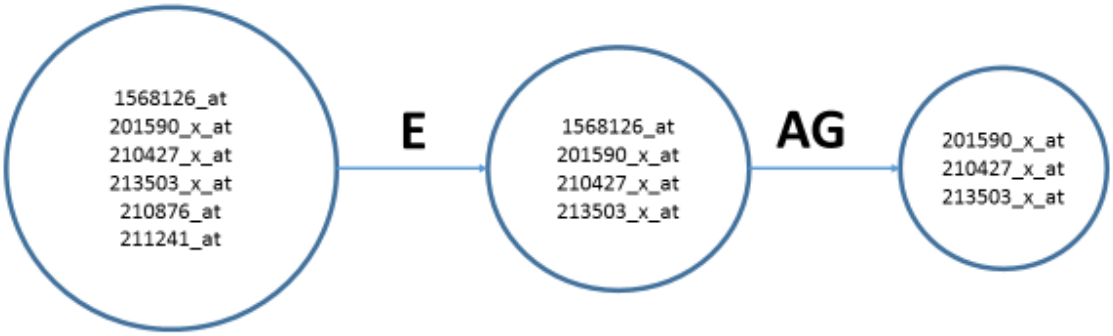


**Figure 5.** Intersection of identifier filters for ANXA2.

If the ENCODE (E) and AffyGrade (AG) filters are applied successively then only three probesets remain.

Expanding upon identifier filtering, the paper presented in Appendix A presents more examples of Affymetrix analysis inconsistency. In this analysis 9, different probeset filtering methods were utilized on the same data set, ovarian serous carcinoma data from TCGA. Each of the probeset filtering methods removed probesets from the analysis due to problems such as cross hybridization, failure to hybridize to a exonic region, and poor sequence complementary matching. A high-quality hybridization should provide a more accurate quantification of gene expression. Therefore, over all genes, those with higher accuracy gene expression values should be more strongly correlated with protein expression for that gene than those with poorer accuracy. The article in Appendix A calculates transcript-to-protein correlations, fits mixture models, and calculates expected utility for each probeset filtering method and their Boolean combinations with the EUFLOW framework.

#### 1.3.2.3 Consequences of threshold selection inconsistency

Another problem in bioinformatics is the inconsistency of threshold selection on the final processed data. Many biological platforms suffer from poor sensitivity and specificity unless some arbitrary threshold is placed on a parameter, which in effect filters features from the final processed data. Examples include selection of a fold-change threshold in microarray data, a minimum read threshold in quantification of RNASeq mapped reads, and the low dynamic range of MS/MS proteomics.[28,65,66] Analysis by Williams et al. (2016) on Drosophila neuron RNASeq gene expression has demonstrated the consequences of changing the thresholds on the scientific conclusion.[66]

A simple threshold scenario is presented in Table 6 for RNASeq read filtering. Users of RNASeq data will often select a minimum number of reads to represent an expressed transcript.

An arbitrary threshold is often selected and changing the threshold number represents a WP. Table 6 represents this scenario with thresholds of 30, 100, and 1000 of reads. If four mRNA are represented by different reads this decision will impact the final processed data.

**Table 6.** Brief threshold selection example.

Example of a threshold selection experiment. For the filtering workflow component (WC) there are three workflow options (WO), remove mRNA that is less than 30 reads, 100 reads and 1000 reads. Choosing the 30 read threshold will result in a reduced final processed data of 3 probesets. This is just an example and is not based on real data. "+" = at this threshold the filter allows the feature, "-"= at this threshold the filter disallows the feature.

| **WC** | **WO** | **WP1** | **WP2** | **WP3** | |
|---|---|---|---|---|---|
| **Threshold selection for number of reads** | **{0-10000}** | **30** | **100** | **1000** | **Features** |
| | | - | - | - | mRNA 1 (25 reads) |
| | | + | - | - | mRNA 2 (75 reads) |
| | | + | + | - | mRNA 3 (150 reads) |
| | | + | + | + | mRNA 4 (1500 reads) |
| Final Processed Data | | mRNA2<br>mRNA3<br>mRNA4 | mRNA3<br>mRNA4 | mRNA4 | |

#### 1.3.2.4 Consequences of RNASeq workflow path inconsistency

Consistency matters when choosing a workflow path to process RNASeq data. RNAseq data analysis has many workflow components such as alignment, assembly, and quantification. Some investigators have reported the impact of workflow path choice on the relative transcriptional abundance.[28,30,67] Other problems include artifacts resulting from gene fusion events, altered reads due to paralogs in the genome, and alignment with introns.[9,44,68] As some workflow options address these issues and others do not, inconsistency results among studies. In Chapter 2 the

inconsistency of RNASeq workflows will be reviewed in greater detail for each of the required workflow components.

#### 1.3.2.5 Other types of workflow inconsistency

The pipelines for variant callers in whole genome sequencing also demonstrate tremendous variation in output. Although this project will not evaluate variant callers, the phenomenon reinforces the hypothesis that methods do matter. Liu et al. (2013) evaluated 7 pipelines for the variant caller endpoint.[38] The seven pipelines differed in Ti/Tv (Transition/Transversion Ratio) and SNP counts and indels by as much as 11% between pipelines. In another recent analysis Alioto et al. (2015) compared combinations of 3 commonly utilized reference genomes, 4 alignment tools, and 2 variant callers and found the workflow paths to have drastic differences in the mutation calls[69]. They compared 19 workflow paths a curated MB.GOLD reference sequence which the authors considered to be the true positive variants.[69] The 19 workflow paths had a precision range for the variants from 0.11 to 0.99. One of the most startling results in next generation pipeline comparison comes from O'Rawe et al.[70] Comparing across five pipelines of whole genome sequencing data and 15 family exomes the investigators reported a 57.4% concordance on single nucleotide variants. This is an extremely disturbing statistic given the rapid march to personalize medicine based on patient genetics for clinical purposes.

### 1.3.3 Quality vs quantity

In the previous section identifier mapping, identifier filtering, threshold selection, and RNASeq were presented to demonstrate that differences exist in workflow paths in the quantity of the output. Also, it is important to understand that the quality of the feature pairs that remain unfiltered may be different between workflow paths. Deciding how to balance quality and quantity of the final processed data really depends upon the user's analysis goals. For example, in Affymetrix identifier filtering, when we remove probesets they are correct or incorrect. A user must decide if they are willing to sacrifice some correct probesets in order to remove more incorrect probesets. Another user might prefer to keep incorrect probesets so they do not throw away correct probesets.

Consider the following in addressing the concerns of users making a choice between quality and quantity (Figure 6). When you are choosing between two workflow paths that remove features some portion of the remaining data is shared (top bar). But when the final processed data is restricted to one workflow path, there will be some true positives and some false positives that are included in the data. Figure 6 presents an unknown choice that a user is making when selection of a particular workflow path is utilized. Choosing WP1 will produce more probeset pairs but worse average quality of data (i.e. large proportion of FP), while selecting WP2 will have more probeset pairs but better average quality (i.e. smaller proportion of FP). A user's utility for this unknown decision should be considered a vital part of any workflow evaluation.

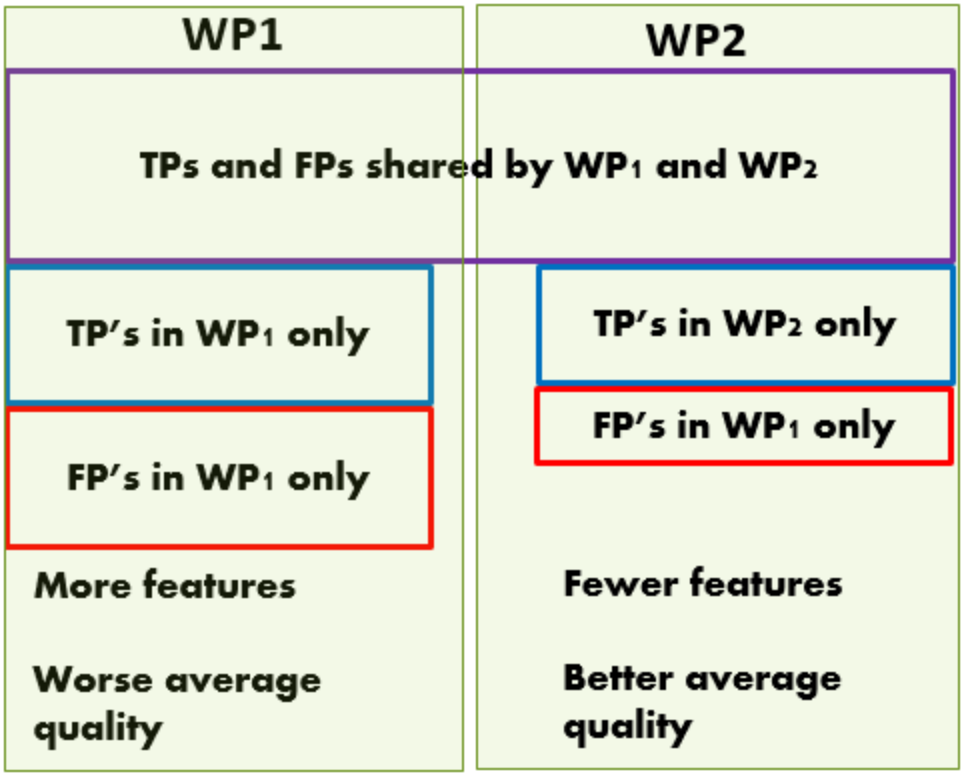


**Figure 6.** Choosing between quantity and quality.

TP = Feature pair correctly included, FP = Feature pair incorrectly included, WP1 = a hypothetical workflow path that produces more unique FP than TP but more unique features overall, WP2 = a hypothetical workflow path that produces slightly more unique TP than FP but less unique features overall.

## 1.4 THESIS

The EUFLOW decision-theoretic framework and software enables bioinformaticians to evaluate alternative workflow paths. It uses real biological data sets, expected biological relationships and user utility to guide optimal workflow path choice. EUFLOW can lead to better analyses across bioinformatics. I test EUFLOW here on gene expression data including workflows of microarray chips and RNASeq platforms.

Figure 7 describes the EUFLOW framework. EUFLOW requires as an input: 1) a set of workflow path evaluation data, 2) reference data, and 3) an identifier map between the reference and the evaluation identifiers. The set of evaluation data contains features (i.e. gene ids, probesets) as the rows and samples as the columns for all workflow paths. The data in the evaluation set is mapped, by the identifier map, to the reference data. The reference data has features (i.e. protein identifiers) as the rows and the same samples as the evaluation set as the columns. A model quality must be calculated at the level of the evaluation feature identifier to reference feature identifier pairs. The user specifies how the model quality is calculated (i.e. correlation) for each pair. The model quality represents the *expected biological relationship* between the pairs. A density of the model quality has TP and FP for every WP, an underlying mixture distribution which includes "+" component (positively correlated, biologically coupled) and a "-"negative component (incorrect or other biological decoupling) can be deconvolved using the procedure described in Chapter 3. An estimation of the posterior probability of a pair in an element of the "+" mixture can be determined using the Expectation Conditional Maximization and the Empirical Bayes Plug-in method, which are described in Chapter 3. A better workflow path will have a higher proportion of pairs that belong in the "+" mixture component. With the user specifying loss from including a false positive and gain from including a true positive EUFLOW is able to calculate the expected utility for each workflow path. The user specifies both the *model quality* criterion and the *user utility* for gain/loss of feature pairs. EUFLOW is implemented as an R package (https://github.com/Kkm5/EUFLOW.git). The EUFLOW R package was prototyped using Affymetrix probeset identifier filtering which is briefly presented in chapter 4 of this dissertation and in its entirety as Appendix A. Further

testing of EUFLOW is presented in this dissertation on TCGA-based RNASeq identifier filtering, threshold selection, and a common workflow path evaluation.

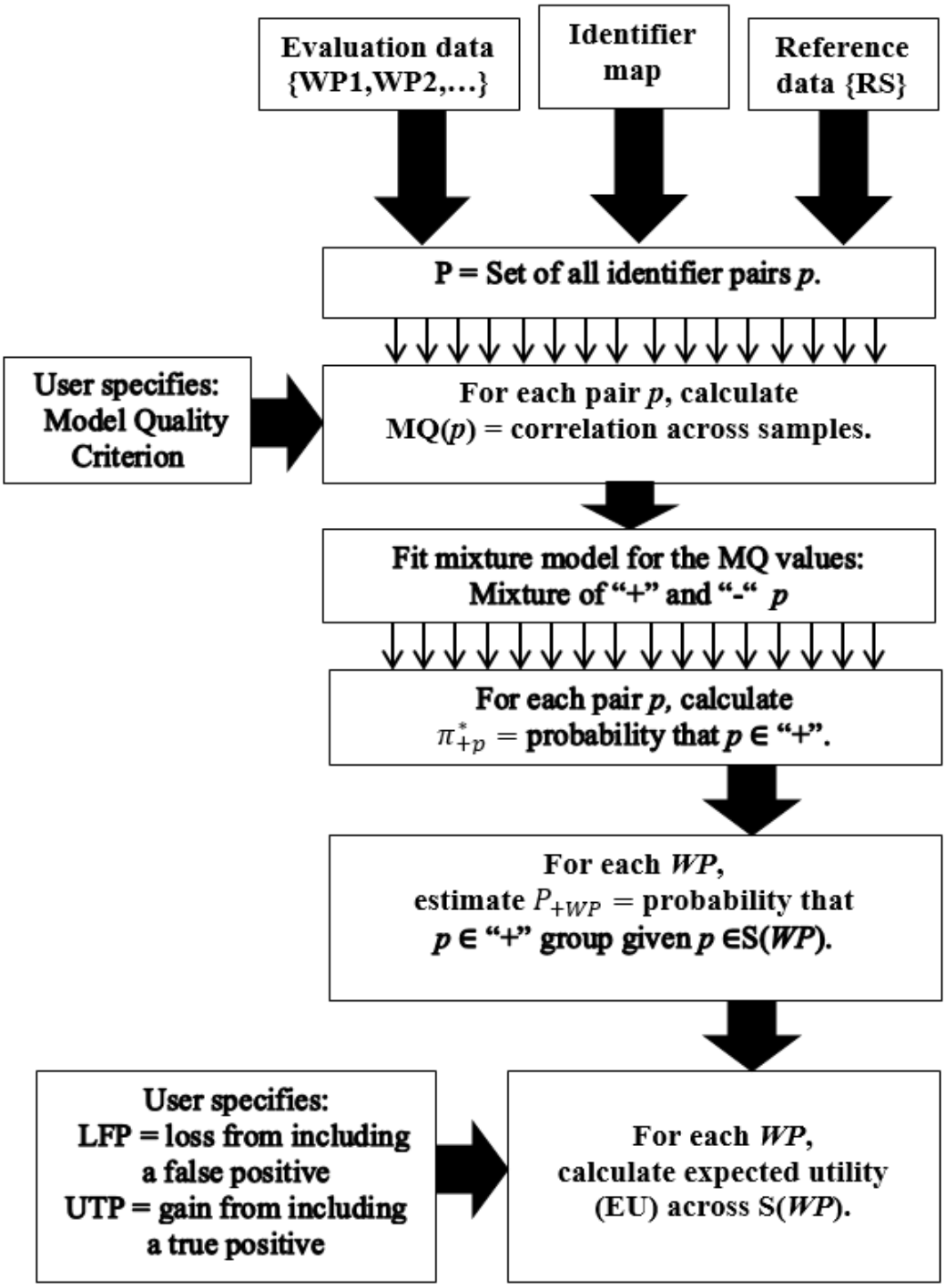


**Figure 7.** EUFLOW framework.

EUFLOW takes three input files and user specifications to determine the expected utility of a workflow path based upon real data, a relationship specified by the user between the workflow path final processed data and the reference data. A complete description of the EUFLOW framework and terms will be provided in Chapter 3.

# 2.0 BACKGROUND

An evaluation of workflows is necessary because: 1) there are so many choices available for a given bioinformatics platforms, 2) these choices differ from one another how the raw data is processed, 3) the current evaluation methodology is incomplete, and 4) real data relationships are easily extracted from the data that is processed. I will begin by discussing the proliferation of bioinformatics workflow paths (Section 2.1). I will next briefly review a selection of workflow components, steps and options available to a user of gene expression microarray and RNASeq processing (Section 2.2). These workflow steps and options can differ drastically in approach, through either computational algorithm or biological relevance. In section 2.3, I will review previous attempts to evaluate gene expression workflow paths and discuss the evaluation gap of gene expression workflows. Previous attempts to evaluate gene expression workflows are limited to simulated data or is not generalizable to other types of workflows. And finally in section 2.4, I will discuss the previous use of mRNA to protein correlation as a measure of data quality, as well as provide an argument for its value to evaluation of gene expression workflows.

## 2.1 PROLIFERATION OF WORKFLOW PATHS

Regardless of the bioinformatics workflow, new components, options and paths are continually developed and utilized. The number of workflow paths that result can be staggering. Due to the magnitude of data types and workflow paths, steps and options there is a great need to organize

the information in the form of data structures and ontologies. One ontology, EDAM, classifies over 2200 bioinformatics concepts including data, identifiers, operations, and topics.[71] The EDAM ontology can be found at http://edamontology.org/page. EDAM refers to WP, WS, and WO's as "operations". Although it provides a framework it does not support evaluation of the bioinformatics operations. A further extension of the EDAM ontology is the BioXSD data exchange format, which provides an XML format to encourage interoperability of bioinformatics web portals.[72]

In addition to structure formats and ontologies, numerous attempts have been made to streamline bioinformatics workflows. The Taverna system uses a search language (Scufl) to execute workflows (WP) through a series of atomic tasks (WS).[73] Another system similar to Taverna is BioWMS.[74] The BioWMS executes workflow and is able to document the process for reproducibility. The Galaxy system is another popular platform which permits hundreds of sequence based processing steps and is able to capture the steps in a protocol.[75] Bioextract and GenePattern 2.0 provide similar executive and reproducibility functions of workflow management tools.[76,77] A new workflow management tool for the genomics community is the Cancer Genomics Cloud by Seven Bridges (http://www.cancergenomicscloud.org). The CGC not only provides a workflow management system, but also is able to directly access the TCGA data and perform a workflow path designed by the user. Advancement has been made in development of workflow organizational structures and management systems. The CGC also has a DREAM Challenge for the quantification of known isoforms and detection gene fusions https://twitter.com/DR_E_A_M.

## 2.2 GENE EXPRESSION WORKFLOW PATHS

Gene expression is one of the most utilized and explored of bioinformatics workflows. In this section, workflow paths, steps and options for three gene expression workflow paths are reviewed: 1) Affymetrix identifier mapping, 2) Affymetrix microarray identifier filtering, and 3) RNAseq analysis workflow paths.

### 2.2.1 Affymetrix identifier mapping

Identifier mapping of Affymetrix probesets is a challenge as the probesets were designed in iterations with redundant probesets.[6,7,78,79] Furthermore, the quality of the redundant probesets can vary greatly which can create challenges in mapping the probesets to other platforms such as protein identifiers.[3] In Day and McDade (2013) we developed the mathematical framework utilized in this dissertation to evaluate four workflow paths from the options EnVison, Netaffx, DAVID, and the union of all three workflow options[1]. Our evaluation was able to utilize the framework to observe that Envision was the best workflow choice of the workflow paths evaluated.[1]

### 2.2.2 Affymetrix microarray probeset filtering

Gene expression analysis on oligonucleotide arrays (i.e. Affymetrix HG-U133 Plus 2.0) has long been impaired by the presence of multiple probesets targeting the same gene, probes of questionable quality, and annotation errors.[80–82] The U133 chip has 54,675 probesets which correspond to 18,000 protein coding genes. Available methods to pare down the feature set from

54,675 probesets to a more accurate and restricted feature set include: (1) consolidation methodologies by trimming the mean, or other outlier reduction methods (ranging from means to medians)[83–86], (2) Probeset redefinition: re-evaluating probes and redefining probesets accordingly[81,87–92], and (3) identifier filtering: removing probesets identified as "bad" based on biological features.[25,5,6] For the purpose of workflow evaluation, this section will focus on identifier filtering. Identifier filtering on Affymetrix probesets represent a motivating example for the development of EUFLOW as there are many workflow options yet no consensus after many years of development in the literature.

Examples of Affymetrix, based probeset filters are presented in Table 4 in Chapter 1. The filters described here are evaluated with the EUFLOW framework and presented in the complete published work in Appendix A. Netaffx is the most common source for identifier filtering of Affymetrix chips and features probeset information directly from the designer of the probesets. One common way to determine the quality of a probeset is to look at the probeset identifier tag. The tags specify the unique probe state, where "_at" is a high quality tag for a probe that binds to one transcript without cross hybridization.[7] Affytag (AG) removes probesets for which the Affymetrix identifier (ID) contains a qualifier; that is, the ID ends in "_[agirxsf]_at", reflecting original doubts concerning the correct and unique hybridization of the probes in each probeset, as documented by Affymetrix when the array was designed.[57,58] AffyGrade (AG), provided by the NetAffx array annotation file, is a quality grade labelled as A, B, C, R, and others. Only probesets with "A" grade were accepted, since "A" grades represent at least 9 "matching probes" to the target mRNA.[57] The NCI Masker[59] filter removes probesets omitted from the NCI "masked" chip description file (CDF). Masker was produced by the NCI Laboratory of Population Genetics. The CDF file eliminates any probes that do not have at least 24 out of 25 nucleotides match the target

GenBank transcript. In addition, it eliminates any nonspecific probes that map to a different chromosome, strand, or are part of a gene cluster that could cause cross hybridization.

Geneannot provides three different identifier filtering tools: 1) probeset sensitivity, 2) probeset specificity, and 3) a probeset quality score.[80] Geneannot Sensitivity (GSEN) is defined as the fraction of the probes in a probeset that match Watson-Crick nucleotide base pairs in the nominal gene. Geneannot Specificity (GSPE) is calculated as a sum over the number of matching probes with lower weight to non-specific probes. Geneannot quality measure (GQ) is determined from the ordinal rank assigned by Geneannot to demonstrate the confirmation of the probeset to mRNA match. A score of "1" is reported to be the "best", which demonstrates that the probes were confirmed using the GeneCard data via Entrez Gene or Ensembl.[80] The worst score is a "6", which is defined as probesets where the only information available is original Netaffx annotation.[57] The EnCode (E) filter utilizes the EnCode[93] project's determination of protein coding status of the target sequence location in the genome, to remove probesets of non-coding targets. The gene status is classified as protein coding, transcribed pseudogene, untranscribed pseudogene, lincRNA, not identified by Genecode, et cetera.[61] Only probesets with the "protein coding" Ensembl code were accepted. The Ensembl codes were matched to the Uniprot accession code present in our analysis. The PlandbAffy (PD) filter utilizes the PlandbAffy [62] database, which uses the probeset sequence and the BLAT database to align probe nucleotide sequences to the target and assign to each probe a grade reflecting alignment mismatches, alignment to other sequences risking cross-hybridization, and intronic versus exonic location. The Jetset (J) filter uses the Jetset[5] assessment, which also considers nucleotide complementarity across the probesets, but also considers splice isoform

coverage, and transcript degradation. In addition, JetSet (J) will score each probeset of a target gene and select the best probeset (of currently defined probesets) for each gene on the chip.

In summary, the workflow options for filtering poor quality probesets can depend upon sequence complementarity (i.e. Jetset, PlandbAffy, Geneannot), probeset design (AffyGrade, AffyTag), or post-hoc probeset processing (i.e. Masker, Encode). Users of one of these filtering methods may filter without evaluating the performance of the filtering workflow option.

### 2.2.3 RNASeq workflow components

RNAseq is widely used in diverse medical domains as a tool of discovery including fields such as cancer, Alzheimer's, and heart disease.[94–96] Yet more workflow paths for RNASeq analysis continue to be developed, each of which argues that some improvement has been made over the old set of standard workflow paths.

RNAseq platforms provide a newer and deeper view gene expression, by providing sequence level information. There are many workflow options to process RNAseq data. There is vigorous competition to develop the fastest, highest performing, and available workflow options in RNAseq analysis. This provides the RNAseq data analysts a diverse selection of methods to choose from. Although there are some popular workflow options in RNAseq analysis, there is no universally accepted way to determine the relative gene expression from RNAseq data. Across the field there are subtle filtering, quality control techniques, and parameter settings that make it very difficult for an analyst to choose workflow options rationally.

There are many well-documented reviews of how the analysis pipelines differ in terms of the process and the resulting gene expression values.[38,70,97,98] In addition to multiple workflow paths there are many platforms which produce unique data types and formats. The Illumina

platform is the most utilized platform but other systems, such as AB SOLiD and Ion torrent, are still considered highly reliable.[99] The RNAseq standard output is a FASTQ file which contains four lines: 1) the read name, 2) raw sequence, 3) optional note space, and 4) the quality identifier of each base. The collection of FASTQ files can contain quite a bit of memory, upwards of 1 TB. Once the FASTQ file is obtained, the next step is to align the FASTQ file to the reference genome. There are many alignment algorithms available and since this is an important focus of this topic it is reviewed in full detail (section 2.2.3.1). Alignment will result in reads stored in a SAM/BAM file. The SAM version of the file is a tab-delimited sequence storage file, BAM is a binary compressed version of the SAM format. Further steps are to assemble the transcriptome by determining which of the exons match the same transcript and determining which isoforms are reliable, thereby providing the user with regions which represent certain transcripts (section 2.2.3.2). Finally, the last stage is the quantification of the reads into some measure of relative transcriptional abundance (section 2.2.3.3).

#### 2.2.3.1 The alignment workflow component

Alignment can be divided into two basic categories, unspliced aligners and spliced aligners. The unspliced aligners will align reads which contain no (large) gaps to the reference genome. The spliced aligners are necessary when the reads map of exon-exon junctions.[98,100] Here I present a review on the functionality of the select alignment tools.

Alignment of reads depends on the ability to search a massive reference genome for each of the reads obtained from the FASTQ file. The number of reads for the Illumina HiSeq platform is typically on the order of 50-200 million reads, which are 32-100 base pairs in length.[101] The first generation alignment methods included hash table based methods such as MAQ, RMAP,

and Soap.[102–104] The hash-based methodologies suffered primary from computational time on the order of thousands of hours.[105]

One of the most widely used short read alignment algorithms is the Burrows Wheeler Alignment tool.[101] The BWA method is able to read paired ends and permits short indels and gaps. It has the advantage over the hash-based methods by using "Full text index in Minute Space " (FM-index), which keeps the memory usage low in the search tree.[106] The FM-index will compress the input index but permits fast substring queries. The Burrows-Wheeler transform methods have the greatest benefit of having performance independent of the size of the reference sequence. Other methods that have benefited from the Burrows Wheeler Transform include BarraCUDA, SOAP2, and Bowtie.[107–109] BarraCUDA uses graphic processing units and the BWA method to process the index tree on graphical processing units rather than computational cluster methods.[109] Since BarraCUDA is using the BWA method, speed is the only real difference between the two methods. SOAP2 is an updated version of the hash table only based SOAP.[108] In SOAP2, the hash table is used to search for the location of the read in the Burrows Wheeler Transform reference index. Ruiquang et al (2009) found that on one human Asian sample similar percentages of paired reads were mapped in about 4% of the time ( SOAP2 828 seconds versus SOAP 19,234 seconds).[108] Another popular hash table based method is MAQ which has developed a mapping quality score.[103] The reads are split and stored in a hash table and the authors suggest 20 to 30 read depth to reduce the false negative rate.[103] The Short Read Mapping Package or SHRiMP takes the reference genome, splits it into "q-grams" and stores the "q-grams" in a hash table.[110] These "q-grams" are then filtered and ordered by size, perfect matches, and total "edits". The "edits" can be indels or SNV's from the reference genome. In this way SHRiMP has gap tolerance unlike previously mentioned methods.[110] The Genomic Next

Generation Universal MAPPer, or GNUMAP, developed by Clement et al. , uses a probabilistic model to align more reads than the previously mentioned methods.[111] GNUMAP stores the reads in a position weight matrix (PWM) and aligns the reads via the Neddleman-Wunsch algorithm. The authors demonstrate an error rate of 1.11% compared to 4.17% using Bowtie and other hash table methods.[111] All of the above mentioned methods tend to focus on speed. Although it is important to evaluate speed of the alignment tool, biological factors such as the ability to align a splice junction, must be considered in evaluation as well. In an effort to address these issues some of the authors have updated the algorithms to address splicing issues.[108,112] However, the BWA-like alignment methods still struggle to align across splice junctions, which is a problem when aligning reads from RNA.

Spliced aligners, or "long" read mappers, attempt to solve the issue of reads that cross a splice junction. Many of these alignment tools start with Bowtie or BWA-like methods to map the non-spliced reads and then a use a new algorithm to sort through the rest of the library to map the spliced reads. This approach is called "exon-first" read mapping and includes methods such as Tophat and MAPSPLICE.[9,98] Tophat relies heavily on Bowtie to map all non-splice junction reads. This is the "exon first" portion of the algorithm.[98] Where Bowtie ignores the unmappable reads, Tophat stores these reads and then applies a "seed and expend" approach. Tophat does not assume that the mapped reads are not alternative splices (as it should not), but rather produces a list of all possible neighbors under a certain threshold distance. The "seed and extend" approach will select about 30 bp upstream from the donor and 30 bp downstream from the acceptor and attempt to align these sites to the "leftovers" in the unmapped set.[98]

Another popular spliced alignment tool is the Mapsplice package by Wang et al. (2010).[9] The Mapsplice algorithm separates what the authors refer to as "tags" (200bp) into read

segments that are mapped considering anchor sites downstream from the read segment. Mapsplice also has a quality score significance based on three components: 1) alignment quality based on direct sequence match to the reference; 2) anchor significance where shorter anchors are considered less significant; and 3) entropy which uses Shannon Entropy to determine the uniformity of the sequence obtained by RNAseq.[9]

There are other splice aligning tools which actually form the basis of the "seed and extend" approach used by Mapsplice and Tophat. QPALMA for example is similar to Tophat, but utilizes a training set of known splice sites.[113] This is acceptable for some investigators which utilize smaller genomes which have extensively cataloged the known splice junctions. However, QPALMA is not ideal for human samples where we do not yet know all possible spice junctions. Another example is ERANGE, Enhanced Read Analysis of Gene Expression. Rather than use the splice sites as a training set, ERANGE appends the known splice junctions to the reference genome.[114] Regardless ERANGE and QPALMA are limited to the definition files of the known splice junctions. Tophat and Mapsplice are not without their own problems, however, as is the case with "exon first" read mapping and retro-transposed pseudogenes. These pseudogenes were once processed mRNA that is reintroduced back into the genome in another location without the introns. Since the exon can match a pseudogene sequence, a read may be mapped to the pseudogene as if it were an exon.

Another successful alignment tool is the STAR algorithm developed by Dolbin et al. (2012). The STAR algorithm is a two phase process: 1) seed searching phase and 2) the clustering/stitching/scoring phase.[115] The seed searching phase finds the Maximal Mappable Prefix, which is the longest possible unique match to a substring. In the second phase the user defines anchor seed windows and then the un-anchored seeds are stitched together and a score is

determined from the possible combinations. To validate their results STAR, TOPHAT, MAPSPLICE were aligned the ENCODE long RNAseq dataset where STAR aligned 94% of the reads compared to TOPHAT finding 71% of the known reads.

#### 2.2.3.2 The transcriptome assembly workflow component

Once alignment/mapping is complete, the next step in an RNAseq workflow is to determine the isoforms that exist in the BAM/SAM file. Isoforms are alternative versions of mRNA molecules which are produced, for example, by different transcriptional start sites or alternative splicing. Gene expression is more granular when determined at the level of the isoform rather than the level of the gene. Assembly can be performed using two different grouping metrics: 1) a reference genome or 2) a de novo method, which uses the mRNA sequences within the data to find the isoform groups. The input to reference genome based methodologies is a BAM/SAM file. The input to a de novo method is the collection of reads as a FASTQ. Cufflinks and Scripture use the reference genome to reconstruct the transcriptome.[100,116] Scripture uses a "connectivity graph" that creates a graph from neighboring nucleotide bases. A probabilistic model will recreate a "transcript graph" and reassemble and report the isoforms based on a likelihood threshold. Cufflinks can perform both reconstruction as well as relative transcriptional abundance. Cufflinks takes a splice alignment input (BAM) from Tophat or another spliced aligner.[100] Cufflinks operates off of an application of the Dilworth Theory (1950) which states that the number of mutually compatible reads is the same as the minimal number of transcripts to explain the reads.[117] Cufflinks uses this principle to take the mutually incompatible reads left over from TOPHAT to determine a "minimum path cover".[100]

A few tools offer the reconstruction of the transcriptome without a reference genome. This type of assembly is called *de novo* assembly. One of the more effective *de novo* assembly

strategies is the Velvet algorithm.[118] Velvet uses a data structure called a de Bruijn graph which organizes k-mers into pairs, where each pair is the node, and the k-mer is displayed along the arc. Since the de Bruijn graph is hampered by the tendency of genomic repeats, which would be represented in the graph as the same chain, the Breadcrumb algorithm is employed by the developers to utilize paired reads in the FASTQ file to determine the assembly.[118] The developers of Velvet mention that it is meant for applications of short reads, which is ideal for RNAseq de novo assembly.

#### 2.2.3.3 The quantification workflow component

Once alignment and assembly into isoforms is complete, the general principle is to determine quantity of a particular transcript. If there are more reads for a particular transcript, this reflects the abundance of the transcript. I will discuss below in the quantification section the different measures of relative transcriptional abundance. The term "relative transcriptional abundance" is utilized in RNAseq analysis for the following reasons: 1) reads may mean that a gene that has twice as many reads is expressed twice as much, 2) one gene is longer and therefore more "read fragments" are available, or 3) some reads may align to paralogs in the reference genome. Some analysts may assume that 2 and 3 should occur is a constant proportion across genes and samples.

Common approaches to quantification include Cufflinks and MISO.[100,103] These two approaches use the assembly to determine the number of counts that map to full length transcripts. This count must be normalized, however, since read depth and fragment size is highly variable. The reads per kilobase of transcripts per million mapped reads (RPKM) is the standard metric in most quantification methodologies, which normalizes the feature by gene

length and total mapped reads. However, paired end reads have a dependency upon one another and therefore the metric fragments per kilobase of transcript per million reads (FPKM) accounts for this in Cufflinks.[100] One more approach to mention is the RSEM[119] algorithm, which is able to calculate abundance with a reference genome or de novo using the Expectation Maximization algorithm by learning a fragment length distribution from the data.

## 2.3 THE EVALUATION GAP

### 2.3.1 RNASeq workflow evaluation with simulated data

One evaluation of RNAseq workflow paths is through the use of the BEERS toolkit.[43] The BEERS simulator takes as an input a set of transcript models and the expected quantification of each transcript and intron in the model. The simulator then creates a FASTQ file with random alternative splice forms from each model and independently introduces nucleotide substitution and indels independently throughout the model. Based on the quantification value of the transcript (provided by user) and intron (also provided by user), a probability is used to generate the simulated reads for the FASTQ file.[43] The resulting simulated data now has a known number of reads by which RNAseq workflow options steps, such as alignment, can be evaluated.

### 2.3.2 RNASeq specific workflow evaluation

There are a few examples of integrated platforms for RNASeq analysis, but none of the platforms provide a complete evaluation of workflows. In each of these examples the BEERS toolkit is used to evaluate the workflow option. Habegger et al. (2010) have developed an integrated and modular framework to complete RNAseq processing called RSEQtools.[120] RSEQtools provides a framework to align FASTQ files from multiple platforms, convert to a mapped read format (MRF) and permit transcript assembly, quantification, aggregation/correlation, and gene fusion identification.[120] RSEQtools also provides visualization on read depth by location. RSEQtools does not, however, provide an evaluation of the workflow paths, but rather a platform to complete the analysis.[120] Another integrated RNAseq workflow tool is ArrayExpressHTS, developed by Goncalves et al. (2011). ArrayexpressHTS provides the user with options to alignment and expression options, such as Tophat, Bowtie, and BWA for alignment and Cufflinks for expression.[98,101,112,121] ArrayexpressHTS enables data quality on the sequence, but not on the complete workflow quality (i.e. alignment, assembly, quantification). Wang et al. (2011) provide a framework called RseqFlow which allows multiple workflow options on RNAseq data including two alignment algorithms, Bowtie and PerM, and three calculated options for RPKM.[122] Wang et al. (2011) evaluated the quality of three expression measures on the criteria of similarity to read length and location databases obtained from ENCODE. RSeqFlow deals with error and ambiguity in RNASeq data by providing three different metrics for gene expression. The RPKM_Uniq value eliminates multi-reads which can over-estimate the gene expression value from alternative isoforms. The RPKM_Random handles the isoforms differently by random assignment of each multi-reads. The RPKM_UM gene expression value does not consider multi-reads and eliminates all unmapped reads.[122] They

determined that the default setting RPKM_UM is the optimal expression measure in sequence similarity from ENCODE. These integrated platforms only assess the quality of the workflow options at isolated stages, rather than permit a comprehensive workflow evaluation.

A powerful new integrated platform called MAP-Rseq created by Kalari et al. (2014) permits multiple preprocessing filtering steps and complete workflow execution including alignment, assembly, and quantification.[123] MAP-Rseq also provides a quality assessment of the process based on the BEERS software to generate simulated paired end RNA sequencing data.[43] With this data the MAP-Rseq RPKM was determined to have a .87 correlation with the simulated BEERS data.

Another large scale comparison across RNAseq workflows has been performed known as the RNA-seq Genome Annotation Assessment Project (RGASP) on 26 alignment protocols.[44] The alignment protocols included multiple spliced aligners, as well as other pipelines utilized widely by users of RNAseq analysis including both versions of TOPHAT, STAR, MAPSPLICE, and a number of other workflow options.[9,98,115] Multiple criteria were utilized to compare to the BEERS simulated transcriptome including alignment yield, assembly performance, coverage, and indel detection. Drastic differences were observed between the alignment yield of the mouse transcriptome and the BEERS transcriptome with a range of 68.4% and 95.1% yield. Also, in terms of ambiguous mapping of reads, one alignment tool reported 37% of sequence reads as being ambiguous. There were great differences in coverage among the workflow options: the simulated data generated 16,554 Ensembl genes and the combined assembly included 17,800 genes. The authors hypothesize that this is due to alignment to over 1,000 pseudogenes. The RGASP results demonstrate that different RNAseq workflow options produce drastically different results.

### 2.3.3 Real data RNASeq comparison and evaluation of workflow steps

Evaluation of RNASeq is certainly a developing area in bioinformatics so complete evaluations are scant. There are many comparisons of RNASeq workflow steps and paths in the literature using real data, but RNASeq real data evaluations are often limited to individual workflow steps. The popular Tophat alignment tool, for example, was validated by a comparative study against ERANGE and Velvet+gmap on mouse real data.[98,114,118] There are also published examples of alignment tools which evaluate performance based upon computational time/memory.[37,109,124] One of these studies, Seyednasrollah et al. (2013), compared runtimes across 8 RNASeq workflow paths (referred to as pipelines) using both human and mouse real data.[37] There are also powerful methods to compare assembly and gene profiling tasks, such as baySeq and edgeR.[68,125] Kvam et al. (2012) developed a real data based comparison using a two-stage Poisson model on four assembly tools on maize RNA.[125] Considering the preceding approaches there are two major gaps that remain in the real data evaluation and comparison of RNASeq workflow paths. One gap is the small number of workflow paths due to massive amount of time and memory necessary, which explains a focus on runtime by RNASeq workflow path comparison and evaluations. The second gap is a lack of focus on quality of the workflow path data. One of the more powerful RNASeq workflow evaluation addresses both of the gaps in the previous literature is Williams et al (2017).[126] They performed an evaluation of 216 different workflow paths for monocyte RNASeq data by determining precision and recall by comparison to microarray and BeadChip assay. They found drastic differences in the workflows but the most drastic differences were observed at the quantification steps of a workflow. It must be noted, however, that referring to microarray data as "truth", which is a required assumption in the Williams et al (2017) evaluation, is a strong

assumption. Other investigators have demonstrated the problems with using microarray as a quality comparison to RNASeq data.[127,128]

## 2.4 CORRELATION AS QUALITY METRIC

Quality of the final processed data should be the guide in choosing a workflow path. Relationships between molecules is the foundation of molecular biology. Widely known concepts, such as the central dogma of molecular biology, can serve as a guide for bioinformatics data processing.[129] The use of correlation to as a Model Quality Criterion is not unique to the EUFLOW framework.[1] Correlation between RNASeq data and Affymetrix data has often been utilized to benchmark new analysis methods in molecular biology.[25,40,130] Correlation of protein and mRNA have also been utilized successfully to learn about mechanisms of cancer biology.[131–134] Protein expression methodology has also benefited from using correlation as a quality metric.[33,39,135,136] In addition, correlation can be utilized to identify or annotate other types of biological molecules such as miRNA and miRNA targets.[137–139,128] In fact, three way correlation analysis can identify the relationship between central dogma pathways.[21] Correlation between protein and mRNA expression levels has been documented to serve as a quality metric is many different applications. Correlation as a model quality criterion can guide users of workflow paths in gene expression or any other bioinformatics workflow when a relationship is expected between molecular pairs.

# 3.0 MATHEMATICAL FRAMEWORK

This section introduces a framework for workflow evaluation, which extends a previously published methodology for identifier filtering and identifier mapping. For more details, see Day and McDade 2013.[1] (In that manuscript, a *WP* is called a "method", and the letter "M" appears where herein we use *WP*.) To evaluate a workflow the user must have the following inputs:

1) **A large number of biological samples from a biological repository**

**2) Two high-throughput data sets created on different platforms**, each with a feature list of identifiers. The two data sets come from the same biological samples

**3) An identifier map which produces pairs of identifiers from the two data sets** (The main example thus far is the pairing of a transcript ID to the ID of a protein that is presumed to be its translation product, Each ID pair selects a pair of features, one from each data set)

**4) A model quality score for each feature pair p**, designated MQ(p). The MQ(p) are treated independently for modeling the mixture distribution

**5) A workflow consisting of a set of workflow components**, where each component has workflow options to select from.

**6) The workflow paths (WP) which we want to evaluate and compare**, each formed by making a set of choices selected from the available options. For each WP, the set of pairs accepted or produced by that method is designated as S(WP). In this framework, WP is a workflow path as previously defined. Membership of a pair p in the set S(WP) means that p is a version or member of that workflow path.

The *Model Quality* score in this application is the correlation of the two measurements across the biological samples. We consider the probability density of the correlation values for all pairs produced by the workflow path (*WP*) (Figure 8, black line). This density is modelled as a mixture with the following three components:

- **"+"**: The transcript feature and the protein feature are correctly identified and mapped, and they are truly *biologically coupled,* which means that transcript abundance and protein abundance are monotonically related for this pair in these samples. The blue line in Figure 8 represents the "+" component.
- **"0"**: The transcript feature and the protein feature are correctly mapped, but *biologically decoupled.* This means that the expected monotonic relationship between a transcript and a protein are not observed. There are many biological reasons for decoupling, including RNA interference by microRNA's, post-translational processing, and any other mechanism causing the protein abundance to fail to reflect the transcript abundance. The green line in Figure 8 represents the distribution of correlations for decoupled pairs, which we refer to as the "0" component.
- ***"x"***: Undesirable pairs. Either a feature was mis-identified, the mRNA/protein pair mapping was incorrect, or the data quality for the feature is poor. The red dashed line in Figure 8 represents the distribution of correlations for misidentified pairs, which we refer to as the "*x*" distribution. Pairs included in this distribution should not be assigned. These assignments may be due to incorrect actions on the part of the identifier mapping or the workflow in general.

However we cannot distinguish the "0" and "*x*" components, so in practice the mixture model fitted has only two components, and the two components "0" and "*x*" are labeled as "-".

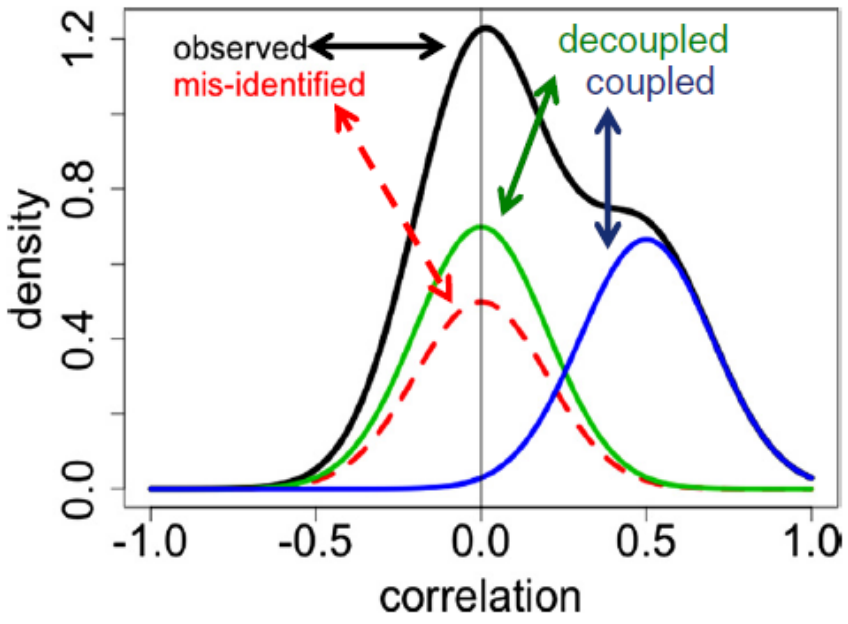


**Figure 8.** Mixture distribution example.

A hypothetical observed correlation density (black) is deconvolved into mixture components, the "+" coupled pairs ("+": blue), a decoupled pairs ("0": green), and mis-identified or poor quality pairs ("*x*"; red)[1].

### 3.1 ESTIMATION OF THE "+" POSTERIOR PROBABILITY

We would like to identify the features in either "+" or "0" for inclusion. However, the data cannot distinguish between the "0" and "*x*" groups. Under mild assumptions, the *WP* with the highest posterior probability for "+" is also the *WP* with the highest posterior probability for "+" or "0". We refer to the combined "0" and "*x*" groups as the "-" group. Even though groups "0" and "x" cannot be distinguished, basing the relative performance of workflow methods on the mixture

distributions from the observed correlations is likely to yield the correct decision; the argument for this statement is previously reported.[1]

Let $G(p)$ be the component, whether "+", "0", or "x", that pair $p$ belongs to. A better workflow path should do a better job at excluding incorrectly mapped pairs (i.e. those with $G(p)$ = "x"). Increasing the probability that $G(p)$ = "+" should reduce the $G(p)$ = "x" component. Let the proportion of pairs in group $g$ be:

$$\Pr(G(p) = g) = \pi_g \text{ for } g \in \{" + ","0","x"\}$$

The mixture model provides the opportunity to estimate $\Pr(G(p) = "+")$ for each pair $p$. This probability provides the metric we need to evaluate alternative workflow paths.

We now assume that the true correlations for all the pairs in group $g$ are distributed as a mixture of normal distributions with mean $\varphi_g$ and variance $V_g$. There is also measurement error, so the correlation of each pair $p$ in group $g$ is normally distributed with marginal mean $\varphi_g$ and marginal variance $\tau_{gp} = V_g + \sigma_p^2$, where $\sigma_p^2$ is the measurement error variance specific to pair $p$. We estimate $\sigma_p^2$ by the bootstrap method, as described in Day & McDade 2013.[1] To estimate the probability of a pair belonging to the "+" group we use an Expectation Conditional Maximization algorithm to determine the following parameters: 1) the prior probability $\pi_+$ of belonging to the "+" group, 2), the within-group true variance $V_+$ of the correlations in "+" group and 3) the within-group true variance $V_- = V_0 + V_x$ . Here "true" signifies without sampling error. This is possible since we are able to constrain the mean of the "0" and "*x*" groups to 0. This constraint defines this algorithm as Expectation Conditional Maximization (ECM). For a complete description of the ECM algorithm see Additional File 1 from Day and McDade 2013[1]. Pseudocode is available to describe the ECM in Appendix B.

Having determined the maximum likelihood estimates of the parameters, we can now calculate for each pair *p* the posterior probability of belonging to the "+" group by defining:

$$\pi^*_{+p} = \Pr(G(p) = + \mid MQ_p \text{ and parameter estimates})$$

$$\pi^*_{-p} = 1 - \pi^*_{+p} = \pi^*_{xp} + \pi^*_{0p}$$

This calculation provides the posterior probability that pair *p* belonging to the "+" component, given the correlation $MQ_p$ and its sampling variance $\sigma_p^2$, estimated from bootstrap sampling. To convert that variance into the variance of the posterior probability, the delta method approximation is used. This consists of multiplying the variance of the correlation times the square of the derivative of the posterior probability as a function of the correlation.

$$v^*_{+p} = var(\pi^*_{+p}) \cong var(MQ_p) \times \left(\frac{d\pi^*_{+p}}{dMQ_p}\right)^2.$$

The expression for the derivative is presented in Additional file 1 of Day and McDade (2013)[1]. A weighted mean of the "+" proportion provides an expected proportion of "+" group pairs for a given workflow path. The weighted mean is estimated using the posterior probabilities of each pair and the variances of these posterior probabilities.

$$P_{+WP} = \sum_{p \in S(WP)} \pi^*_{+p} (v^*_{+p})^{-1} \Big/ \sum_{p \in S(WP)} (v^*_{+p})^{-1}$$

This value provides a basis for the application of the Bayes expected loss calculation to calculate a user utility for each WP, which will be introduced in the next section.

## 3.2 EXPECTED UTILITY FOR AN ANALYSIS GOAL

It is important to consider that different analysts have different analysis goals. One workflow path may include a pair or a feature while another excludes it. The pair will be either a "true positive" of the first *WP* or a "true negative" of the second. The relative value of including a true positive versus excluding a false positive will be different for different scientific goals. Utility values can express these valuations. We utilize the Bayesian decision principle of maximizing expected utility. This principle is useful for selecting a *WP* as illustrated in the next section. The Bayesian expected loss calculation for a particular *WP* is:

$$Mean\ EU_{WP} = U_{TP}P_{+WP} - L_{FP}P_{-WP}$$

Define $U_{TP}$ as the utility of the user for a true positive pair, $L_{FP}$ as the loss of a user for including a false positive pair, $P_{+WP}$ as the expected proportion of "+" group pairs for a given workflow path and $P_{-WP}$ as the expected proportion of "-" group pairs for a given workflow path. These values enable the calculation of the Mean Expected Utility (MEU). As an alternative, the analyst may choose to use Total Expected Utility (TEU), which is the product of the number of workflow paths in the evaluation and mean expected utility. Define $n_{WP}$ as the number of workflow paths to evaluate.

$$Total\ EU_{WP} = n_{WP} \times (U_{TP}P_{+WP} - L_{FP}P_{-WP})$$

## 3.3 COMPOSITE FILTERING STRATEGIES

Boolean conjunction (intersection; “and”) and disjunction (union; “or”) operators, can create composite filtering strategies, which are easily evaluated as well. An analyst may consider whether the union or intersection of two or more filtering methods is worth the extra effort. Given a current strategy, for each so-far-unused method, one can automatically construct and evaluate the strategies formed by conjoining this method to the current strategy via conjunction or disjunction. A forward selection assesses the expected utility for each of these conjoined strategies, and chooses the one with the highest expected utility. This is referred to as “greedy” selection because it takes the immediate best step, in sequence. In contrast is the exhaustive search of every Boolean combination of the methods, which in principle could find better strategies, but is sometimes impractical. To see an example of using Boolean combinations of filtering strategies in context see Figure 16 and 17 in Appendix A.

# 4.0 EUFLOW R PACKAGE

We have developed the EUFLOW package to assist users who have multiple versions of data that arise from different workflow paths. The package can be downloaded via github at https://github.com/Kkm5/EUFLOW.git. In Chapter 4, this package will be described including the development, requirements, as well as a basic vignette of the R package.

## 4.1 DEVELOPMENT PRIOR TO PACKAGE

EUFLOW is an extension of prior projects which evaluated identifier mapping and filtering. The package depends upon some functionality of IdMappingAnalysis, which was utilized in evaluation of the identifier mapping evaluation in Section 1.2.2.1.[140] IdMappingAnalysis enables users to evaluate identifier mapping, while EUFLOW extends the evaluation to identifier filtering, threshold selection and general workflow path evaluation.[140] EUFLOW functionality was inspired by the identifier filtering evaluation performed in Appendix A[2]. In this Chapter the resources for the package development are presented as well as a vignette "How to use the EUFLOW package?"

## 4.2 RESOURCES FOR PACKAGE DEVELOPMENT

The following resources are utilized to: 1) obtain data, 2) process workflow options, 3) develop the package, and 4) version control and backup.

Data for testing and development of EUFLOW was obtained through the data portal of The Cancer Genome Atlas (TCGA).[4] TCGA has thousands of patients across 34 types of cancer on numerous platforms including DNA sequencing, miRNA sequencing, RNA sequencing, protein expression, DNA methylation, and copy number variation[4]. Levels of data are available from complete processed data (gene expression) to the raw unprocessed files (sequence files). An application process is necessary for sensitive Level 1 information, like DNA sequence data. The Cancer Genomics hub (CGHub) through the National Cancer Institute serves as the secure repository for large and protected files[4]. Although TCGA has a download matrix to obtain data files other means exist, like cBio R packages, to facilitate program based data retrieval.[32]

RStudio is the most widely utilized IDE for the R programming language. R studio can be downloaded open source at https://www.rstudio.com/. All required packages are able to be installed within one working environment and developed as an R project. The source code for the EUFLOW utilizes Git within Rstudio and all code is updated to the Github server at https://github.com/Kkm5/EUFLOW.git. The package includes the following: 1) a data folder for vignette input, 2) inst folder for input data file download outside of the package, 3) man folder for help files, 4) R folder for the source code, and 5) vignette folder for an example of package use.

R Bioconductor is also a value resource for users that process RNASeq data in R. Many of the workflow paths and options discussed in Chapter 2 are available as a package that can be obtained through R Bioconductor. There are more RNASeq workflow paths and options available through Bioconductor than any other open source platform. This creates and environment that is convenient as users can run a workflow path on raw data with R

Bioconductor packages and then evaluate the different workflow path output files with the EUFLOW package.

## 4.3 EUFLOW VIGNETTE

*Note: this a copy of the vignette in the EUFLOW package. It is a stand-alone demonstration of an evaluation of two RNASeq workflow paths and is available in the package as the file "How-to-use-the-EUFLOW-package.R"*

### 4.3.1 Introduction

The data in bioinformatics is often in some "raw" form which is not yet ready for analysis. Processing this data often involves several steps, called variously a workflow, pipeline, or protocol. EUFLOW *does not* process raw data but rather serve as an evaluation on the final processed data from alternative workflow paths.

To evaluate a workflow the user must have the following inputs:

- A large number of *biological samples* from a biological repository, such as TCGA, or a private collection of biological samples.
- Two high-throughput data sets created on different platforms, each with a feature list of identifiers; the two data sets come from the same *biological samples.*

- *EvaluationExperimentSet* is an R dataframe, which contains features from different workflow paths on the same samples. For example gene expression data from two different workflow paths (format described in the next section).
- *ReferenceSet* is a R dataframe, which contains features which correspond to the *EvaluationExperimentSet* on the same samples as the *EvaluationExperimentSet*

- An identifier map which produces pairs of identifiers from the two data sets. (The main example thus far is the pairing of a transcript ID to the ID of a protein that is presumed to be its translation product). Each ID pair selects a pair of features, one from each data set (format described in the next section).
  - *IdentifierMap* is an R dataframe, which contains a list of reference feature identifiers mapped to a comma separated list of features. The *IdentifierMap* can be constructed using the IdMappingAnalysis Bioconductor package[140]. If a user does not have an identifier map and the features identifiers are the same ( i.e. Gene Symbols for the *ReferenceSet* and the *EvaluationExperimentSet* are the same) then a EUFLOW function WorkflowPathMap enables a user to construct an *IdentifierMap*.
- A model quality score for each feature pair *p*, designated *MQ*(*p*). The *MQ*(*p*) are treated independently for modelling the mixture distribution. In applications thus far, this score is a correlation coefficient between the two features. In this vignette we will demonstrate Pearson and Spearman correlation.
  - Currently supported for correlation, person spearman, or kappa.

### 4.3.2 RNASeq evaluation demonstration

RNASeqV1 and RNASeqV2 are workflow paths that process raw FASTQ RNASeq data to obtain a measure of gene expression in TCGA data. The differences between the workflow options that are employed by these workflow paths is discussed in greater detail in Chapter 2, but the primary difference between the workflow paths is the handling of alternative splicing. The evaluation of these two workflow paths is a simple example for the illustration of EUFLOW functionality. Table 7 shows the workflow options for each of the workflow paths evaluated in this vignette. The RNASeq workflow requires the alignment of reads, the assembly into transcripts, and the quantification of the sample RNA. RNASeqV1 uses the Burrows-Wheeler algorithm for alignment and Samtools for assembly and quantification by determining the RPKM (Reads Per Kilobase Million).[124,141] The RNASeqV2 workflow path, however, using an assembly method which considers gene isoforms which determines gene expression at the gene level as fragments per kilobase million (FPKM). In this example there are 67 mRNA features considered as only 67 corresponding Reverse Phase Protein Assay (RPPA) antibodies were available in TCGA as the reference dataset. Only 66 mRNA features were considered for RNASeqV2 due to missing datafor some samples. The identifier map creates paired features of the 133mRNA/protein pairs.

**Table 7.** Demonstration of EUFLOW input.

A table for the demonstration evaluation of EUFLOW. EUFLOW does not process the workflow but compares the workflow paths on the same evaluation data. In this example RNASeqV1 and RNASeqV2 workflow paths are evaluated. The input to EUFLOW is the boxed section. The EVALUATION DATA VALUES produced from the two workflow paths form the input for the EUFLOW package as a merged R dataframe. The REFERENCE DATA VALUES are from another platform with an expected relationship. The identifier map pairs the features in the evaluation set.

| **One Workflow = A Sequence of Workflow Components** | | ***Workflow Paths = A Sequence of Workflow Options*** | |
|---|---|---|---|
| ***Workflow Component*** | ***Workflow Options*** | ***RNASeqV1 Workflow path*** | ***RNASeqV2 Workflow path*** |
| Alignment | Bowtie,BWA | BWA | Bowtie |
| Assembly | Mapslice, Samtools | Samtools | Mapsplice |
| Quantification | RSEM, RPKM | RPKM | RSEM |
| EVALUATION DATA VALUES | | RNASeqV1<br>Gene expression values<br>67 mRNA features<br>198 samples | RNASeqV2<br>Gene expression values<br>66 mRNA features<br>198 samples |
| REFERENCE DATA VALUES | | RPPA Fold change<br>67 protein features<br>198 samples | |
| IDENTIFIER MAP | | 133 mRNA (GENE SYMBOL)<br>67 proteins (GENE SYMBOL) | |

Users do not input raw data into EUFLOW but rather must process the data in R or input from outside of the R environment. A small section of the *EvaluationExperimentSet* sample data of RNASeqV1 and RNASeqV2 are represented in Figure 9.

```
data(RNASEQDATA)
RNASEQDATA[1:9,1:3]
```

```
##           X TCGA.04.1348 TCGA.04.1357
## 1  ACACA_v1        2.747         1.78
## 2   AKT1_v1       59.020        56.39
## 3   AKT2_v1       38.490        20.88
## 4   AKT3_v1        1.125         1.30
## 5  ANXA1_v1       77.479       120.54
## 6     AR_v1        0.892         1.80
## 7    BAX_v1       20.940        26.15
## 8   BCL2_v1        0.670         3.99
## 9 BCL2L1_v1       82.593        85.88
```

**Figure 9.** Sample evaluation data for EUFLOW input.

The first column (X) are the feature identifiers. Two sample identifiers are represented as gene expression.

The RNASEQDATA file represents the *EvaluationExperimentSet* for this vignette. Each feature identifier has two pieces of information, the first is the mRNA in this example separated by a "_" character and the workflow path identifier. For example in the first row, "ACACA" is the mRNA for acetyl-CoA carboxylase alpha and "v1" is the designation for RNASeqV1. Each new workflow path dataset must be appended to the *EvaluationExperimentSet* with new workflow path tags.

The reference dataset in this demonstration will use the TCGA RPPA protein expression data on the samples.[20] The same 198 samples are represented in the same order. In our example the Level 3 data for gene expression and protein expression uses the same identifier. For example in the RPPADATA.original data file, the "ACACA" represents the protein expression for acetyl-CoA carboxylase alpha. So in this example two pairs exist for acetyl-CoA carboxylase alpha, ACACA_v1/ACACA and ACACA_v2/ACACA. Users that would like to evaluate complex identifier maps should utilize the Bioconductor package IdMappingAnalysis before workflow path evaluation in EUFLOW.[140] To see an example of a complex identifier mapping evaluation please see Day and McDade (2013).[1] Figure 10 is an example output for the reference data as an input to

EUFLOW. It is required that the same sample identifier be utilized to label the sample columns (i.e. TCGA.04.1348 must be in both the reference and the evaluation data files).

```
data(RPPADATA.original)
RPPADATA<-RPPADATA.original
RPPADATA[1:9,1:3]
```

```
##        X TCGA.04.1348 TCGA.04.1357
## 1  ACACA       0.1370      -1.8782
## 2   AKT1       0.1644       0.8931
## 3   AKT2       0.1644       0.8931
## 4   AKT3       0.1644       0.8931
## 5  ANXA1      -0.1690       0.0967
## 6     AR      -0.3593       0.2772
## 7    BAX       0.0118       0.7261
## 8   BCL2      -0.7044       1.3982
## 9 BCL2L1       0.3587       1.7334
```

**Figure 10.** Sample reference data for EUFLOW input.

An example of EUFLOW reference data. The first column (X) are the identifiers for reference features that are mapped to the evaluation data. The next two columns represent the RPPA fold change values for two sample identifiers.

Now that we have the data for our example, the WorkflowPathData function will modify the separate dataframes into one data structure to prepare to calculate the model quality and perform the evaluation (Figure 11). The first item in the list is the reference data and the second item in the list is the evaluation data.

```
Workflow.Path.Data<-WorkflowPathData(EvaluationExperimentSet,ReferenceSet)
Workflow.Path.Data[[1]][1:9,1:3]
```

```
##        Symbol TCGA.04.1348 TCGA.04.1357
## ACACA   ACACA       0.1370      -1.8782
## AKT1     AKT1       0.1644       0.8931
## AKT2     AKT2       0.1644       0.8931
## AKT3     AKT3       0.1644       0.8931
## ANXA1   ANXA1      -0.1690       0.0967
## AR         AR      -0.3593       0.2772
## BAX       BAX       0.0118       0.7261
## BCL2     BCL2      -0.7044       1.3982
## BCL2L1 BCL2L1       0.3587       1.7334
```

**Figure 11.** Workflow data structure for EUFLOW.

Workflow.Path.Data has the first indexed list as the reference data and each indexed item that follows as a workflow path from the evaluation data set.

New labels are assigned using the BuildEvaluationStructure function to create a data structure that can sort by reference identifiers and evaluation identifiers. The user determines the tags based upon the workflow path. The selection of the tags are for distinguishing between workflow paths and will be used in the output to present the workflow path decision metrics provided by EUFLOW.

```
Evaluation.Structure<-BuildEvaluationStructure(Workflow.Path.Data,EvaluationTag=c("RN
ASeqv1","RNASEQv2"))
Evaluation.Structure[1:9,1:3]
```

```
##                   Symbol TCGA.04.1348 TCGA.04.1357
## ACACA_reference    ACACA       0.1370      -1.8782
## AKT1_reference      AKT1       0.1644       0.8931
## AKT2_reference      AKT2       0.1644       0.8931
## AKT3_reference      AKT3       0.1644       0.8931
## ANXA1_reference    ANXA1      -0.1690       0.0967
## AR_reference          AR      -0.3593       0.2772
## BAX_reference        BAX       0.0118       0.7261
## BCL2_reference      BCL2      -0.7044       1.3982
## BCL2L1_reference BCL2L1       0.3587       1.7334
```

**Figure 12.** Evaluation dataframe.

Dataframe with row names tagged by the user to distinguish workflow paths in EUFLOW output.

The function WorkflowPathModelQuality creates a map between the reference ids and the evaluation ids. The Path.Model.Quality object contains all of the pairs across the two platforms.

```
WorkflowPathMap(Evaluation.Structure)
```

```
##              reference                                workflow_paths_combined
## 1      ACACA_reference         ACACA_path_RNASeqv1_1,ACACA_path_RNASeqv2_2
## 2       AKT1_reference           AKT1_path_RNASeqv1_1,AKT1_path_RNASeqv2_2
## 3       AKT2_reference           AKT2_path_RNASeqv1_1,AKT2_path_RNASeqv2_2
## 4       AKT3_reference           AKT3_path_RNASeqv1_1,AKT3_path_RNASeqv2_2
## 5      ANXA1_reference         ANXA1_path_RNASeqv1_1,ANXA1_path_RNASeqv2_2
## 6         AR_reference               AR_path_RNASeqv1_1,AR_path_RNASeqv2_2
## 7        BAX_reference             BAX_path_RNASeqv1_1,BAX_path_RNASeqv2_2
## 8       BCL2_reference           BCL2_path_RNASeqv1_1,BCL2_path_RNASeqv2_2
## 9     BCL2L1_reference       BCL2L1_path_RNASeqv1_1,BCL2L1_path_RNASeqv2_2
```

```
Path.Model.Quality<-WorkflowPathModelQuality(Evaluation.Structure)
```

**Figure 13.** Workflow identifier map.

Reference column contains the reference identifier. The workflow_paths_combined column contains a comma separated value list of evaluation features.

Next, using the function ModelQualityPairs on the object Path.Model.Quality the user can determine the appropriate model quality for this evaluation. Model.Quality.Values is an dataframe which contains the model quality values for each of the pairs. How the values are determined is specified by the user. In this example Pearson correlations are calculated for each Reference-Evaluation pair across all samples.

```
Model.Quality.Values<-ModelQualityPairs(Path.Model.Quality,method="pearson")
head(as.data.frame(Model.Quality.Values))
```

```
##         reference workflow_paths_combined pearson
## 1 ACACA_reference   ACACA_path_RNASeqv1_1  0.5491
## 2 ACACA_reference   ACACA_path_RNASEQv2_2  0.5546
## 3  AKT1_reference    AKT1_path_RNASeqv1_1  0.6291
## 4  AKT1_reference    AKT1_path_RNASEQv2_2  0.5957
## 5  AKT2_reference    AKT2_path_RNASeqv1_1 -0.0787
## 6  AKT2_reference    AKT2_path_RNASEQv2_2 -0.0603
```

**Figure 14.** Pearson model quality values from EUFLOW.

Other model quality values can be calculated using the "method" argument in the function. Spearman r values are calculated in this example.

```
Model.Quality.Values<-ModelQualityPairs(Path.Model.Quality,method="spearman")
head(as.data.frame(Model.Quality.Values))
```

```
##         reference workflow_paths_combined spearman
## 1 ACACA_reference   ACACA_path_RNASeqv1_1   0.5389
## 2 ACACA_reference   ACACA_path_RNASEQv2_2   0.5645
## 3  AKT1_reference    AKT1_path_RNASeqv1_1   0.5293
## 4  AKT1_reference    AKT1_path_RNASEQv2_2   0.5478
## 5  AKT2_reference    AKT2_path_RNASeqv1_1  -0.0465
## 6  AKT2_reference    AKT2_path_RNASEQv2_2  -0.0452
```

**Figure 15.** Spearman model quality values from EUFLOW.

Next the correlation values and the reference-evaluation pairs are the input to the EstimatePosteriorProbability function. The first step of this function is to apply a bootstrapping procedure to obtain a resampled standard deviation and bias of the model quality values. Next the vector of correlations, the variance, and the bias are the input to the EM procedure to estimate the posterior probability and posterior probability variance of belonging to the "+" component. Figure

9 is the output of the EstimatePosteriorProbability function. It is a mixture distribution is estimated that has 2 components where one component represents the "+" component and the 0 centered component represents the "-" and "0" component. The dataframe Posterior.Probability has a column for the posterior probability and variance of the posterior probability of the “+” component.

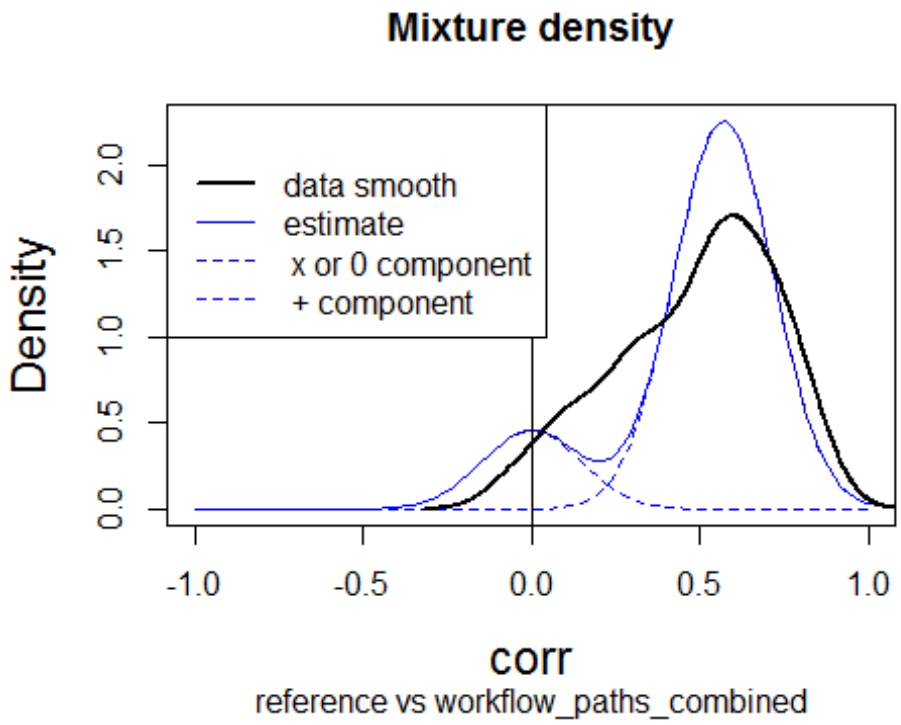


**Figure 16.** Mixture distribution for vignette.

A sample of the mixture distribution plot obtained by running the *EstimatePosteriorProbability* function in EUFLOW. The black data smooth line is the empirical correlation density, the solid blue line is the mixture fit estimate, and the dashed lines represent the two predicted components where the “x or 0" component has a 0 centered mean.

```
head(Posterior.Probability)
```

```
##         reference workflow_paths_combined postProbs postProbVar    corr
## 1 ACACA_reference   ACACA_path_RNASeqv1_1  0.999820    7.92e-08  0.5491
## 2 ACACA_reference   ACACA_path_RNASEQv2_2  0.999731    2.29e-07  0.5546
## 3  AKT1_reference    AKT1_path_RNASeqv1_1  0.999946    9.87e-09  0.6291
## 4  AKT1_reference    AKT1_path_RNASEQv2_2  0.999909    2.57e-08  0.5957
## 5  AKT2_reference    AKT2_path_RNASeqv1_1  0.000144    3.27e-08 -0.0787
## 6  AKT2_reference    AKT2_path_RNASEQv2_2  0.000301    1.66e-07 -0.0603
##       sd      bias
## 1 0.0588 -0.001878
## 2 0.0723  0.002329
## 3 0.0766 -0.002303
## 4 0.0708 -0.011713
## 5 0.0438 -0.000379
## 6 0.0483  0.002412
```

**Figure 17.** Posterior probability output from EUFLOW.

For each feature pair the posterior probability of belonging to the "+" component (postProbs), the variance of that probability (postProbVar), model quality (corr), the standard deviation (sd) of the model quality, and the bias (bias) of the model quality are calculated.

A user can now input values for the Utility of a true positive (UTP) and Loss of a false positive (LFP) for the estimation of the Expected Utility of each method. Evaluation.table will take the input value of Posterior.dataframe to calculate the following values: 1) nPairs, 2) PrPlus, 3) PrTrue, 4) PrFalse, 5) Utrue, 6) Lfalse, 7) Eutility1, and 8) Eutility.

The number of pairs used in the workflow path is defined as nPairs. PrPlus is $P_{+WP}$ from Chapter 3 and is the optimally weighted mean of the proportion of pairs that are estimated to be in the + component for each workflow path. PrTrue is equal to $P_{+WP}$ unless a deltaPlus factor is applied by the user. As defined previously, the deltaPlus is proportion of valid pairs that the user expects to be in the + component. Prfalse is $P_{-WP} = 1 - P_{+WP}$, Utrue is calculated as the product $U_{TP}P_{+WP}$ which is the first part of the Bayesian expected loss calculation. LFalse is the second part of the Bayesian expected loss $L_{FP}P_{-WP}$. Eutility1 is the complete Bayes expected loss and represents the Mean Expected Utility $U_{TP}P_{+WP} - L_{FP}P_{-WP}$. Eutility is the product of the nPairs and the Bayes expected loss and represents the TEU, $n_{WP} \times (U_{TP}P_{+WP} - L_{FP}P_{-WP})$.

In the sample data, RNASeqv1 has 67 gene expression features mapped to 67 RPPA protein expression features and nPairs in the calculation below represents the number of pairs with model quality scores determined from correlation. A PrPlus value of 0.945 was estimated using the unknowns estimated from the ECM, and this value is the optimally weighed mean of the proportion of pairs belonging to the + component for RNASeqv1 pairs. PrTrue is equal to the PrPlus as the deltaPlus parameter was set at 1 for this example. The value of 0.0546 for PrFalse is simply 1-PrTrue and is the optimally weighted mean of the proportion of pairs belonging to the "-" component. Provided these values for the data the user specified values are now used to determine the Bayes expected loss as UTP =1 and LFP = 1, which in this example simplifies the Bayes expected loss to PrTrue – PrFalse and for RNASeqv1 is 0.891. And finally the Total Expected Utility is the product of nPairs and Eutility1. For this worked example RNASeqv2 has the maximum value of Mean Expected Utility and Total Expected Utility and is the suggested workflow path using default EUFLOW parameters.

```
Evaluation.table<-WorkflowEvaluationTable(Posterior.Probability)
Evaluation.table
```

```
##          nPairs PrPlus PrTrue PrFalse Utrue Lfalse Eutility1 Eutility
## RNASeqv1     67  0.945  0.945  0.0546 0.945 0.0546     0.891     59.7
## RNASeqv2     66  0.965  0.965  0.0347 0.965 0.0347     0.931     61.4
```

**Figure 18.** EUFLOW Evaluation table.

For each workflow path the number of feature pairs (nPairs), proportion of pairs belonging to the "+" component, proportion of feature pairs belonging to the "+" component with delta factor (PrTrue), proportion of feature pairs belonging to the "-" component, the utility portion of the Bayes expected loss (UTrue), the loss portion of the Bayes Expected loss (Lfalse), the mean expected utility (Eutility1), and the Total Expected Utility (Eutility).

# 5.0 EUFLOW RNASEQ EVALUATION EXPERIMENTS

In Chapter 2, many RNASeq workflow paths were reviewed to demonstrate the proliferation and systematic differences of workflow paths. In this section, usage of EUFLOW (https://github.com/Kkm5/EUFLOW.git) is demonstrated through evaluations of several RNASeq workflows, each with a single workflow component. These include: an identifier filtering workflow component, a threshold selection workflow component, and a workflow component consisting of an entire RNASeq pipeline. An evaluation can be performed at a quantifiable stopping point which has some model quality criterion to the reference data. For simplicity the evaluations performed in Chapter 5 end at the final processed data of gene expression values. All of these evaluations are available within the vignette "How_to use_the_EUFLOW_package.Rmd"

## 5.1 DATA FOR EVALUATION

The evaluations performed to demonstrate EUFLOW include breast invasive carcinoma (BRCA) (406 total samples) and ovarian serous carcinoma (OV) (198 total samples). The BRCA and OV sample pools were used for this evaluation example due to: 1) the high number of available samples, 2) the availability of Illumina HiSeq RNASeq data, and 3) the availability of protein expression data. Matched samples, which have both RNAseq and RPPA data are utilized. If data

was not available for either the transcript or the protein then it was excluded from the evaluation. The following evaluations will be performed 1) RNASeq TCGA BRCA identifier filtering, 2) RNASeq TCGA BRCA Threshold selection, and RNASeq TCGA OV common workflow path evaluation.

The IlluminaHiSeq_RNASeqV1 and IlluminaHiSeq_RNASeqV2 data were downloaded using the cBio R package and the TCGA data portal, respectively. These "ready to go" datasets are the final processed datasets of curated workflow paths. RNASeqV1 is processed by the workflow path developed by Li et al. (2010).[8] RNASeqV2 is processed using RSEM and MapSplice developed by Wang et al (2010).[9] Another workflow path included in the evaluation is the TCGA BRCA data set. This data was processed using the SALMON workflow path.[10] The workflow path was executed by David Boone, PhD at the Department of Biomedical Informatics, School of Medicine, University of Pittsburgh.

The Reference Set file is the protein expression Reverse Phase Protein Assay (RPPA) fold change data.[20] This dataset was downloaded using the cdgsr Bioconductor package.[35] Since the quality of RPPA data is highly dependent upon the binding of antibody only validated antibody status is included in this analysis. Appendix B.2 has the antibody list, which are classified as high quality using the MDAnderson standard antibody list. This list was produced using a procedure similar to Tibes et al. (2006) validation of RPPA antibodies.[20]

The results for these evaluations are presented in Section 5.2 (Identifier filtering), Section 5.3 (Threshold selection) and Section 5.4 (Common workflow paths).

## 5.2 EVALUATING AND COMPARING RNASEQ IDENTIFIER FILTERS

In bioinformatics workflows, features that do not meet certain biological standards can be removed from the analysis. However, this practice is inconsistent and can hinder meta-analysis and clinical implementation. A simple evaluation removing feature pairs which have transmembrane and low complexity regions is presented below.

### 5.2.1 Identifier filtering workflow paths

The identifier filtering evaluation of RNASeq workflow paths utilized the Salmon version of the popular Sailfish workflow path on TCGA breast cancer data of 406 samples.[10] The BRCA RPPA fold change data was obtained through the TCGA data warehouse on the same 406 samples. For simplicity, the identifier map is limited to an identity relation in which the transcript and the protein use the same HGNC identifiers in Appendix B.2. The features produced by the workflow paths and the reference data are already in the format of a HGNC identifier so the id map was simplified in this evaluation. Biomart was utilized to search the ENCODE database for 62 feature pairs and the classification of the TMHMM algorithm and the SEG complexity[142,143]. These filters were selected due to the biological impact of transmembrane protein and low complexity regions in an RNASeq workflow. Transmembrane proteins are difficult to measure due to the loss of stabilization of the phospholipid membrane in the structure. Low complexity regions impact the identification of protein coding regions[144]. Considering these groups the three workflow paths for this evaluation are 1) No filtering, 2) filter transmembrane feature pairs, and 3) filter high complexity feature pairs.

**Table 8.** SALMON RNASeq filtered workflow component evaluation workflow paths.

A table for the evaluation of a filtering workflow component applied to SALMON processed RNASeq gene count data. The final quantified gene counts are then filtered based upon two filtering categories to remove features that are transmembrane (TM.) and high complexity (HC) according to the TMHMM and SEG algorithm from the BioMart database. The boxed section is the input for EUFLOW.

| **Workflow** | | ***Workflow Paths*** | | |
|---|---|---|---|---|
| ***Workflow Component*** | ***Workflow Options*** | ***SALMON No filter*** | ***SALMON Filter TM*** | ***SALMON Filter HC*** |
| Alignment | Read –free alignment | Read –free alignment | Read –free alignment | Read –free alignment |
| Assembly | SALMON | SALMON | SALMON | SALMON |
| Quantification | SALMON | SALMON | SALMON | SALMON |
| Filter | TMHMM,SEG | None<br>All 62 mRNA features remain | TMHMM = FALSE<br>48 mRNA features remain | SEG =FALSE<br>50 mRNA features remain |
| EVALUATION DATA VALUES | | SALMON<br>Gene count values<br>62 mRNA<br>406 samples | SALMON<br>Gene count values<br>48 mRNA<br>406 samples | SALMON<br>Gene count values<br>50 mRNA<br>406 samples |
| REFERENCE DATA VALUES | | 62 RPPA fold change protein features | | |
| IDENTIFIER MAP | | 62 mRNA (GENE SYMBOL)<br>62 proteins (GENE SYMBOL) | | |

### 5.2.2 Identifier filtering model quality

The input files to EUFLOW includes 1) a 62 X 407 RPPA fold change data file where the rows are the 62 features and the first column of the file contains the Hgnc_symbols followed by 406 TCGA sample names, 2) a 186 X 407 RNASeq feature gene count data file where the rows are the features with a tag (_v1, _v2, _v3, _v4, _v5) for features which meet the identifier filtering groups in Table 8. The identifier map in this example is the matched gene identifiers (i.e ANXA2_reference mapped to ANXA2_v1). Model quality is determined via Pearson correlation for each of the pairs. The density for the Pearson correlations is in Figure 10 (black line). Using a bootstrap procedure variance and bias are estimated and the mixture distribution is deconvolved via Expectation-Maximization, the resulting mixture components are represented by the dotted blue lines in Figure 19, and the resulting posterior probability is for each pair.

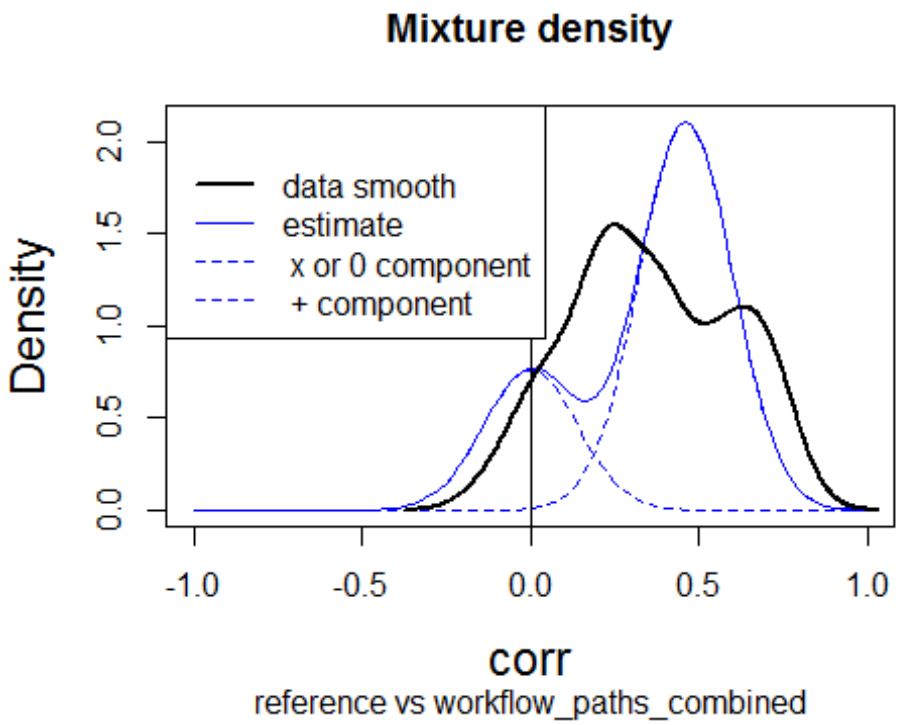


**Figure 19.** Mixture distribution for identifier filtering of RNASeq breast cancer data.

Mixture distribution plot obtained by running the *EstimatePosteriorProbability* function in EUFLOW on the model quality parameters from the RNASeq/RPPA pairs. The black data smooth line is the empirical correlation density, the solid blue line is the mixture fit estimate, and the dashed lines represent the two predicted components.

### 5.2.3 Identifier filtering expected utility

Given the default parameters (UTP = 1, LFP = 1, and delta = 1), mean expected utility (MEU) and total expected utility (TEU) were calculated using the WorkflowEvaluationTable from the EUFLOW package. Each of the filter feature sets and the unfiltered set are provided in Table 9. Each filter represents a workflow path. Key observations include the PrPlus of the non-transmembrane feature pairs of 0.962. This is consistent with non-transmembrane protein expression reliability, as transmembrane proteins mis-folding in the absence of phospholipid

membrane stabilization. Considering equal treatment of false positives and false negatives the optimal workflow path with a MEU 0.925 is NTM. However, if TEU is the criterion then not filtering at all is the optimal workflow path (TEU=50.87).

**Table 9.** Utility table for RNASeq identifier filtering example.

The number of pairs used in the workflow path is defined as nPairs. PrPlus is $\boldsymbol{P_{+WP}}$ from Chapter 3 and is the optimally weighted mean of the proportion of pairs that are estimated to be in the + component for each workflow path. PrTrue is equal to $\boldsymbol{P_{+WP}}$ unless a deltaPlus factor is applied by the user. As defined previously, the deltaPlus is proportion of valid pairs that the user expects to be in the + component. Prfalse is $\boldsymbol{P_{-WP}} = \mathbf{1} - \boldsymbol{P_{+WP}}$, Utrue is calculated as the product $\boldsymbol{U_{TP}P_{+WP}}$ which is the first part of the Bayesian expected loss calculation. LFalse is the second part of the Bayesian expected loss $\boldsymbol{L_{FP}P_{-WP}}$. MEU is the complete Bayes expected loss and represents the Mean Expected Utility $\boldsymbol{U_{TP}P_{+WP} - L_{FP}P_{-WP}}$. TEU is the product of the nPairs and the Bayes expected loss and represents the Total Expected Utility, $\boldsymbol{n_{WP} \times (U_{TP}P_{+WP} - L_{FP}P_{-WP})}$.

| *Workflow Path* | *Number of pairs* | *PrPlus* | *PrTrue* | *PrFalse* | *Utrue* | *Lfalse* | *MEU* | *TEU* |
|---|---|---|---|---|---|---|---|---|
| AllGeneIDs | 62 | 0.910 | 0.910 | 0.0897 | 0.910 | 0.0897 | 0.821 | 50.87 |
| NTM | 48 | 0.962 | 0.962 | 0.0376 | 0.962 | 0.0376 | 0.925 | 44.39 |
| LC | 50 | 0.907 | 0.907 | 0.0927 | 0.907 | 0.0927 | 0.815 | 40.73 |

## 5.3 EVALUATION OF RNASEQ THRESHOLD SELECTION

The next evaluation considers the same breast cancer data, but uses a different application of threshold selection. Three cutoff points (1000, 5000, 10000 gene count) separate the feature pairs into 4 groups that are evaluated as separate workflow paths. Table 10 lists the workflow components and the 4 workflow paths. The workflow paths differ only in the threshold step where a mean gene count is determined across the samples and the thresholds of 1000, 5000, and 10000 create 4 different workflow paths with 62, 59, 38, and 16 feature pairs, respectively.

**Table 10.** SALMON RNASeq threshold evaluation workflow paths and data input.

A table for the evaluation of a threshold workflow component applied to SALMON processed RNASeq gene count data. The final quantified gene counts are included if the mean gene counts across the samples is above the threshold of 1000, 5000, or 10000 gene counts from the SALMON workflow. The boxed section is the input for EUFLOW.

| **Workflow** | | ***Workflow Paths*** | | | |
|---|---|---|---|---|---|
| ***Workflow Component*** | ***Workflow Options*** | ***SALMON*** | ***SALMON gene counts over 1000*** | ***SALMON gene counts over 5000*** | ***SALMON gene counts over 10000*** |
| Alignment | Read –free alignment | Read –free alignment | Read –free alignment | Read –free alignment | Read –free alignment |
| Assembly | SALMON | SALMON | SALMON | SALMON | SALMON |
| Quantification | SALMON | SALMON | SALMON | SALMON | SALMON |
| Threshold | 1000 count mean<br>5000 count mean<br>10000 count mean | None | 59 mRNA features with sample mean over 1000 | 38 mRNA features with sample mean over 5000 | 16 mRNA features with sample mean over 10000 |
| EVALUATION DATA VALUES | | SALMON Gene count values<br>62 mRNA<br>406 samples | SALMON Gene count values<br>59 mRNA<br>406 samples | SALMON Gene count values<br>38 mRNA<br>406 samples | SALMON Gene count values<br>16 mRNA<br>406 samples |
| REFERENCE DATA VALUES | | 62 RPPA fold change protein features | | | |
| IDENTIFIER MAP | | 62 mRNA (GENE SYMBOL)<br>62 proteins (GENE SYMBOL) | | | |

Model quality is determined via Pearson correlation for each of the pairs. The correlation density is in Figure 20 (black line). Using a bootstrap procedure variance and bias are estimated and the mixture distribution is deconvolved via Expectation-Maximization, the resulting mixture components are represented by the dotted blue lines in Figure 20, and the resulting posterior probability is estimated.

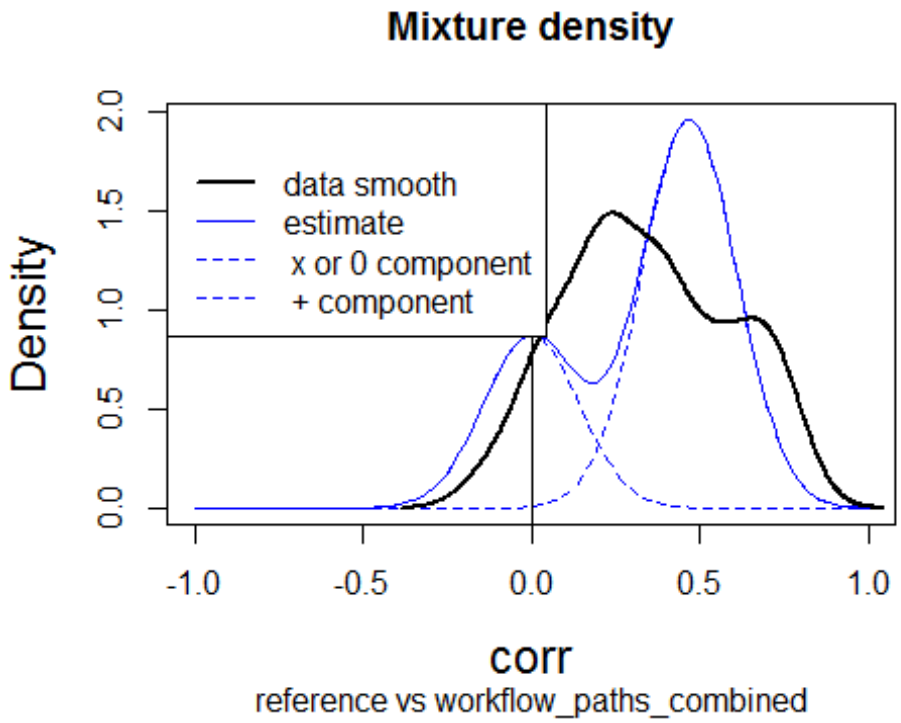


**Figure 20.** Mixture distribution from the threshold evaluation.

Mixture distribution plot obtained by running the EstimatePosteriorProbability function in EUFLOW on the model quality parameters from the RNASeq/RPPA pairs. The black data smooth line is the empirical correlation density, the solid blue line is the mixture fit estimate, and the dashed lines represent the two predicted components.

**Table 11.** Utility table for RNASeq threshold example.

The number of pairs used in the workflow path is defined as nPairs. PrPlus is $\boldsymbol{P_{+WP}}$ from Chapter 3 and is the optimally weighted mean of the proportion of pairs that are estimated to be in the + component for each workflow path. PrTrue is equal to $\boldsymbol{P_{+WP}}$ unless a deltaPlus factor is applied by the user. As defined previously, the deltaPlus is proportion of valid pairs that the user expects to be in the + component. Prfalse is $\boldsymbol{P_{-WP}} = \mathbf{1} - \boldsymbol{P_{+WP}}$, Utrue is calculated as the product $\boldsymbol{U_{TP}P_{+WP}}$ which is the first part of the Bayesian expected loss calculation. LFalse is the second part of the Bayesian expected loss $\boldsymbol{L_{FP}P_{-WP}}$. MEU is the complete Bayes expected loss and represents the Mean Expected Utility $\boldsymbol{U_{TP}P_{+WP} - L_{FP}P_{-WP}}$. TEU is the product of the nPairs and the Bayes expected loss and represents the Total Expected Utility, $\boldsymbol{n_{WP} \times (U_{TP}P_{+WP} - L_{FP}P_{-WP})}$.

| *Workflow Path* | *Number of pairs* | *PrPlus* | *PrTrue* | *PrFalse* | *Utrue* | *Lfalse* | *MEU* | *TEU* |
|---|---|---|---|---|---|---|---|---|
| **Allfeatures** | **62** | **0.879** | **0.879** | **0.121** | **0.879** | **0.121** | **0.758** | **47.0** |
| *Over1000* | *59* | *0.896* | *0.896* | *0.104* | *0.896* | *0.104* | *0.791* | *46.7* |
| Over5000 | 38 | 0.824 | 0.824 | 0.176 | 0.824 | 0.176 | 0.647 | 24.6 |
| Over10000 | 16 | 0.862 | 0.862 | 0.138 | 0.862 | 0.138 | 0.723 | 11.6 |

Table 11 is the evaluation table from EUFLOW for the number of feature pairs, parameters and Mean Expected Utility and Total Expected Utility. From the posterior probability and application of the default utility parameters, we are able to calculate the MEU and TEU for each threshold cut point. If TEU is the criterion for deciding the threshold then not filtering at all is the WP decision. However, if MEU is the criterion then a threshold of 1000 is the WP decision. It is important to remember that TEU selects workflow paths in filtering/threshold examples by being a stringent criterion and is optimal users that prefer not to lose data (See Chapter 3 for details).

## 5.4 EVALUATION OF COMMON RNASEQ WORKFLOW PATHS

EUFLOW can also be applied in a very different context, such as different versions of a workflow path. If two workflow paths produce different levels of gene expression the level of different values

then the correlation values can be very different as well. This can result in data that is inconsistent and may influence the data quality. One of the more common choice that users of RNASeq data must make is between multiple versions of the same dataset, simply processed with a different workflow path. It is important to mention users of TCGA RNASeq when it was first available downloaded RNASeqV1 data as the default Level 3 data. When the Mapsplice RNASeqV2 data was available users were able to download both versions of the data. As of 2017, only RNASeqV2 data is available for download directly from TCGA, but the original data (RNASeqV1) is archived at cBio[32]. RNASeqV2 has become the 'de facto' standard for TCGA RNASeq data. However, it is not clear that this workflow path is optimal for all users. Many of the workflow paths presented in Chapter 2, may be optimal but these workflow paths must be recalculated from Level 1 raw data. A user may be interested in whether the "ready to go" Level 3 data is sufficient for their analysis goals or whether they should choose to reprocess the data with one of the countless available workflow paths.

**Table 12.** General and RNASeq specific definition of a workflow.

A table for the evaluation of common RNASeq workflow paths. Alignment, Assembly and Quantification workflow components have different workflow options for the three workflow paths RNASeqV1, RNASeqV2, and Piccolo. The dark outline box represents the input to the EUFLOW package.

| **Workflow** | | ***Workflow Paths*** | | |
|---|---|---|---|---|
| ***Workflow Component*** | ***Workflow Options*** | ***RNASeqV1 Workflow path*** | ***RNASeqV2 Workflow path*** | ***PICCOLO Workflow path*** |
| Alignment | Bowtie,BWA | BWA | Bowtie | |
| Assembly | Mapslice, Samtools | Samtools | Mapsplice | |
| Quantification | RSEM, RPKM | RPKM | RSEM | |
| EVALUATION DATA VALUES | | RNASeqV1 Gene expression values 67 mRNA features 198 samples | RNASeqV2 Gene expression values 66 mRNA features 198 samples | PICCOLO Gene expression values 65 mRNA Features 198 samples |
| REFERENCE DATA VALUES | | RPPA Fold change 67 protein features 198 samples | | |
| IDENTIFIER MAP | | 198 mRNA (GENE SYMBOL) 67 proteins (GENE SYMBOL) | | |

This evaluation is a three way RNASeq workflow path evaluation on TCGA ovarian samples across three different workflow paths, RNASeqV1, RNASeqV2, and the PICCOLO workflow path. The RNASeqV1 data uses the RPKM method which quantifies gene expression by normalizing for total read length and the number of sequencing reads[8]. RNASeqV2 carefully considers splice junctions using Mapsplice and RSEM to quantify gene expression[9]. The Piccolo workflow path uses the Rsubread package and reports the data via feature counts determined from the FKPM[34]. OV RNASeqV1, OV RNASeqV2, OV RPPA data were obtained using the

cgdsr R Bioconductor package (https://CRAN.R-project.org/package=cgdsr). The Piccolo data was obtained from the Gene Expresion Omnibus (GSE62944).

Table 12 lists the workflow components and options employed for the evaluation of sample of common RNASeq workflow paths. Model quality is determined via Pearson correlation for each of the pairs. The density for the Pearson r values is in Figure 21 (black line). Using a bootstrap procedure variance and bias are estimated and the mixture distribution is deconvolved via Expectation-Maximization, the resulting mixture components are represented by the dotted blue lines in Figure 21, and the resulting posterior probability is for each pair. Threes feature pairs contained missing data; RNASeqv2 was evaluated for 66 feature pairs and Piccolo was evaluated for 65 feature pairs.

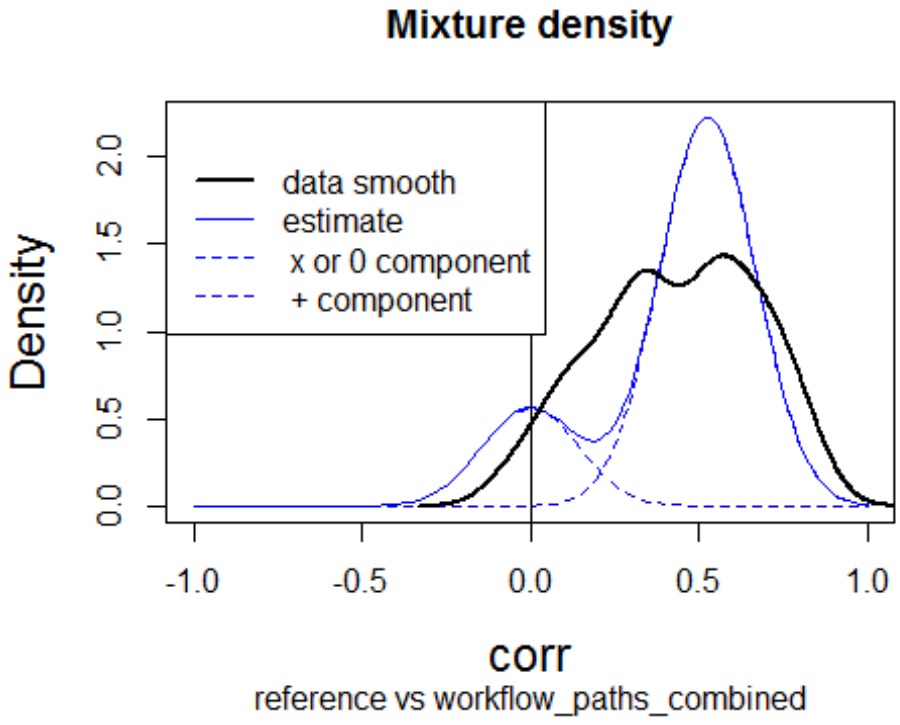


**Figure 21.** Mixture distribution from the Ovarian TCGA workflow path evaluation.

Mixture distribution plot obtained by running the EstimatePosteriorProbability function in EUFLOW on the model quality parameters from the RNASeq/RPPA pairs. The black data smooth line is the empirical correlation density, the solid blue line is the mixture fit estimate, and the dashed lines represent the two predicted components.

**Table 13.** Utility table for RNASeq workflow evaluation example.

The number of pairs used in the workflow path is defined as nPairs. PrPlus is $\boldsymbol{P_{+WP}}$ from Chapter 3 and is the optimally weighted mean of the proportion of pairs that are estimated to be in the + component for each workflow path. PrTrue is equal to $\boldsymbol{P_{+WP}}$ unless a deltaPlus factor is applied by the user. As defined previously, the deltaPlus is proportion of valid pairs that the user expects to be in the + component. Prfalse is $\boldsymbol{P_{-WP}} = \mathbf{1} - \boldsymbol{P_{+WP}}$, Utrue is calculated as the product $\boldsymbol{U_{TP}P_{+WP}}$ which is the first part of the Bayesian expected loss calculation. LFalse is the second part of the Bayesian expected loss $\boldsymbol{L_{FP}P_{-WP}}$. MEU is the complete Bayes expected loss and represents the Mean Expected Utility $\boldsymbol{U_{TP}P_{+WP} - L_{FP}P_{-WP}}$. TEU is the product of the nPairs and the Bayes expected loss and represents the Total Expected Utility, $\boldsymbol{n_{WP} \times (U_{TP}P_{+WP} - L_{FP}P_{-WP})}$.

| *Workflow Path* | *Number of pairs* | *PrPlus* | *PrTrue* | *PrFalse* | *Utrue* | *Lfalse* | *MEU* | *TEU* |
|---|---|---|---|---|---|---|---|---|
| RNASeqv1 | 67 | 0.950 | 0.950 | 0.0504 | 0.950 | 0.0504 | 0.899 | 60.2 |
| ***RNASeqv2*** | ***66*** | ***0.972*** | ***0.972*** | ***0.0279*** | ***0.972*** | ***0.0279*** | ***0.944*** | ***62.3*** |
| PICCOLO | 65 | 0.878 | 0.878 | 0.1218 | 0.878 | 0.1218 | 0.756 | 49.2 |

In Table 13, RNASeqv2 provided the highest MEU and TEU, primarily due to the high PrPlus and low PrFalse values. All three WPs had very high PrPlus values likely due to the filtering and threshold steps incorporated into these workflow paths.

# 6.0 DISCUSSION

The multitude of workflow paths available for bioinformatics provides diverse and unique ways to process raw data. Here I have demonstrated the evaluation of different types of workflow evaluation in three publications and the subsequent development of the EUFLOW package. EUFLOW guides users to utilize a workflow path by carefully considering error tolerance and biological relationships within paired data.

## 6.1 RELEVANCE TO BIOMEDICINE

Biotechnology tools such as RNASeq gene profiling could be extremely powerful in diagnosis and treatment of disease. However, if two widely accepted workflow paths produce clinical results which are inconsistent then clinical implementation will appropriately be questioned. Reproducibility of data is a major obstacle in effective clinical adoption of high-throughput genomic and proteomic data.[145,146] One argument to increase reproducibility is standardization of workflows. Standardization of a workflow path is often difficult because the standard must be determined to be the most reliable and accurate among the choices. Gene expression microarray and RNASeq analysis are two examples where hundreds of workflow path tools have been developed yet no standard analysis workflow path has emerged. Another solution is transparency and demanding that workflows are not only published but also evaluated against multiple other

workflows. An alternative argument can also be posed, rather than instill a standard workflow path (which unlikely to be universally accepted) provide a means to evaluate alternative workflows. RNAseq is at the cusp of being clinically implemented as a diagnostic tool, but before RNAseq is effectively implemented in the clinical environment, problems in the RNAseq workflow path consistency should be addressed.

If another workflow path were utilized, would the same data lead to similar conclusions in the new workflow path? If the answer to this question is no, then the clinical relevance of RNAseq would be called into question. Relevance could only be retrieved if we knew that one workflow was the most reliable. In order for RNAseq data to be clinically relevant and generalizable we must determine a way to evaluate workflow paths in RNAseq analysis. In Chapter 2 there were many examples of the vast differences between final processed datasets, however little work has been performed to demonstrate the downstream impact of the differences in these datasets on the scientific conclusions. I would like to highlight that workflow path choice impact on analysis outcome is an underdeveloped area of research in bioinformatics.

## 6.2 INNOVATION

The EUFLOW package, to the best of my knowledge, is the first methodology to evaluate bioinformatics workflow paths that 1) is ***usable*** to address any bioinformatics workflow choice issue, 2) uses ***real data*** on biological samples to perform the evaluation, and 3) allows the user to select the best workflow path for a preferred explicit trade-off between correctly including and incorrectly including a feature, reflecting the goals of the analyst. Threshold selection is also a

common task as the low count sensitivity of gene expression and dynamic range of proteomics demands a line to be drawn to decide the detectability of certain molecules. Identifier mapping is a less common workflow component, but especially valuable for integrated analyses.

In the identifier filtering of Affymetrix probsets, not only did EUFLOW evaluate the workflow paths using real data, but it was also able to evaluate combinations of filters created by intersections and unions. Enabling the evaluation of combinations of methods is powerful.

Using real data for evaluation is uniquely valuable. Chapter 2 has demonstrated evaluation of RNASeq workflow paths using simulated data, but the limitations of our biological understanding casts doubt about the degree of realism that simulation can provide; discovery of novel features in molecular biology continues. The use of transcript-to-protein correlation is imperfect (see limitations below), but it does correspond to a basic expectation connecting the genome to phenotypes. EUFLOW can go beyond this with refined models and model quality scores that consider other biology, such as miRNA and DNA methylation.

Finally, in regard to utility as a feature of EUFLOW it is important to consider that not all users have the same purpose in an analysis. In the identifier filter example, removing the transmembrane feature pairs from an analysis gives an improvement on the mean expected utility from 0.821 to 0.925. However the Total Expected Utility is not improved by using the non-transmembrane proteins only. EUFLOW enables utility and loss to be factored into the workflow path choice, as well as the criterion for the choice.

## 6.3 LIMITATIONS

The EUFLOW package is limited by the availability of data in two parallel platforms with a *Model quality* that makes sense to the user. It is often difficult to find protein expression data or other reference data on the same samples as your evaluation. When reference data is available then it is often in a smaller feature size that the evaluation data. In the RNASeq filtering evaluation only 62 verified antibodies were available for the breast cancer data as opposed to 1109 gene identifiers from the evaluation dataset. Currently the availability of paired data across platforms is limited to a few repositories, such as TCGA. EUFLOW is enhanced by the availability of paired data, and hopefully more data in this format is available in the future.

Another limitation is that real data based correlation is not a flawless model quality. Biological interference such as miRNA switching off translation, the impact of proteasomes destroying protein after it is translated, and other biological deregulating factors, mRNA and protein are not always expected to have a positive correlation. The EUFLOW framework accounts for discordant but biologically connected data, however, the impact of this factor in a particular dataset is unknown.

## 6.4 FUTURE WORK

For future development of the EUFLOW package, my first step would be to complete a usability test for users of RNASeq workflow paths to identify needs for future versions of the EUFLOW package. Many extensions are desirable, More flexibility in choice of model quality criterion would be desirable. More visual representations of the effects of the utility parameters on the MEU

or TEU could be helpful. When threshold selection is the goal, visualizing the effect of changing the threshold one unit at a time and plots the TEU or MEU. And finally a more developed input functionality that would guide the processing of workflow paths from within R will enhance usability. This automated piping should integrate with command line tools like SALMON.

My next step would be to develop the EUFLOW package for incorporation into Bioconductor. Bioconductor is open source and open development hub for bioinformatics package development. Due to the fact that Bioconductor contains many complete workflow procedures for many of the use cases mentioned in this proposal, Bioconductor is the ideal location for the EUFLOW package.[35] Finally, conducting workflow path evaluations in workflows such as miRNA target identification, peptide identification in mass spectrometry, and many RNASeq workflow paths would be highly valuable research.

## 6.5 CONCLUSION

There are too many workflow paths to evaluate exhaustively for any bioinformatics based process. Literature searches should be conducted, investigators should communicate, and new workflow paths should be developed. However, after an investigator has completed the searching, communication, and looked at what is new; arbitrary decisions remain. The investigator has specific analysis aims that should also be considered. Without an evaluation framework for these choices users are left to go back and change certain filtering steps or lower certain parameter thresholds to see if this changes the outcome. The EUFLOW package provides the user with a "prior to final analysis" tool to plan a workflow path and any workflow options to be included in the data processing stage. Furthermore the EUFLOW package provides a less arbitrary means of

deciding how an investigator goes forward in future experiments. In conclusion, EUFLOW enables users of bioinformatics workflows to evaluate alternative workflow paths guided by inherent biological relationships and user utility.

# APPENDIX A

## IMPROVING CANCER GENE EXPRESSION DATA QUALITY THROUGH A TCGA DATA-DRIVEN EVALUATION OF IDENTIFIER FILTERING.

*Note: This paper was an aim of the dissertation and is provided here as an appendix for reference in the main document*[2].

Data quality is a recognized problem for high-throughput genomics platforms, as evinced by the proliferation of methods attempting to filter out lower quality data points. Different filtering methods lead to discordant results, raising the question, which methods are best? Astonishingly little computational support is offered to help analysts decide which filtering methods are optimal for the research question at hand.

To evaluate them, we begin with a pair of expression data sets, transcriptomic and proteomic, on the same samples. The pair of data sets form a test bed for the evaluation. Identifier mapping between the data sets creates a collection of feature pairs, with correlations calculated for each pair. To evaluate a filtering strategy, we estimate posterior probabilities for the correctness of probesets accepted by the method. An analyst can set expected utilities that represent the trade-off between the quality and quantity of accepted features.

We tested nine published probeset filtering methods and combination strategies. We used

two test beds from cancer studies providing transcriptomic and proteomic data. For reasonable utility settings, the Jetset filtering method was optimal for probeset filtering on both test beds, even though both assay platforms were different. Further intersecting with a second filtering method was indicated on one test bed but not the other.

## A.1 INTRODUCTION

Do commonly utilized methods to process raw data from the high-throughput genomic platforms differ much from each other? Does it matter which methods are utilized to process the data? Repositories of information such as the Gene Expression Omnibus, cBioPortal, and The Cancer Genome Atlas (TCGA) contain hundreds of platforms and thousands of patient samples[31–33]. These platforms include measurement of gene expression, copy number variation, protein expression and post-translational modification. All of this information is available to users in "levels" of data, where, for most users, only the processed data is available. Some workflow options are "ready-to-go": the data available are pre-processed, such as Affymetrix HU133 Plus 2.0 data, RNA-Seq data, methylation data, and many other data types in TCGA. Alternatively, many analysts prefer to start with raw data and apply a customized workflow consisting of their preferred sequence of processing steps. *Workflow options* include ready-to-go workflows, custom workflows, individual processing steps, or tuning parameters in particular steps. Any change in a workflow step or change of a parameter setting constitutes a new workflow option. To what extent do these choices affect the final dataset to be analyzed? If the datasets differ substantially, will they differ in quality? If so, how can we tell which is best? Finally, will soundness of the scientific conclusions be harmed by sub-optimal workflow choices, and improved by better choices? Surprisingly these questions are scarcely addressed in the

bioinformatics literature. Aside from the obvious benefit that the quality of analyses could be improved, there is the issue of comparing results from different studies. When two investigations report on comparable data sets, a third party may wish to compare or contrast the results, for example for scientific validation. The choice of different workflows in the two studies generates a potential confounder in comparing them. Greater consensus on workflow choices would help alleviate this problem.

An example of a data setting burdened by a poor understanding of workflow option choices is the Affymetrix microarray. Affymetrix expression data is publicly available for over 35,000 datasets, and is an immensely valuable resource for almost every type of cancer research[147]. However, there is no de facto standard of determining the gene expression values from raw data. Many processing and normalization options can yield values of gene expression on about 18,000 gene products from an ambiguous set of 54,675 probesets. A critical step in an Affymetrix workflow is to remove, or "filter", poor quality probesets. This process of removing "bad" measurement points has been defined previously as identifier filtering[3].

Identifier filtering applied to Affymetrix chips presents an opportunity to evaluate workflow options concisely. We previously performed a comparison, not an evaluation, of identifier filtering. In identifier filtering, the user removes features (i.e. probesets) judged to do a poor job reflecting expression of their intended gene products. Table 14 outlines the identifier filtering implementations tested here, including PlandbAffy (PD), JetSet (J), AffyTag (AT), AffyGrade (AG), Masker (M), EnCode (E), and three methods deriving from GeneAnnot (GSPE,GSEN, GQ).[60,62,5,64,148]. Table 14 also provides the abbreviation utilized in this paper for each of the nine filtering strategies. These methods apply diverse criteria that consider nucleotide complementarity, probe design, and cross hybridization of probe to off-target gene product.

This article presents a comprehensive evaluation of workflows consisting of probeset identifier filtering methods and their combinations. The methodology is previously published[149]; and this paper is an application of the methodology to an important problem in bioinformatics practice. As a test-bed for evaluation, it utilizes transcript expression data paired with protein expression data. However, the goal of this work is not specifically to guide analysis of paired datasets, but rather a much broader goal, to provide guidance for feature filtering in transcript expression experiments.

Prior research by our laboratory group has documented disagreements among resources that map between identifiers for probesets and identifiers for proteins[150]. This work was implemented as a Bioconductor[35]package, IdMappingAnalysis[140]. We showed that the quality of the mappers could be compared based on real biological data[150]. Subsequent methodological work created a more general decision-theory-based approach, and demonstrated how other workflow elements besides identifier mappers, including filtering methods and threshold choices, can also be evaluated [149].

**Table 14.** Identifier filtering methods and the scores utilized for filtering

| Filter Symbol | Description | Developer Criteria | Identifier Filtering |
|---|---|---|---|
| **AT**[57,58] | Affytag - Pre-2004 Affymetrix annotation for the Affymetrix HGU133 Plus 2.0 array | Original annotation determined by mapping to UniGene and Locus Link. "_at is considered unique. | Filter al annotation tags that begin with "_[agirxsf]_at" |
| **AG**[57,58] | Affy Grade - Netaffx Transcript Assignment Pipeline | "A" grade is the highest grade where ≥ 9 probes match transcript sequence. | Filter grades not equal to A. |
| **M**[59] | Masker - National Cancer Institute alternative chip definition file (CDF) masking out probesets with poor target location | A CDF file which eliminates a probe when more than 2 nucleotides to not match the target as well as nonspecific probes | Filter any probeset that has no remaining probes on the mask |
| **GSEN**[60] | GeneAnnot Sensitivity | The fraction of the probes in a probeset that match Watson-Crick nucleotide base pairs in the nominal gene | Filter probesets with Geneannot Sensitivity < 90% |
| **GSPE**[60] | GeneAnnot Specificity | Sum over the number of matching probes with lower weight to non-specific probes | Filter probesets with Geneannot Specificity ≤ 50% |
| **GQ**[60] | Geneannot Quality Score | A pipeline which confirms the probeset annotation with GeneCard data. | GQ= 1 is confirmed entirely with GeneCard data; Filter probesets with a GQ = [2-6] |
| **E**[61] | Encode - Encyclopedia of DNA elements | Protein coding genes are determined by human curation, RNA sequence and comparative genomics | Filter all probesets that map to a non-"Protein coding" target |
| **PD**[62] | PlandbAffy database | BLAT of target to the probe and evaluation of nucleotide mismatch or exon location | Filter all probesets with a proportion of "good" probes <30% |
| **J**[5] | Jetset Bioconductor package | Determines features such as robustness of the probe, coverage, as well as nucleotide alignment with the reference genome | Filter all except the highest-scoring probeset among those annotated for target gene. |

For a variety of reasons, previous investigators have examined correlations between data from pairs of expression platforms, for example relating RNA-Seq to oligonucleotide data, and relating oligonucleotide data to protein expression data[151–154]. A natural assumption is that greater transcript expression will lead to greater protein expression. There are, however, biological reasons that a particular mRNA species might have weak or no correlation with the expression of the

correctly mapped protein [149,153–157] . The evaluation method applied here takes that into account, as we shall see.

## A.2 METHODS

For reference, an overview of the methodology appears in Figure 21.

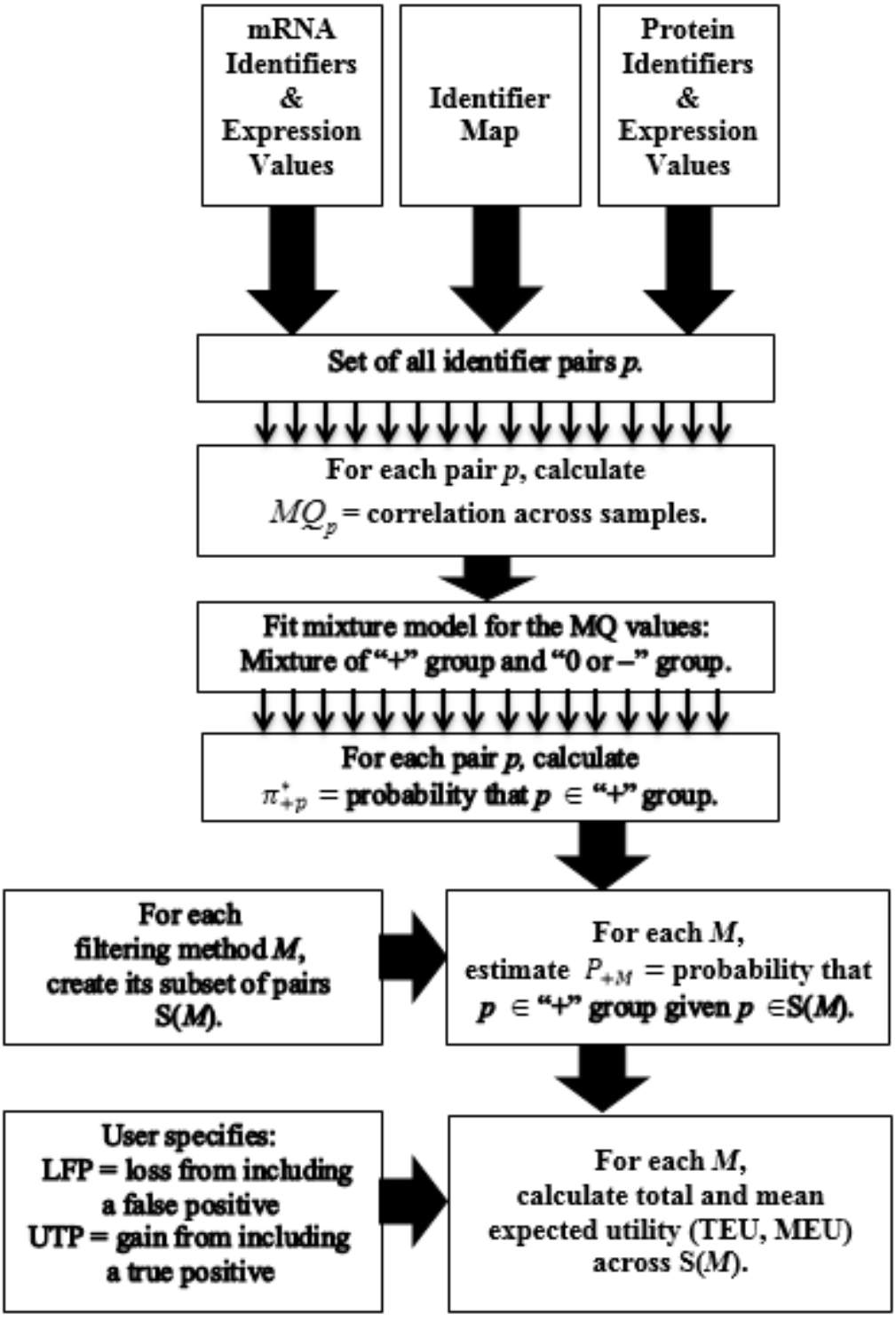


**Figure 22.** Identifier filtering flowchart

Two cancer datasets were utilized as test-beds for this evaluation of filter methods. The first is a data set of 91 endometrial cancer samples and 7 normal endometrium samples, studied with tandem mass spectrometry proteomic data and Affymetrix U133 2.0 Plus expression data from the Gynecologic Cancer Center of Excellence (GynCOE)[150]. The second is a data set of 401 ovarian serous cystadenocarcinoma samples with (RPPA) protein assay and Affymetrix U133A mRNA data from The Cancer Genome Atlas (TCGA)[135,158]. These data sets differ substantially in sample size, number of features, and platforms. Proteomic and mRNA feature identifiers are paired across platforms using the IdMappingRetrieval Bioconductor package[159,160]. The EnVision mapping was selected based on the results from our previously published evaluation of identifier mapping resources [150,161].

The endometrial cancer biomarker studies were performed by the Gynecologic Cancer Center of Excellence [150,162,163]. The tissue samples were subjected to trypsin digest at the University of Pittsburgh. Tryptic peptide digests were separately analysed in duplicate by LC-MS/MS with an LTQ-FT (ThermoFisher Scientific, Inc, San Jose CA) and an LTQ-Orbitrap (ThermoFisher Scientific Inc.) mass spectrometer. The combined analyses yielded 12,288 distinct protein UniProt accessions across all samples and both instruments. The gene expression data was performed on the Affymetrix U133 2.0 Plus array. For complete details of the microarray and proteomic studies, see Day et al[150].

For the second test-bed, we turned to TCGA. TCGA has multiple levels of genomic, transcriptomic, somatic mutation, and protein expression data for many types of cancer data. The ovarian serous cystadenocarcinoma sample dataset is especially useful here. The ovarian cancer data has 401 samples with various types of genomic, transcriptomic, and proteomic data. The data utilized here comes from two platforms: the U133A Affymetrix array, with 22,277 probesets, and

the RPPA studies on 68 proteins performed by M.D.Anderson Cancer Center. The proteins selected for the RPPA studies were chosen for their cancer relevance. Using the IdMappingRetrieval Bioconductor package[164], we obtained 151 probeset-to-protein pairs.

Nine filtering methods were evaluated and compared and they are listed in Table 14. Affytag (AG) removes probesets for which the Affymetrix identifier (ID) contains a qualifier; that is, the ID ends in "_[agirxsf]_at", reflecting original doubts concerning the correct and unique hybridization of the probes in each probeset, as documented by Affymetrix when the array was designed[57,58]. Although the identifier tags were initially used as the 'de facto' quality measure, these tags had reliability problems. We include this tag-based probeset quality measure to verify that our quality assessment paradigm can detect the expected deficiency of performance in a superseded method relative to the more recent measures. AffyGrade (AG), provided by the NetAffx array annotation file, is a quality grade labelled as A, B, C, R, and others. Only probesets with "A" grade were accepted, since "A" grades represent at least 9 "matching probes" to the target mRNA[57]. The NCI Masker [59] filter removes probesets omitted from the NCI "masked" chip description file (CDF). Masker was produced by the NCI Laboratory of Population Genetics. The CDF file eliminates any probes that do not have at least 24 out of 25 nucleotides match the target GenBank transcript. In addition, it eliminates any nonspecific probes that map to a different chromosome, strand, or are part of a gene cluster that could cause cross hybridization.

We test three filters utilizing Geneannot[60], a database of gene expression annotations and quality which evaluates the Affymetrix probesets on the following criteria. For each of the probesets on the Affymetrix chip sensitivity, specificity, and overall quality score is determined. Sensitivity is defined as the fraction of the probes in a probeset that match Watson-Crick nucleotide base pairs in the nominal gene. This classification is labelled as Geneannot Sensitivity (GSEN).

The next classification is labelled Geneannot Specificity (GSPE) and is a sum over the number of matching probes with lower weight to non-specific probes. Thresholds defining GSEN and GSPE were, respectively, sensitivity metric $\geq 0.9$ and specificity metric $\geq 0.5$, each chosen by maximizing expected utility. Finally, the Geneannot quality measure (GQ) is determined from the ordinal rank assigned by Geneannot to demonstrate the confirmation of the probeset to mRNA match. A score of "1" is reported to be the "best", which demonstrates that the probes were confirmed using the GeneCard data via Entrez Gene or Ensembl. The worst score is a "6", which is defined as probesets where the only information available is original Netaffx annotation[57]. For the purposes of this study GQ accepts only probesets with a "1" score. Our EnCode (E) filter utilizes the EnCode[93] project's determination of protein coding status of the target sequence location in the genome, to remove probesets of non-coding targets. The files are available at http://encodeproject.org. The GENCODE version 12 annotation files were utilized to determine gene status from human genome build 37. The gene status is classified as protein coding, transcribed pseudogene, untranscribed pseudogene, lincRNA, not identified by Genecode, et cetera[61]. Only probesets with the "protein coding" Ensembl code were accepted. The Ensembl codes were matched to the Uniprot accession code present in our analysis.

The PlandbAffy (PD) filter utilizes the PlandbAffy [62] database, which uses the probeset sequence and the BLAT database to align probe nucleotide sequences to the target and assign to each probe a grade reflecting alignment mismatches, alignment to other sequences risking cross-hybridization, and intronic versus exonic location. The PlandbAffy filter was defined to accept a probeset if 30% of the probes within the probeset were classified as perfect exonic, non-cross hybridizing matches. To set the threshold, we maximized expected utility, as described in Day and McDade[149]. The Jetset (J) filter uses the Jetset [5] assessment, which also considers

nucleotide complementarity across the probesets but also considers splice isoform coverage, and transcript degradation. In addition, JetSet (J) will score each probeset of a target gene and select the best probeset (of currently defined probesets) for each gene on the chip. Therefore, Jetset (J) is a stringent eliminator of probesets.

The identifier filtering evaluation of probesets uses a previously published methodology for comparison of bioinformatics workflow options to determine the evaluation metric. The steps in this application to identifier filtering are summarized in Figure 21. For more details, see Day and McDade 2013[149]. The method requires the following inputs:

- A large number of *biological samples* from a biological repository, such as TCGA, or a private collection of biological samples.
- Two high-throughput platforms each with a feature list of identifiers; the two platforms must be on the same *biological samples.*
- A planned set of workflow options to compare.
- An identifier map, which connects the pairs of data across the platforms (i.e. transcript to protein).
- A *Model Quality Score* for each pair *p,* designated $MQ_p$. The *MQ* scores are treated independently for modelling the mixture distribution. In applications thus far, this score is a correlation coefficient.
- For each method *M,* the set of pairs accepted or produced by that method is designated as *S*(*M*).

In the current application, each pair is an mRNA transcript feature paired with a protein feature linked through the EnVision identifier mapping resource. Membership of a pair *p* in the set *S*(*M*) means that method *M* claims that the transcript feature in *p* should be included for any data analysis. The two platforms, respectively, assess the two processes gene expression and protein expression.

The *Model Quality* score in this application is the correlation of the two measurements across the biological samples. We consider the probability density of the correlation values for all pairs produced by the methods *M* (Figure 23, black line). This density is modelled as a mixture with the following components:

- "+": The transcript feature and the protein feature are correctly identified and they are truly *biologically coupled.* This means that a pair in this component is correctly mapped between transcript and protein identifier, and transcript abundance and protein abundance are monotonically related. The blue line in Figure 23 represents the "+" component.
- "0": The transcript feature and the protein feature are correctly mapped, but *biologically decoupled.* This means that the expected monotonic relationship between a transcript and a protein are not observed. There are many biological reasons for decoupling, including RNA interference by microRNA's, post-translational processing, and any other mechanism causing the protein abundance to fail to reflect the transcript abundance. The green line in Figure 23 represents the distribution of correlations for decoupled pairs, which we refer to as the "0" component.
- "*x*": An incorrect mRNA/protein pair relationship was assigned. The red dashed line in Figure 23 represents the distribution of correlations for misidentified pairs, which we refer to as the "*x*" distribution. Pairs included in this distribution should not be assigned. These

assignments may be due to incorrect actions on the part of the identifier mapping or the workflow in general.

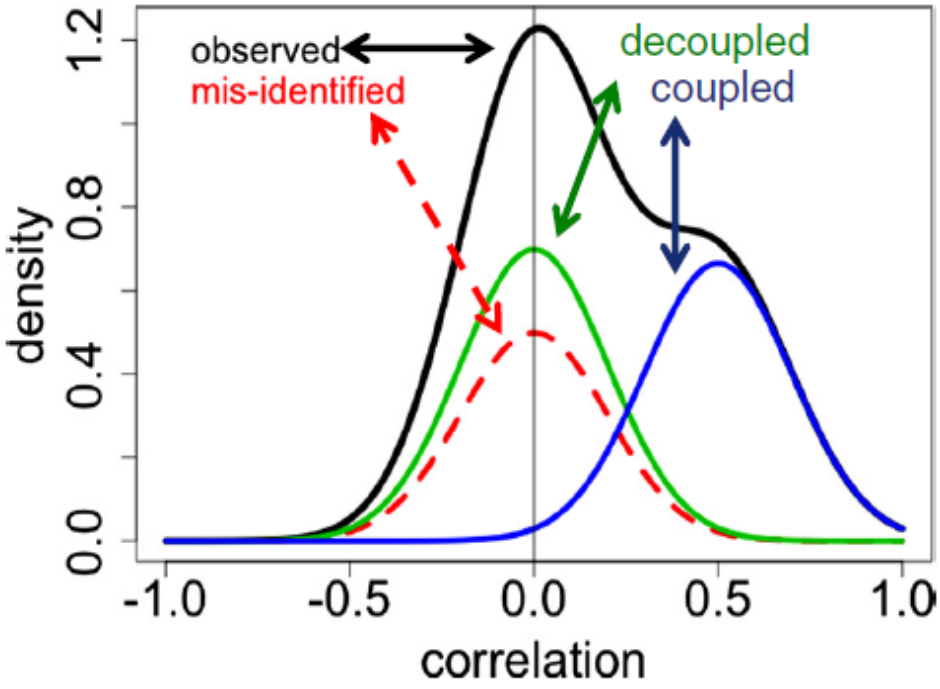


**Figure 23.** Mixture distribution example from Day and McDade (2013).

Observed (black): marginal density of correlations. Mis-identified (red, dotted): density of correlations where either feature is mis-identified, or they are incorrectly mapped. Decoupled (green): density of correlations of pairs correctly mapped but biologically uncorrelated ("discordant"). Coupled (blue): density of correlations of pairs correctly mapped and biologically coupled.

We would like to identify the features in either "+" or "0" for inclusion. However, the data cannot distinguish between the "0" and "*x*" groups. Under mild assumptions, the method with the highest posterior probability for "+" is also the method with the highest posterior probability for "+" or "0". We refer to the combined "0" and "*x*" groups as the "-" group. Even though groups "0" and "*x*" cannot be distinguished, basing the relative performance of workflow methods on the

mixture distributions from the observed correlations is likely to yield the correct decision; the argument for this statement is previously reported[149].

Let $G(p)$ be the component, whether "+", "0", or "*x*", that pair $p$ belongs to. A better workflow option should do a better job at excluding incorrectly mapped pairs (i.e. those with $G(p)$ = "*x*"). Increasing the probability that $G(p)$ = "+" should reduce the $G(p)$ = "*x*" component. Let the proportion of pairs in group $g$ be $\Pr(G(p) = g) = \pi_g$ for $g \in \{" + ", "0", "x"\}$. The mixture model provides the opportunity to estimate $\Pr(G(p) = "+")$ for each pair $p$. This probability provides the metric we need to evaluate workflow options.

We now assume that the true correlations for all the pairs in group $g$ are distributed as a mixture of normal distributions with mean $\varphi_g$ and variance $V_g$. There is also measurement error, so the correlation of each pair $p$ in group $g$ is normally distributed with marginal mean $\varphi_g$ and marginal variance $\tau_{gp} = V_g + \sigma_p^2$, where $\sigma_p^2$ is the measurement error variance specific to pair $p$. We estimate $\sigma_p^2$ by the bootstrap method, as described in Day & McDade 2013[149]. To estimate the probability of a pair belonging to the "+" group we use an ECM algorithm to determine the following parameters: 1) the prior probability $\pi_+$ of belonging to the "+" group, 2), the within-group true variance $V_+$ of the correlations in "+" group and 3) the within-group true variance $V_- = V_0 + V_x$ . Here "true" signifies without sampling error. This is possible since we are able to constrain the mean of the "0" and "*x*" groups to 0. For a complete description of the ECM algorithm see Additional File 1 from Day and McDade 2013[149].

Having determined the maximum likelihood estimates of the parameters, we can now calculate for each pair $p$ the posterior probability of belonging to the "+" group by defining:

$$\pi^*_{+p} = \Pr(G(p) = + \mid MQ_p \text{ and parameter estimates})$$

$$\pi^*_{-p} = 1 - \pi^*_{+p} = \pi^*_{xp} + \pi^*_{0p}$$

This calculation provides the posterior probability that pair $p$ belonging to the "+" component, given the correlation $MQ_p$ and its sampling variance $\sigma_p^2$, estimated from bootstrap sampling. To convert that variance into the variance of the posterior probability, the delta method approximation is used. This consists of multiplying the variance of the correlation times the square of the derivative of the posterior probability as a function of the correlation.

$$v^*_{+p} = var\left(\pi^*_{+p}\right) \cong var\left(MQ_p\right) \times \left(\frac{d\pi^*_{+p}}{dMQ_p}\right)^2$$

A weighted mean of the "+" proportion provides an expected proportion of "+" group pairs for a given identifier filtering method. The weighted mean is estimated using the posterior probabilities of each pair and the variances of these posterior probabilities.

$$P_{+M} = \sum_{p \in S(M)} \pi^*_{+p} (v^*_{+p})^{-1} \Big/ \sum_{p \in S(M)} (v^*_{+p})^{-1}$$

This quantity provides the basis for comparing the methods, $M \in \{M_1, \ldots, M_K\}$.

It is important to consider that different analysts have different analysis goals. One method may include a pair or a feature while another excludes it. The pair will be either a "true positive" of the first method or a "true negative" of the second. The relative value of including a true positive versus excluding a false positive will be different for different scientific goals. Utility values can express these valuations. We utilize the Bayesian decision principle of maximizing expected utility. This principle is useful for selecting a single filtering method, choosing a threshold for a

method (Geneannot, PlandbAffy), or selecting a Boolean composite filtering strategy (described in the next section).

For this study, we set the following values:

$L_{FP}$ = the loss associated with a "false positive" = 1,

$U_{TP}$ = the utility of including a "true positive"=2,

We explored sensitivity of the comparisons between methods to these three values, and found that the comparisons are relatively insensitive (data not shown). The Bayesian expected loss calculation is:

$$EU = U_{TP}P_{+M} - L_{FP}P_{-M}$$

This is the Mean Expected Utility (MEU). As an alternative the analyst may choose to use Total Expected Utility (TEU), which simply is the product of the number of methods compared and mean expected utility.

Boolean conjunction (intersection; "and") and disjunction (union; "or"). operators, can create composite filtering strategies, which are easily evaluated as well. An analyst may consider whether the union or intersection of two or more filtering methods is worth the extra effort. Given a current strategy, for each so-far-unused method, one can automatically construct and evaluate the strategies formed by conjoining this method to the current strategy via conjunction or disjunction. A forward selection assesses the expected utility for each of these conjoined strategies, and chooses the one with the highest expected utility. This is referred to as "greedy" selection because it takes the apparent best step, in sequence. In contrast is the exhaustive search of every

Boolean combination of the methods, which in principle could find better strategies, but is impractical.

## A.3 RESULTS

The nine filtering methods are far from redundant. Many analysts who use one of the filtering methods listed in Table 14 might expect only minimal differences in the probesets retrieved and retained. Instead, the nine filtering method strategies do not demonstrate similar probeset decisions. Table 15 compares the classifications of each pair of methods. Panel A: all probesets on Affymetrix HGU133 Plus 2.0 array. Panel B: only 887 probesets from the ID pairs in the endometrial sample.

Each table entry is the odds ratio from the 2x2 table cross-classifying probesets as either filtered or retained by the two methods. The odds ratio is the product of the agreements divided by the product of the disagreements. An odds ratio of 1.0 indicates that the two classifications are providing independent information; an odds ratio much larger than one indicates redundant information, and an odds ratio much smaller than one indicates contradictory information. For example, the odds ratio of 5.88 comparing Jetset to Encode in Table 15 Panel A indicates considerable redundant information: the odds of a probeset being excluded by Encode is 5.88 times greater if the probeset is also excluded by Jetset versus if it is included by Jetset. In contrast the odds ratio of 1.07 comparing PlandbAffy to Masker indicates nearly independent information: knowing whether Masker includes a probeset says almost nothing about whether

PlandbAffy does. More remarkable still is the odds ratio of 0.345 for Jetset and Masker, which indicates that knowing that Masker includes a probeset considerably *decreases* the odds that Jetset includes it; Jetset and Masker provide contradictory information. (One might hope that they usefully complement each other. The analysis of Boolean combinations will address that hope.) (For details about odds ratios, see Szumilas (2010)[165]. )

For each of the test-beds (endometrial and ovarian), a correlation mixture model was fitted to all feature pairs as described in the Methods section. Figure 24 shows the fitted mixture components for the two test-beds. They appear considerably different. Nevertheless, as we will see, the two mixture models lead to similar comparative evaluations of the filtering methods, suggesting that the best practices conclusions we are seeking may have general application.

**Table 15.** Odds ratio chart for probeset filtering

The table cell entries are the odds ratios assessing the degree of association of each pair of filtering methods. Each table entry is the odds ratio from the 2x2 table cross-classifying probesets as either filtered or retained by the two methods. The odds ratio is the product of the agreements divided by the product of the disagreements. For details of the interpretation of the odds ratios, see text. Panel A: all probesets on Affymetrix HGU133 Plus 2.0 array. Panel B: the 887 probesets from the ID pairs in the endometrial sample.

**A**

| *Filter* | *J* | *E* | *PD* | *GSEN* | *GSPE* | *GQ* | *AT* | *AG* | *M* |
|---|---|---|---|---|---|---|---|---|---|
| *J* | - | *5.9* | *3.7* | *3.1* | *24.7* | *29.3* | *1.2* | *20.3* | *0.35* |
| *E* | | - | *11.4* | *29.3* | *32.5* | *46.4* | *0.3* | *14.1* | *0.60* |
| *PD* | | | - | *6.8* | *6.8* | *6.4* | *0.5* | *4.7* | *1.07* |
| *GSEN* | | | | - | *50.0* | *1103.0* | *0.2* | *301.0* | *0.45* |
| *GSPE* | | | | | - | *760.0* | *0.3* | *53.2* | *0.79* |
| *GQ* | | | | | | - | *0.3* | *50.1* | *0.69* |
| *AT* | | | | | | | - | *0.3* | *0.93* |
| *AG* | | | | | | | | - | *1.85* |
| *M* | | | | | | | | | - |

**B**

| *Filter* | *J* | *E* | *PD* | *GSEN* | *GSPE* | *GQ* | *AT* | *AG* | *M* |
|---|---|---|---|---|---|---|---|---|---|
| *J* | - | *1.87* | *2.34* | *0.957* | *2.62* | *2.73* | *4.91* | *2.81* | *5.05* |
| *E* | | - | *4* | *7.34* | *4.24* | *21.6* | *0.682* | *16.2* | *1.49* |
| *PD* | | | - | *1.23* | *35.9* | *2.13* | *2.19* | *1.45* | *1.46* |
| *GSEN* | | | | - | *0.744* | *Inf* | *0.421* | *70.8* | *1.26* |
| *GSPE* | | | | | - | *Inf* | *5.23* | *2.08* | *1.29* |
| *GQ* | | | | | | - | *0.073* | *127* | *1.32* |
| *AT* | | | | | | | - | *0.312* | *2.77* |
| *AG* | | | | | | | | - | *0.907* |
| *M* | | | | | | | | | - |

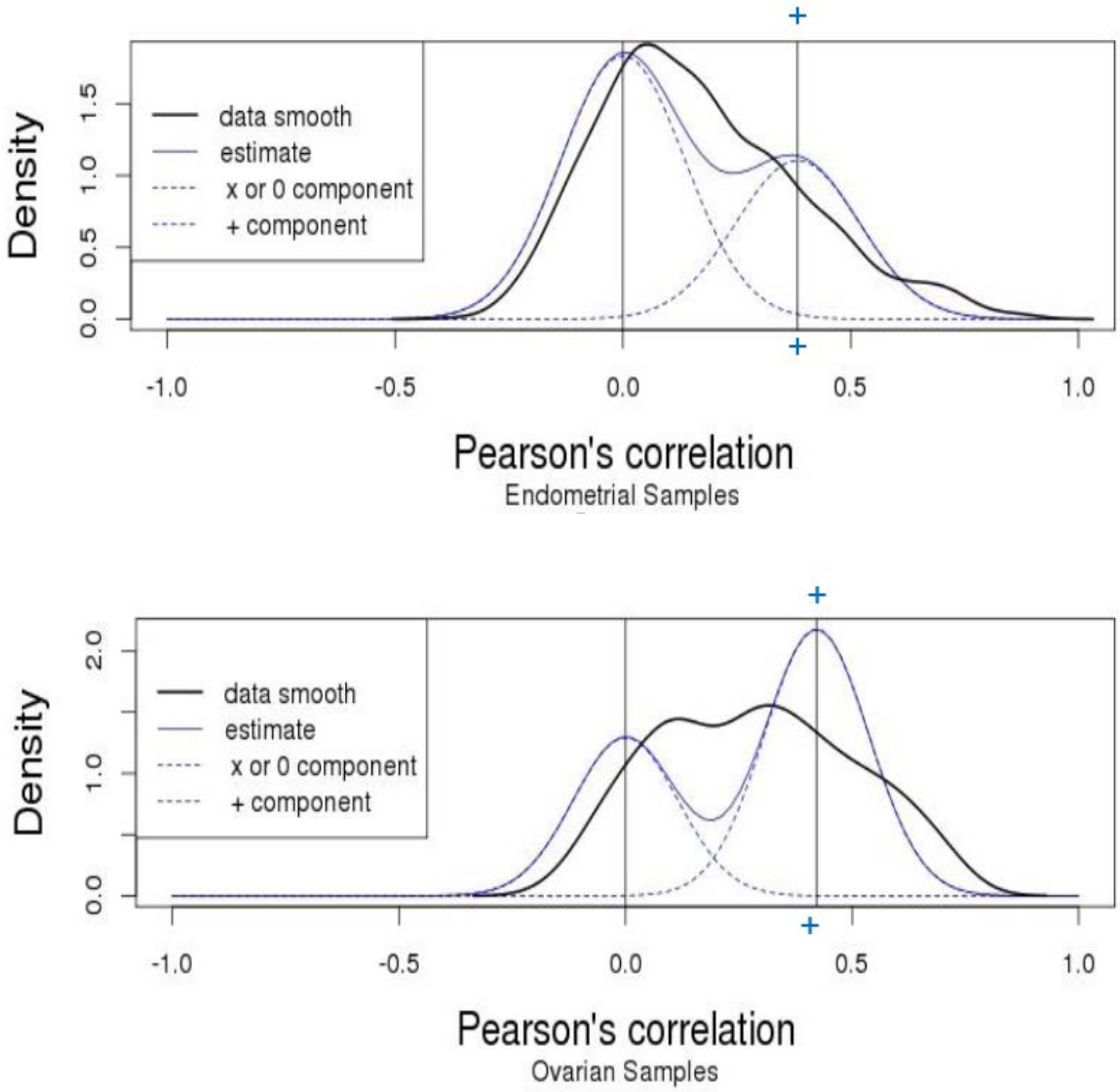


**Figure 24.** Observed and fitted density distributions for probeset filtering example

A) GynCOE Endometrial Experiment, B) TCGA Ovarian Experiment**.** In each panel, the horizontal axis represents the Pearson's correlation for pairs of mRNA expression and protein expression features (887 pairs in panel A, 151 pairs in panel B). The solid black line "data smooth" is a non-parametric estimate of the probability density of observed correlations. The solid blue line is a mixture distribution estimate of the probability density of the true correlations, determined from the generalized EM algorithm, which deconvolves the error term with individual variances for each correlation. The dotted lines are the mixture components. The mixture component labelled "*x* or 0" is interpreted as incorrect or decoupled feature pairs (probability =0.624 in endometrial samples and probability = 0.373 in ovarian samples). The mean is constrained to zero (see Methods). The mixture component labelled "+" is interpreted as correct and coupled feature pairs (probability = 0.376 in endometrial samples and probability =0.627 in the ovarian samples). Vertical lines are placed at the mean correlation of each component.

Each mixture distribution has two components: one centered at zero, and the other with mean > 0. The right-most component, labeled "+", corresponds to the pairs where both features are correctly identified and mapped, and also biologically "coupled" through the translation process, protein synthesis. For each pair, we calculated the posterior probability for belonging to the "+" component. Summing or averaging across the pairs accepted by a filtering method, we calculated the expected utility for that method.

The purpose of probeset filtering is to remove incorrectly identified or ineffectively designed probesets without removing too many correct probesets. Some investigators may want to apply stringent filtering criteria, for example to reduce multiple comparisons penalties and false discoveries, while others would be more concerned with missing a true discovery. For purposes of illustration, we fix a utility of a true positive (UTP) = 2, and a loss of a false positive (LFP) = 1 (see Methods). This implies that an investigator would wish to include a true positive at the cost of including a false positive feature, but not at the cost of including 3 false positive features, with indifference if the cost is two false positives. The different quantity-quality priorities of investigators are represented by two ways of combining expected utilities: the total expected utility (TEU) and the mean expected utility (MEU). An analyst choosing TEU wants as many features as possible, perhaps driven by the need to feed some systems biology algorithm. An analyst choosing MEU is more concerned with the quality of the resulting data set. Summary figures demonstrate the greedy forward selection for the endometrial and ovarian data sets (Figure 25). For each of the filters applied in Figure 25A there is a removal of poor quality

probesets with a gain in TEU. Figure 25B illustrates the MEU over a series of filters, each successive filter reduces probesets with the increase in expected utility.

Figure 26 demonstrates a more detailed picture with each circle data point representing a set of probesets obtained from the application of an identifier filtering method. The two paths represent using TEU or MEU as the metric for the greedy forward selection. For the endometrial data, Figure 26A plots the estimated proportion of true coupled (quality) vs. the number of pairs remaining (quantity), for the endometrial data. The point at the upper left corresponds to including all 887 features pairs obtained with no filtering. The proportion of "+" pairs is only 0.30, which implies that the total expected utility TEU is -81.9 and the mean expected utility MEU -0.0923. The conclusion is that, without filtering, one should not analyze these data. The labeled points correspond to reduced feature sets created by a single filtering method. The paths correspond to successive application of filters selected by a greedy forward selection of intersections and unions.

Jetset filtering provided the best single-method strategy for both TEU and MEU criteria (label = J). It is notable that Jetset was optimal even for TEU despite removing roughly half of the probesets (from 887 to 434; 51.1% probesets removed). For Jetset, TEU = 80.3 and MEU = 0.185, both in the positive zone, suggesting at least that after filtering a data set is of sufficient quality to deserve analysis. Figure 26 shows the subsequent improvements by greedy selection of higher order Boolean combinations for the TEU (Panel A) and MEU (Panel B) criteria, Intersecting Jetset with GSPE was the best next step for the TEU criterion (filtering away 56.1% of the probesets) 138.9 TEU; intersecting with PlandbAffy is the best next step for MEU

(filtering away 59.8%) 0.3868 MEU. Further selection did not improve either criterion noticeably (maximum TEU 148.5, maximum MEU 0.4864).

In the endometrial data set the estimated proportion of true coupled (Pr("+")) is .303 with 887 mRNA-Protein pairs. The endometrial greedy forward selection shows a very similar path and in fact after one greedy node both greedy search modes find Jetset as the methodology of option, increasing the Pr("+") from .303 to .396. The optimal set for total expected utility is Pr("+") = .496, while the mean expected utility finds a set with Pr("+") = 0.503.

In the ovarian cancer data set (Figure 26B), Jetset filtering again provided the best single-method strategy for MEU criteria. Jetset reduced the number of probesets even more severely, from 151 to 47 (78.9% probesets filtered away) for the MEU selection criterion. The benefit in terms of the quality was quite dramatic but the cost in terms of pair reduction actually decreased the total expected utility from 290 to 131. Figure 26B shows the subsequent improvements by greedy selection of higher order Boolean combinations. Intersecting Jetset with Encode was the best next step for the TEU criterion (filtering away 18.5% of the probesets) 1.58 TEU; taking a

union with Jetset after the encode intersection restored 4 probesets and increased the TEU very slightly to 1.60. No further union or intersection provided any improvement.

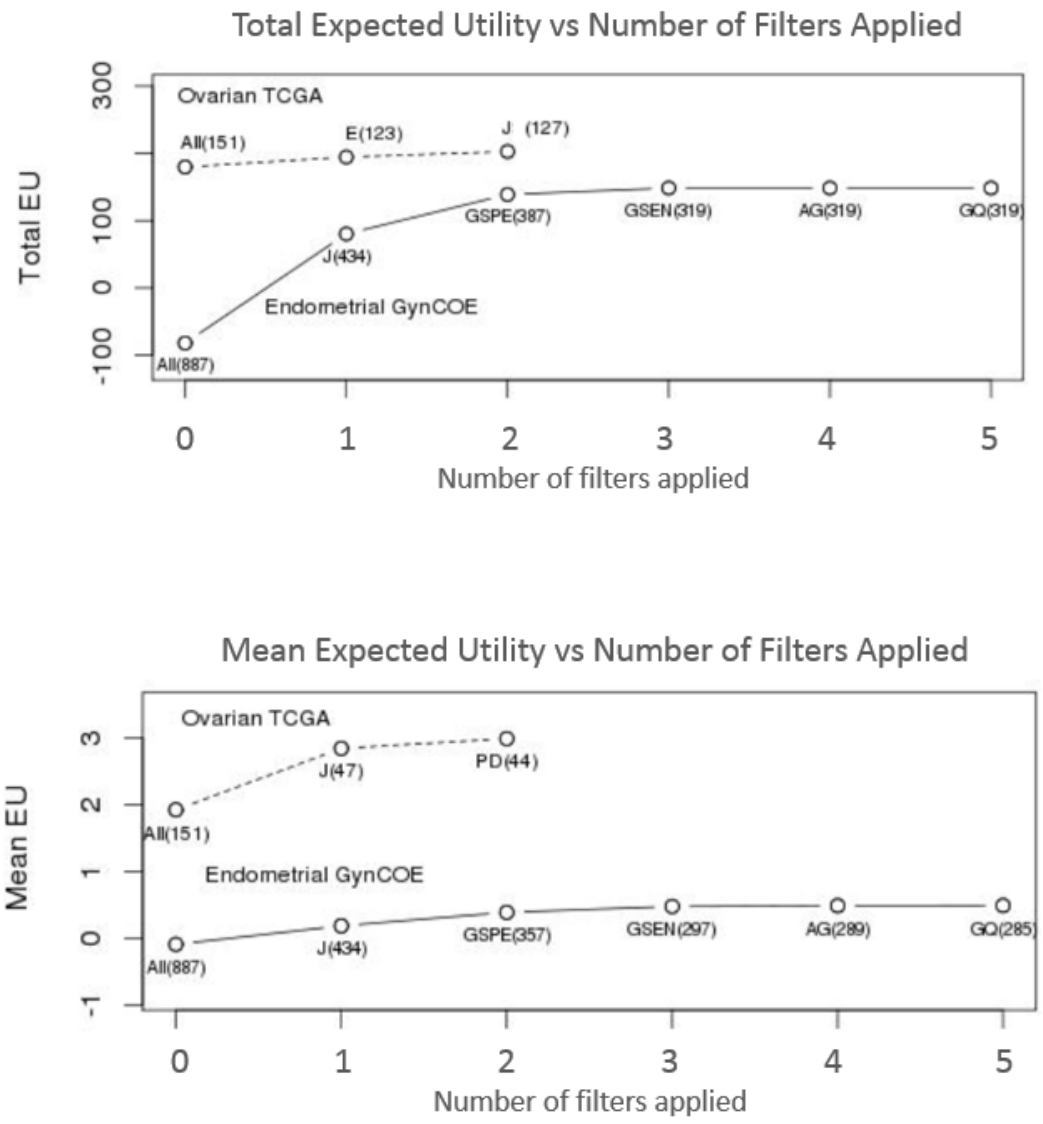


**Figure 25.** Greedy forward selection for probeset filtering example

Starting with all probesets the filters are applied to each cancer type using a greedy forward selection. The numbers of probesets are shown above each data point. The filter number represents the next filter intersection in the greedy forward selection, a union or an intersection. A) The total expected utility is the greedy forward selection criterion. B) The mean expected utility is the greedy forward selection criterion.

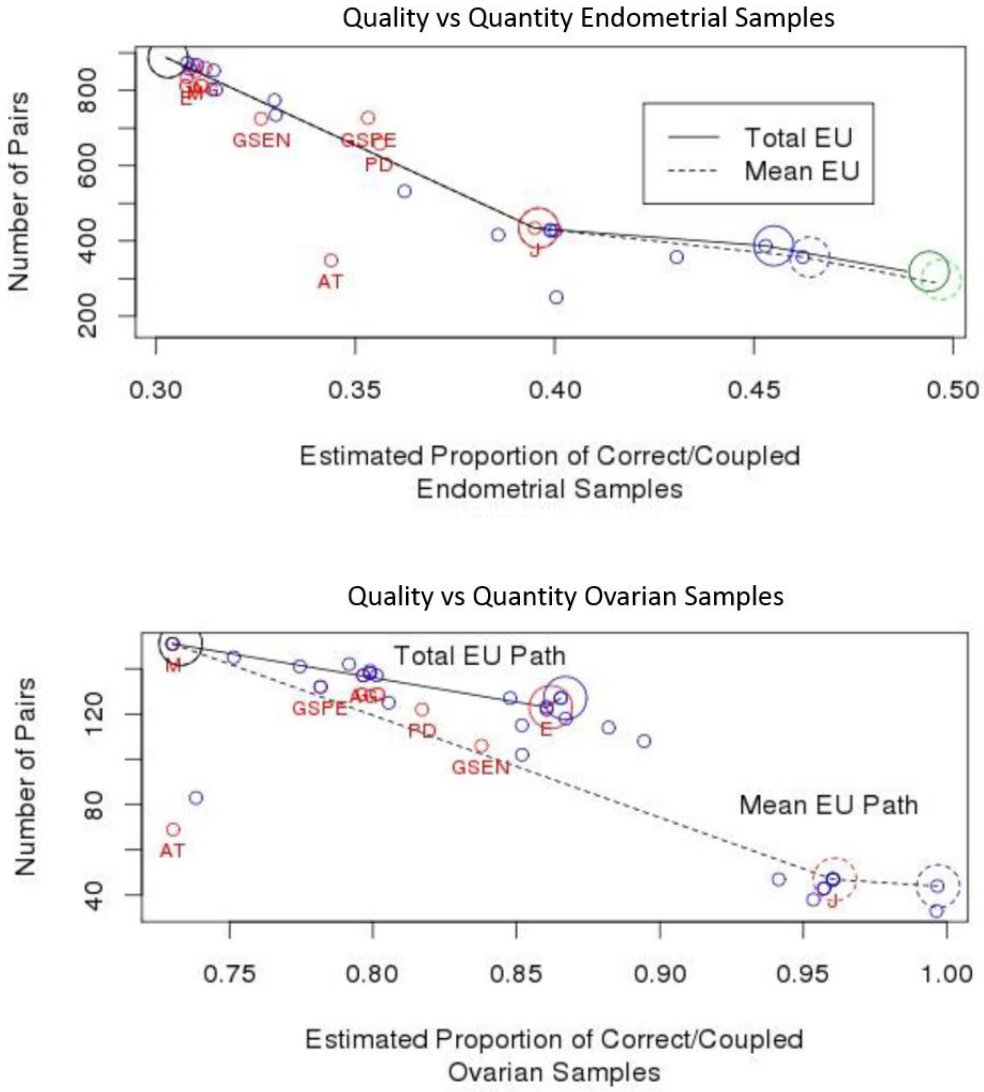


**Figure 26.** Quantity versus quality for probeset filtering example

Points are plotted for filtering strategies constructed from filtering methods by Boolean operators. A Level 1 strategy (red; "1") is a single filtering method. J: Jetset; GQ: Geneannot Quality; GSPE: Geneannot Specificity; GSEN: Geneannot Sensitivity; M: Masker; PD: Plandbaffy; AG:Affymetrix Grade; AT:Affymetrix Tag; E:Encode . A Level 2 strategy (blue; "2") is the intersection or union of two Level 1 strategies, and so forth. The lines connect the best strategies (circled) at each Boolean complexity level according to greedy forward selection.

In the Ovarian Serous Carcinoma TCGA dataset the Affymetrix to reverse phase protein assay data provide 151 pairs at a high Pr("+") = .733. Unlike the endometrial data the TEU and the MEU provide 2 different paths to "best practice" of probeset filtering. The MEU path chooses the Jetset filter method by throwing away all but the 47 pairs in the Jetset optimization with a Pr("+") of .961. After 2 levels the MEU maximizes to a Pr("+") = .997 and eliminated all but 44 pairs. The TEU favors quantity by keeping 123 pairs and a Pr("+") of .862. The TEU actually adds back in the union of the Jetset of 4 probesets to bring the total pairs to 127 and a Pr("+") of .867. Whether filtering with two methods rather than one is worth the extra effort is, of course, the judgment of the analyst.

## A.4 DISCUSSION

The goal of this work is to provide guidance for choosing a probeset filtering strategy in transcript expression experiments. It is not to guide analysis of paired datasets. The usefulness of linking with the protein abundance data is specifically to help evaluate and compare different filtering methods.

Many investigators assume that the filtering methods available are close enough that choice of filtering method will have a negligent effect on the overall results of an analysis. The odds comparing the methods pairwise demonstrate that some of the methods differ considerably. The two test-beds both selected Jetset as the best single-method strategy for the MEU criterion. This happens despite the fact that both the mRNA expression platform and the proteomics

platform are different between the two test-beds, and the range of correlations is quite different as well. This result provides encouragement that the evaluation methodology applied here can produce best-practices conclusions that can be useful for external microarray data sets.

Jetset eliminates as many as 80% of probesets. This may seem extreme to a user of microarrays but consider this: There are roughly three times the probesets as there are protein-coding genes in the human genome, so if (as Jetset does) a method selects only the best probeset for each gene, then a minimum of 2/3 of the probesets must be eliminated. When the goal of the biological research requires more features than a strict filtering method like Jetset would allow, and then Jetset would not be used. Our method reflects this, by granting the user goal-specific utility values that will penalize false negatives more stringently. If increasing this penalty leads to an unacceptable number of false positives, then the research goal cannot be achieved, and it is best that the investigator know this.

In the ovarian dataset, an investigator leery of discarding such a large proportion of probesets would be attracted to using Encode, guided by our TEU criterion. In both test-beds, Encode removes few probesets, but in the ovarian test-bed the probesets removed are of especially poor quality. This may be related to the fact that the mass spectrometry platform in the endometrial test-bed is not designed for accurate quantification. In contrast, the RPPA platform utilizes selected validated antibodies, so that one source of poor correlations is greatly reduced. Since RPPA data is a ligand based local protein expression assay the sensitivity for an individual protein is much higher than the LC MS/MS data. This method is sensitive to correlation of

mRNA expression to protein and the RPPA data has a more reliable protein measurement at low protein expression.

## A.5 CONCLUSIONS

The evaluation methodology applied here has some major virtues. Conclusions are developed on real, not simulated, data. Conclusions can be subject to replication independently on multiple test-beds. Conclusions are responsive to the needs of investigators through the decision theory framework, which helps an investigator decide how much data to filter away based on mRNA to protein correlation.

Many investigators utilize publicly available data, such as the TCGA data warehouse, to unlock discoveries at the genome, transcript, and protein levels of cancer biology. Previously, in merging and analyzing data from an expression data set and proteomic data on the same samples, our team found startling differences in the identifier mapping services. We developed a principled, data-grounded method to evaluate and compare these services. This method has broad generalizability to evaluating many kinds of data pipeline choices and strategies, including identifier filtering methods and read filtering methods to remove erroneous or poor quality features, and tuning parameter settings in pipelines. We are developing a new package that will support much wider applications to all kinds of workflow options. That package will include the decision theory component as well.

# APPENDIX B

## B.1 ECM ALGORITHM

Given model quality values (MQ) of feature pairs are distributed

$$MQ_{(p)} \sim N(\varphi_{G(p)}, \tau_{G(p)p})$$

Define variance as a convolution of measurement error variances of the pairs and group variance

$$\tau_{G(p)p} = \sigma_p^2 + V_{G(p)}$$

Where $\varphi_{G(p)}$ is the mean of group $\{" + ", "0", "x"\}$

Define the prior probability of a pair as a member of group g

$$\Pr(G(p) = g) = \pi_g \text{ for } g \in \{" + ", "0", "x"\}$$

Define the prior probability of the "-" mixture component as

$$\pi_- = \pi_0 + \pi_x$$

$$\Pr(G(p) = g) = \pi_g \text{ for } g \in \{" + ", " - "\}$$

Define group membership G(p) as missing data

Define the complete data likelihood per observation *k*

$$\Pr\left[MQ_p, G(p)\middle|\varphi, V, \pi\right] = \Pr[MQ_p|G(p)] \times \Pr[G(p)]$$

$$\propto \exp(-\frac{(MQ_p - \varphi_{G(p)})^2)}{2\tau_{G(p)p}} \times (\tau_{G(p)p})^{-1/2} \times \pi_{G(p)}$$

Set mean of the "0" component and "x " component to 0

$$\varphi_0 = \varphi_x = 0$$

Define posterior probability of a pair membership in the "-" component

$$\pi^*_{-p} = 1 - \pi^*_{+p} = \pi^*_{xp} + \pi^*_{0p}$$

Define free variable

$$\phi = (\varphi_- = 0, \varphi_+, \pi_+, V_-, V_+)$$

# E-step

Calculate expectation of the complete-data log likelihood

$$Q(\phi,\phi^*)=\frac{1}{2}\sum_k\sum_g \pi_{gk}^*(-\frac{(MQ_p-\varphi_{G(p)})^2}{(\sigma_p^2+V_g)}-\log(\sigma_p^2+V_g)+N_0^*\log\pi_0+N_1^*\log\pi_1$$

Where the posterior odds and solve expectation

$$\frac{\pi_{-p}^*}{\pi_{+p}^*}=\frac{\Pr(MQ_p,G(p)="-"|\varphi_-^*,V_-^*,\pi_-^*)}{\Pr(MQ_p,G(p)="+"|\varphi_+^*,V_+^*,\pi_+^*)}$$

$$=\frac{\pi_-^*}{\pi_+^*}\times\frac{\exp\left(-\frac{(MQ_p-\varphi_-^*)^2}{2(V_-^*+\sigma_p^2)}\right)/\sqrt{V_-^*+\sigma_p^2}}{\exp\left(-\frac{(MQ_p-\varphi_+^*)^2}{2(V_+^*+\sigma_p^2)}\right)/\sqrt{V_+^*+\sigma_p^2}}$$

$$E^*N_-=\sum_p \pi_{-p}^*$$

$$E^*N_+=\sum_p \pi_{+p}^*$$

Set Q partial derivatives to zero:

$$\frac{\partial Q}{\partial\varphi_g}=\sum_k \pi_{gp}^*\left((MQ_p-\varphi_g)(\sigma_p^2+\widehat{V_g})^{-1}\right)=0$$

$$\frac{\partial Q}{\partial V_g}=\frac{1}{2}\sum_k \pi_{gp}^*\left((MQ_p-\varphi_g)^2(\sigma_k^2+V_g)^{-2}-(\sigma_k^2+V_g)^{-1}\right)=0$$

# M-step

Iterate n-steps

Set $V_g = \hat{V}_g$

Solve

$$\hat{\varphi}_g = \frac{\sum_p \pi^*_{gp} MQ_p (\sigma_p^2 + \hat{V}_g)^{-1}}{\sum_p \pi^*_{gp} (\sigma_p^2 + \hat{V}_g)^{-1}}$$

Set $\varphi_g = \hat{\varphi}_g$

Solve

$$\hat{V}_g = \max(0, \sum\nolimits_p \pi^*_{gp} (MQ_p - \hat{\varphi}_g)^2 - \sum\nolimits_p \pi^*_{gp} \sigma_p^2)/N^*_g$$

## B.2 FEATURES FOR EVALUATION

**Table 16.** Features for RNASeq identifier filtering evaluation.

Hgnc Symbols, Descriptions and Filter status for the 62 features from the Identifier filtering and threshold evaluation on SALMON TCGA Breast cancer samples. Transmembrane = 0 is a non-transmembrane gene product based on BioMart results from TMHMM, Transmembrane =1 is a transmembrane gene product based on Biomart results from TMHMM, Low complexity = 0 detects a random intended amino acid sequence based on SEG results from BioMart, Low complexity =1 detects sequences with an ordered sequence structure similar to protein coding sections.

| *Hgnc_symbol* | *Description* | *Transmembrane* | *Low complexity region* |
|---|---|---|---|
| ACACA | acetyl-CoA carboxylase alpha | 0 | 1 |
| AKT1 | AKT serine/threonine kinase 1 | 0 | 0 |
| AKT2 | AKT serine/threonine kinase 2 | 0 | 0 |
| AKT3 | AKT serine/threonine kinase 3 | 0 | 0 |
| ANXA1 | annexin A1 | 0 | 0 |
| AR | androgen receptor | 0 | 1 |
| BAX | BCL2 associated X, apoptosis regulator | 1 | 1 |
| BCL2 | BCL2, apoptosis regulator | 1 | 1 |
| BCL2L1 | BCL2 like 1 | 0 | 0 |
| BCL2L11 | BCL2 like 11 | 1 | 1 |
| BECN1 | beclin 1 | 0 | 1 |
| CAV1 | caveolin 1 | 1 | 1 |
| CCNB1 | cyclin B1 | 0 | 1 |
| CCND1 | cyclin D1 | 0 | 1 |
| CCNE1 | cyclin E1 | 0 | 1 |
| CDH1 | cadherin 1 | 1 | 1 |
| CDH2 | cadherin 2 | 1 | 1 |
| CDKN1B | cyclin dependent kinase inhibitor 1B | 0 | 0 |
| CLDN7 | claudin 7 | 1 | 1 |
| COL6A1 | collagen type VI alpha 1 chain | 0 | 1 |
| CTNNA1 | catenin alpha 1 | 0 | 1 |
| CTNNB1 | catenin beta 1 | 0 | 1 |
| DVL3 | dishevelled segment polarity protein 3 | 0 | 1 |
| EEF2K | eukaryotic elongation factor 2 kinase | 0 | 1 |
| EGFR | epidermal growth factor receptor | 1 | 1 |
| EIF4E | eukaryotic translation initiation factor 4E | 0 | 1 |
| EIF4EBP1 | eukaryotic translation initiation factor 4E binding protein 1 | 0 | 0 |
| ERBB2 | erb-b2 receptor tyrosine kinase 2 | 1 | 1 |

| | | | |
|---|---|---|---|
| ERBB3 | erb-b2 receptor tyrosine kinase 3 | 1 | 1 |
| ERRFI1 | ERBB receptor feedback inhibitor 1 | 0 | 1 |
| ESR1 | estrogen receptor 1 | 0 | 1 |
| GATA3 | GATA binding protein 3 | 0 | 1 |
| GSK3A | glycogen synthase kinase 3 alpha | 0 | 1 |
| GSK3B | glycogen synthase kinase 3 beta | 0 | 1 |
| IGFBP2 | insulin like growth factor binding protein 2 | 0 | 1 |
| IRS1 | insulin receptor substrate 1 | 0 | 1 |
| KDR | kinase insert domain receptor | 1 | 1 |
| KIT | KIT proto-oncogene receptor tyrosine kinase | 1 | 1 |
| MAP2K1 | mitogen-activated protein kinase kinase 1 | 0 | 1 |
| MAPK1 | mitogen-activated protein kinase 1 | 0 | 1 |
| MAPK14 | mitogen-activated protein kinase 14 | 0 | 1 |
| MTOR | mechanistic target of rapamycin | 0 | 1 |
| NOTCH1 | notch 1 | 0 | 1 |
| PCNA | proliferating cell nuclear antigen | 0 | 0 |
| PECAM1 | platelet and endothelial cell adhesion molecule 1 | 1 | 1 |
| PGR | progesterone receptor | 0 | 1 |
| PRKAA1 | protein kinase AMP-activated catalytic subunit alpha 1 | 0 | 1 |
| PRKCA | protein kinase C alpha | 0 | 0 |
| PTEN | phosphatase and tensin homolog | 0 | 1 |
| PXN | paxillin | 0 | 1 |
| RAF1 | Raf-1 proto-oncogene, serine/threonine kinase | 0 | 1 |
| RPS6 | ribosomal protein S6 | 0 | 1 |
| RPS6KB1 | ribosomal protein S6 kinase B1 | 0 | 0 |
| SMAD1 | SMAD family member 1 | 0 | 1 |
| SMAD3 | SMAD family member 3 | 0 | 1 |
| SMAD4 | SMAD family member 4 | 0 | 1 |
| SRC | SRC proto-oncogene, non-receptor tyrosine kinase | 0 | 1 |
| STAT5A | signal transducer and activator of transcription 5A | 0 | 1 |
| STMN1 | stathmin 1 | 0 | 0 |
| SYK | spleen associated tyrosine kinase | 0 | 0 |
| TP53 | tumor protein p53 | 1 | 1 |
| YBX1 | Y-box binding protein 1 | 0 | 1 |

## BIBLIOGRAPHY


1. Day RS, McDade KK. A decision theory paradigm for evaluating identifier mapping and filtering methods using data integration. *BMC Bioinformatics*. 2013;14(1):223. doi:10.1186/1471-2105-14-223.
2. McDade KK, Chandran U, Day RS. Improving cancer gene expression data quality through a TCGA data-driven evaluation of identifier filtering. *Cancer Inform*. 2015;14:149-161. doi:10.4137/CIN.S33076.
3. Day RS, McDade KK, Chandran UR, et al. Identifier mapping performance for integrating transcriptomics and proteomics experimental results. *BMC Bioinformatics*. 2011;12(1):213. doi:10.1186/1471-2105-12-213.
4. National Cancer Institute Center for Bioinformatics. The Cancer Genome Atlas.
5. Li Q, Birkbak NJ, Gyorffy B, Szallasi Z, Eklund AC. Jetset: selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics*. 2011;12(1):474. doi:10.1186/1471-2105-12-474.
6. Nurtdinov RN, Vasiliev MO, Ershova AS, Lossev IS, Karyagina AS. PLANdbAffy: probe-level annotation database for Affymetrix expression microarrays. *Nucleic Acids Res*. 2010;38(Database issue):D726-D730. doi:10.1093/nar/gkp969.
7. Liu G, Loraine AE, Shigeta R, et al. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res*. 2003;31(1):82-86.
8. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. 2010;26(4):493-500. doi:10.1093/bioinformatics/btp692.

9. Wang K, Singh D, Zeng Z, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*. 2010;38(18):e178. doi:10.1093/nar/gkq622.

10. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol*. 2014;32(5):462-464. doi:10.1038/nbt.2862.

11. Zhang L, Wei Q, Mao L, Liu W, Mills GB, Coombes K. Serial dilution curve : a new method for analysis of reverse phase protein array data. 2009;25(5):650-654. doi:10.1093/bioinformatics/btn663.

12. Sun M, Lai D, Zhang LI, Huang X. Modified SuperCurve Method for Analysis of Reverse-Phase Protein Array Data. 2015;22(8):765-769. doi:10.1089/cmb.2015.0007.

13. Coulibaly L, Henry E, He B, Barillot E. NormaCurve : A SuperCurve-Based Method That Simultaneously Quantifies and Normalizes Reverse Phase Protein Array Data. 2012;7(6). doi:10.1371/journal.pone.0038686.

14. Wang W, Lin F, Chang W, Lin K, Huang H, Lin N. miRExpress : Analyzing high-throughput sequencing data for profiling microRNA expression. 2009;13:1-13. doi:10.1186/1471-2105-10-328.

15. Yang J, Shao P, Zhou H, Chen Y, Qu L. deepBase : a database for deeply annotating and mining deep sequencing data. 2010;38(December 2009):123-130. doi:10.1093/nar/gkp943.

16. Lienhard M, Grasse S, Rolff J, et al. QSEA — modelling of genome-wide DNA methylation from sequencing enrichment experiments. 2017;45(6). doi:10.1093/nar/gkw1193.

17. Niazi U, Geyer KK, Vickers MJ, Hoffmann KF, Swain MT. DISMISS : detection of stranded methylation in MeDIP-Seq data. *BMC Bioinformatics*. 2016:1-12. doi:10.1186/s12859-016-1158-7.

18. Affymetrix. Data Sheet. *GeneChip Hum Genome U133 Arrays*. 2007:1-8. https://cancergenome.nih.gov/abouttcga/aboutdata/platformdesign/affymetrixU133array.

19. Illumina. HiSeq ® 2000 System User Guide. 2014;(November).

20. Tibes R, Qiu Y, Lu Y, et al. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Mol Cancer Ther*. 2006;5(10):2512-2521. doi:10.1158/1535-7163.MCT-06-0334.

21. Clarke C, Henry M, Doolan P, et al. Integrated miRNA, mRNA and protein expression analysis reveals the role of post-transcriptional regulation in controlling CHO cell growth rate. *BMC Genomics*. 2012;13(1):656. doi:10.1186/1471-2164-13-656.

22. Weisenberger DJ, Berg D Van Den, Pan F, et al. Comprehensive DNA Methylation Assay Platform.

23. Tosta FE, Braganholo V, Murta L, Mattoso M. Improving workflow design by mining reusable tasks. *J Brazilian Comput Soc*. 2015. doi:10.1186/s13173-015-0035-y.

24. Streit M, Member S, Lex A, Member S, Schmalstieg D, Schumann H. Model-Driven Design for the Visual Analysis of Heterogeneous Data. 2012;18(6):998-1010.

25. Mieczkowski J, Tyburczy ME, Dabrowski M, Pokarowski P. Probe set filtering increases correlation between Affymetrix GeneChip and qRT-PCR expression measurements. *BMC Bioinformatics*. 2010;11:104. doi:10.1186/1471-2105-11-104.

26. Calza S, Raffelsberger W, Ploner A, Sahel J, Leveillard T, Pawitan Y. Filtering genes to improve sensitivity in oligonucleotide microarray data analysis. 2007;35(16). doi:10.1093/nar/gkm537.

27. Aanes H, Winata C, Moen LF, et al. Normalization of RNA-sequencing data from samples with varying mRNA levels. *PLoS One*. 2014;9(2):e89158. doi:10.1371/journal.pone.0089158.

28. Mcnutt P, Gut I, Hubbard K, Beske P. A novel method to prioritize RNAseq data for post-hoc analysis based on absolute changes in transcript abundance. 2015;14(3):227-241. doi:10.1515/sagmb-2014-0018.

29. Razumovskaya J, Olman V, Xu D, et al. A computational method for assessing peptide-identification reliability in tandem mass spectrometry analysis with SEQUEST. *Proteomics*. 2004;4(4):961-969. doi:10.1002/pmic.200300656.

30. St Laurent G, Shtokalo D, Tackett MR, et al. On the importance of small changes in RNA expression. *Methods*. 2013;63(1):18-24. doi:10.1016/j.ymeth.2013.03.027.

31. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus : NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-210.

32. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2(5):401-404. doi:10.1158/2159-8290.CD-12-0095.

33. Akbani R, Ng PKS, Werner HMJ, et al. A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat Commun*. 2014;5(May):3887. doi:10.1038/ncomms4887.

34. Rahman M, Jackson LK, Johnson WE, Li DY, Bild AH, Piccolo SR. Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. 2015;31(July):3666-3672. doi:10.1093/bioinformatics/btv377.

35. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80. doi:10.1186/gb-2004-5-10-r80.

36. Ruffalo M, LaFramboise T, Koyutürk M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*. 2011;27(20):2790-2796. doi:10.1093/bioinformatics/btr477.

37. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform*. December 2013. doi:10.1093/bib/bbt086.

38. Liu X, Han S, Wang Z, Gelernter J, Yang B-Z. Variant callers for next-generation sequencing data: a comparison study. *PLoS One*. 2013;8(9):e75619. doi:10.1371/journal.pone.0075619.

39. Irmler M, Hartl D, Schmidt T, et al. An approach to handling and interpretation of ambiguous data in transcriptome and proteome comparisons. *Proteomics*. 2008;8(6):1165-1169. doi:10.1002/pmic.200700741.

40. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18(9):1509-1517. doi:10.1101/gr.079558.108.

41. Vescovo V Del, Meier T, Inga A, Denti MA, Borlak J. A Cross-Platform Comparison of Affymetrix and Agilent Microarrays Reveals Discordant miRNA Expression in Lung Tumors of c-Raf Transgenic Mice. 2013;8(11). doi:10.1371/journal.pone.0078870.

42. Shedden K, Chen W, Kuick R, et al. Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. *BMC Bioinformatics*. 2005;6:26. doi:10.1186/1471-2105-6-26.

43. Grant GR, Farkas MH, Pizarro AD, et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*. 2011;27(18):2518-2528. doi:10.1093/bioinformatics/btr427.

44. Engström PG, Steijger T, Sipos B, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods*. 2013;10(12):1185-1191. doi:10.1038/nmeth.2722.

45. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*. 2010;11:94. doi:10.1186/1471-2105-11-94.

46. Marot G, Castel D, Estelle J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. 2012;14(6). doi:10.1093/bib/bbs046.

47. Jiang N, Leach LJ, Hu X, et al. Methods for evaluating gene expression from Affymetrix microarray datasets. *BMC Bioinformatics*. 2008;9:284. doi:10.1186/1471-2105-9-284.

48. Gertz EM, Sengupta K, Difilippantonio MJ, Ried T, Schäffer A a. Evaluating annotations of an Agilent expression chip suggests that many features cannot be interpreted. *BMC Genomics*. 2009;10:566. doi:10.1186/1471-2164-10-566.

49. Yu V, Fagan L, Wraith S. Antimicrobial selection by a computer. A blinded evaluation by infectious diseases experts. *JAMA J …*. 1979. http://ukpmc.ac.uk/abstract/MED/480542. Accessed October 19, 2012.

50. Draghici S, Sellamuthu S, Khatri P. Babel's tower revisited: a universal resource for cross-referencing across annotation databases. *Bioinformatics*. 2006;22(23):2934-2939.

51. Huang da W, Sherman BT, Tan Q, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res*. 2007;35(Web Server issue):W169-W175.

52. Kahlem P, Clegg A, Reisinger F, et al. ENFIN--A European network for integrative systems biology. *Comptes Rendus Biol*. 2009;332(11):1050-1058. doi:10.1016/j.crvi.2009.09.003.

53. Côté RG, Jones P, Martens L, et al. The Protein Identifier Cross-Referencing ( PICR ) service : reconciling protein identifiers across multiple source databases. 2007;14:1-14. doi:10.1186/1471-2105-8-401.

54. Gao J, Zhang C, Iersel M Van, et al. BridgeDb app : unifying identifier mapping services for Cytoscape [ v1 ; ref status : indexed , http://f1000r.es/3qb ]. 2014;100039:1-7. doi:10.12688/f1000research.4521.1.

55. Saetrom O, Snøve O, Saetrom P. Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. *RNA*. 2005;11(7):995-1003. doi:10.1261/rna.7290705.

56. Sá PHCG De, Veras AAO, Carneiro AR, et al. The impact of quality filter for RNA-Seq. *Gene*. 2015;563(2):165-171. doi:10.1016/j.gene.2015.03.033.

57. Netaffx T, Genechip A. Transcript Assignment for NetAffx ™ Annotations. 2006;(Figure 1):1-9.

58. Sleuthing With the Affymetrix NetAffx ™ Website: Identifing and Examing Probe Sets and Their Genomic Context.

59. Zhang, Jinghui; Finney, Richard; Beutow K. Custom Chip Definition Files (CDF) for Unified Gene Expression Analysis with Affymetrix Chips. 2005. http://masker.nci.nih.gov/ev/.

60. Ferrari F, Bortoluzzi S, Coppe A, et al. Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics*. 2007;8:446. doi:10.1186/1471-2105-8-446.

61. Harrow J, Denoeud F, Frankish A, et al. GENCODE: producing a reference annotation for ENCODE. *Genome Biol*. 2006;7 Suppl 1:S4.1-9. doi:10.1186/gb-2006-7-s1-s4.

62. Nurtdinov RN, Vasiliev MO, Ershova AS, Lossev IS, Karyagina AS. PLANdbAffy: probe-level annotation database for Affymetrix expression microarrays. *Nucleic Acids Res*. 2010;38(Database issue):D726-D730. doi:10.1093/nar/gkp969.

63. The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*. 2011;9(4):e1001046. doi:10.1371/journal.pbio.1001046.

64. Sleuthing With the Affymetrix NetAffx ™ Website: Identifing and Examing Probe Sets and Their Genomic Context. http://www.affymetrix.com/support/technical/whitepapers/Sleuthing_NetAffx_whitepaper.pdf.

65. Buschmann V, Kapusta P, Erdmann R. Use of Time-Resolved Fluorescence To Improve Sensitivity and Dynamic Range of Gel-Based Proteomics. 2016. doi:10.1021/acs.analchem.5b03805.

66. Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics*. 2016:1-13. doi:10.1186/s12859-016-0956-2.

67. Molnar M, Ilie L. Correcting Illumina data. *Brief Bioinform*. September 2014. doi:10.1093/bib/bbu029.

68. Vijay N, Poelstra JW, Künstner A, Wolf JBW. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol Ecol*. 2013;22(3):620-634. doi:10.1111/mec.12014.

69. Ginsbach P, Drews R, Paramasivam N, et al. detection in cancer using whole-genome sequencing. 2015. doi:10.1038/ncomms10001.

70. O'Rawe J, Jiang T, Sun G, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med*. 2013;5(3):28. doi:10.1186/gm432.

71. Jonassen I, Bolser D, Uludag M, et al. Databases and ontologies EDAM : an ontology of bioinformatics operations , types of data and identifiers , topics and formats. 2013;29(10):1325-1332. doi:10.1093/bioinformatics/btt113.

72. Kalaš M, Puntervoll P, Joseph A, et al. BioXSD : the common data-exchange format for everyday bioinformatics web services. 2010;26:540-546. doi:10.1093/bioinformatics/btq391.

73. Oinn T, Addis M, Ferris J, et al. Taverna : a tool for the composition and enactment of bioinformatics workflows. 2004;20(17):3045-3054. doi:10.1093/bioinformatics/bth361.

74. Bartocci E, Corradini F, Merelli E, Scortichini L. BioWMS : a web-based Workflow Management System for bioinformatics. 2007;14:1-14. doi:10.1186/1471-2105-8-S1-S2.

75. Goecks J, Nekrutenko A, Taylor J, Team TG. Galaxy : a comprehensive approach for supporting accessible , reproducible , and transparent computational research in the life sciences. 2010.

76. Lushbourgh C, Bergman MK, Lawernce CJ, et al. BioExtract Server — An Integrated Workflow - Enabling System to Access and Analyze Heterogeneous , Distributed Biomolecular Data. 2016:1-4. doi:10.1109/TCBB.2008.98.

77. Chapman SJ, Khor CC, Davies WH, et al. GenePattern 2.0. 2006;38(5):500-501.

78. Sandberg R, Larsson O. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics*. 2007;8:48. doi:10.1186/1471-2105-8-48.

79. Transcript Assignment for NetAffx Annotations : Affymetrix GeneChip IVT Array Whitepaper Collection. 2006;2.3:1-9.

80. Ferrari F, Bortoluzzi S, Coppe A, et al. Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics*. 2007;8:446. doi:10.1186/1471-2105-8-446.

81. Carter SL, Eklund AC, Mecham BH, Kohane IS, Szallasi Z. Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics*. 2005;6:107. doi:10.1186/1471-2105-6-107.

82. Ballester B, Johnson N, Proctor G, Flicek P. Consistent annotation of gene expression arrays. *BMC Genomics*. 2010;11:294. doi:10.1186/1471-2164-11-294.

83. Hu Z, Willsky GR. Utilization of two sample t-test statistics from redundant probe sets to evaluate different probe set algorithms in GeneChip studies. *BMC Bioinformatics*. 2006;7:12. doi:10.1186/1471-2105-7-12.

84. Jaksik R, Polańska J, Herok R, Rzeszowska-Wolny J. Calculation of reliable transcript levels of annotated genes on the basis of multiple probe-sets in Affymetrix microarrays. *Acta Biochim Pol*. 2009;56(2):271-277. http://www.ncbi.nlm.nih.gov/pubmed/19436837.

85. Li H, Zhu D, Cook M. A statistical framework for consolidating "sibling" probe sets for Affymetrix GeneChip data. *BMC Genomics*. 2008;9:188. doi:10.1186/1471-2164-9-188.

86. Schneider S, Smith T, Hansen U. SCOREM: statistical consolidation of redundant expression measures. *Nucleic Acids Res*. 2012;40(6):e46. doi:10.1093/nar/gkr1270.

87. Cambon AC, Khalyfa A, Cooper NGF, Thompson CM. Analysis of probe level patterns in Affymetrix microarray data. *BMC Bioinformatics*. 2007;8:146. doi:10.1186/1471-2105-8-146.

88. van den Berg BHJ, McCarthy FM, Lamont SJ, Burgess SC. Re-annotation is an essential step in systems biology modeling of functional genomics data. *PLoS One*. 2010;5(5):e10642. doi:10.1371/journal.pone.0010642.

89. de Leeuw WC, Rauwerda H, Jonker MJ, Breit TM. Salvaging Affymetrix probes after probe-level re-annotation. *BMC Res Notes*. 2008;1:66. doi:10.1186/1756-0500-1-66.

90. Gautier L, Cope L, Bolstad BM, Irizarry R a. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004;20(3):307-315. doi:10.1093/bioinformatics/btg405.

91. Liu H, Zeeberg BR, Qu G, et al. AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets. *Bioinformatics*. 2007;23(18):2385-2390. doi:10.1093/bioinformatics/btm360.

92. Liu X, Milo M, Lawrence ND, Rattray M. Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics*. 2006;22(17):2107-2113. doi:10.1093/bioinformatics/btl361.

93. Harrow J, Denoeud F, Frankish A, et al. GENCODE: producing a reference annotation for ENCODE. *Genome Biol*. 2006;7 Suppl 1:S4.1-9. doi:10.1186/gb-2006-7-s1-s4.

94. Yang K-C, Yamada K a, Patel AY, et al. Deep RNA sequencing reveals dynamic regulation of myocardial noncoding RNAs in failing human heart and remodeling with mechanical circulatory support. *Circulation*. 2014;129(9):1009-1021. doi:10.1161/CIRCULATIONAHA.113.003863.

95. Wong K-K, Izaguirre DI, Kwan S-Y, et al. Poor survival with wild-type TP53 ovarian cancer? *Gynecol Oncol*. 2013;130(3):565-569. doi:10.1016/j.ygyno.2013.06.016.

96. Mills JD, Nalpathamkalam T, Jacobs HIL, et al. RNA-Seq analysis of the parietal cortex in Alzheimer's disease reveals alternatively spliced isoforms related to lipid metabolism. *Neurosci Lett*. 2013;536:90-95. doi:10.1016/j.neulet.2012.12.042.

97. Han L, Vickers KC, Samuels DC, Guo Y. Alternative applicationns for distinct RNA sequencing strategies. *Brief Bioinform*. September 2014. doi:10.1093/bib/bbu032.

98. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105-1111. doi:10.1093/bioinformatics/btp120.

99. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57-63. doi:10.1038/nrg2484.

100. Trapnell C, Williams B a, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511-515. doi:10.1038/nbt.1621.

101. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324.

102. Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics*. 2008;24(5):713-714. doi:10.1093/bioinformatics/btn025.

103. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. 2008;18(11):1851-1858. doi:10.1101/gr.078212.108.

104. Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics*. 2008;9:128. doi:10.1186/1471-2105-9-128.

105. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods*. 2011;8(6):469-477. doi:10.1038/nmeth.1613.

106. Lam TW, Sung WK, Tam SL, Wong CK, Yiu SM. Compressed indexing and local alignment of DNA. *Bioinformatics*. 2008;24(6):791-797. doi:10.1093/bioinformatics/btn032.

107. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25. doi:10.1186/gb-2009-10-3-r25.

108. Li R, Yu C, Li Y, et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*. 2009;25(15):1966-1967. doi:10.1093/bioinformatics/btp336.

109. Klus P, Lam S, Lyberg D, et al. BarraCUDA - a fast short read sequence aligner using graphics processing units. *BMC Res Notes*. 2012;5(1):27. doi:10.1186/1756-0500-5-27.

110. Rumble SM, Lacroute P, Dalca A V, Fiume M, Sidow A, Brudno M. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol*. 2009;5(5):e1000386. doi:10.1371/journal.pcbi.1000386.

111. Clement NL, Snell Q, Clement MJ, et al. The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics*. 2010;26(1):38-45. doi:10.1093/bioinformatics/btp614.

112. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357-359. doi:10.1038/nmeth.1923.

113. De Bona F, Ossowski S, Schneeberger K, Rätsch G. Optimal spliced alignments of short sequence reads. *Bioinformatics*. 2008;24(16):i174-i180. doi:10.1093/bioinformatics/btn300.

114. Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. 2008;5(7):1-8. doi:10.1038/NMETH.1226.

115. Dobin A, Davis C a, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635.

116. Guttman M, Garber M, Levin JZ, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol*. 2010;28(5):503-510. doi:10.1038/nbt.1633.

117. Dilworth R. A decomposition theorem for partially ordered sets. *Annu Math*. 1950;51:161-166.

118. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 2008;18(5):821-829. doi:10.1101/gr.074492.107.

119. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323. doi:10.1186/1471-2105-12-323.

120. Habegger L, Sboner A, Gianoulis T a, et al. RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics*. 2011;27(2):281-283. doi:10.1093/bioinformatics/btq643.

121. Goncalves a., Tikhonov a., Brazma a., Kapushesky M. A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics*. 2011;27(6):867-869. doi:10.1093/bioinformatics/btr012.

122. Wang Y, Mehta G, Mayani R, et al. RseqFlow: Workflows for RNA-Seq data analysis. *Bioinformatics*. 2011;27(18):2598-2600. doi:10.1093/bioinformatics/btr441.

123. Kalari KR, Nair A a, Bhavsar JD, et al. MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing. *BMC Bioinformatics*. 2014;15(1):224. doi:10.1186/1471-2105-15-224.

124. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26(5):589-595. doi:10.1093/bioinformatics/btp698.

125. Kvam, Vanessa; Liu,Peng; Si Y. A Comparison of Statistical Methods for Detecting Differentially expressed Genes From RNA-Seq Data. *Am J Bot*. 2012;99(2):248-256. doi:10.3732/ajb.1100340.

126. Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics*. 2017:1-12. doi:10.1186/s12859-016-1457-z.

127. Huber W, Gentleman R. Using oligonucleotide microarray reporter sequence information for preprocessing and quality assessment. 2010:1-6.

128. Fu X, Fu N, Guo S, et al. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*. 2009;10:161. doi:10.1186/1471-2164-10-161.

129. Francis H.S. Crick. On Protein Synthesis. *Symp Soc Exp Biol*. 1958;12:138-163.

130. Turro E, Bochkina N, Hein A-MK, Richardson S. BGX: a Bioconductor package for the Bayesian integrated analysis of Affymetrix GeneChips. *BMC Bioinformatics*. 2007;8:439. doi:10.1186/1471-2105-8-439.

131. Shankavaram UT, Reinhold WC, Nishizuka S, et al. Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Mol Cancer Ther*. 2007;6(3):820-832. doi:10.1158/1535-7163.MCT-06-0650.

132. Rogers S, Girolami M, Kolch W, et al. Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics*. 2008;24(24):2894-2900. doi:10.1093/bioinformatics/btn553.

133. Zeeberg BR, Kohn KW, Kahn A, et al. Concordance of Gene Expression and Functional Correlation Patterns across the NCI-60 Cell Lines and the Cancer Genome Atlas Glioblastoma Samples. 2012;7(7):1-9. doi:10.1371/journal.pone.0040062.

134. Tainsky M a. Genomic and proteomic biomarkers for cancer: a multitude of opportunities. *Biochim Biophys Acta*. 2009;1796(2):176-193. doi:10.1016/j.bbcan.2009.04.004.

135. Gao J. Correlating Protein Phosphorylation with Genomic Alterations in Cancer RPPA : Reverse phase protein arrays. *TCGA Meet*. https://www.genome.gov/Multimedia/Slides/TCGA1/TCGA1_Gao.pdf.

136. Reeves G a, Talavera D, Thornton JM. Genome and proteome annotation: organization, interpretation and integration. *J R Soc Interface*. 2009;6(31):129-147. doi:10.1098/rsif.2008.0341.

137. Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. *Nature*. 2008;455(7209):58-63. doi:10.1038/nature07228.

138. Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. *Nature*. 2008;455(7209):64-71. doi:10.1038/nature07242.

139. Miles GD, Seiler M, Rodriguez L, Rajagopal G, Bhanot G. Identifying microRNA/mRNA dysregulations in ovarian cancer. *BMC Res Notes*. 2012;5(1):164. doi:10.1186/1756-0500-5-164.

140. Lisovich A, Day RS. The IdMappingAnalysis package in Bioconductor : Critically comparing identifier maps retrieved from bioinformatics annotation resources. *Version 121 Bioconductor Release 211*. 2012:1-18. http://www.bioconductor.org/packages/release/bioc/html/IdMappingAnalysis.html .

141. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352.

142. Dozmorov MG, Adrianto I, Giles CB, et al. Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data. *BMC Bioinformatics*. 2015;16(Suppl 13):S10. doi:10.1186/1471-2105-16-S13-S10.

143. Sonnhammer ELL, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. 1998.

144. Wan H, Li L, Federhen S, Wootton JC. Discovering Simple Regions in Biological Sequences Associated with Scoring Schemes. *J Comput Biol*. 2003;10(2):171-185.

145. Mcnutt M. Reproducibility. *Science (80- ).* 2014;343(January):229. doi:10.1126/science.1250475.

146. Stodden BV, Mcnutt M, Bailey DH, et al. Enhancing reproducibility for computational methods.

147. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus : NCBI gene expression and hybridization array data repository. 2002;30(1):207-210.

148. Encode T, Consortium P. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*. 2011;9(4):e1001046. doi:10.1371/journal.pbio.1001046.

149. Day RS, McDade KK. A decision theory paradigm for evaluating identifier mapping and filtering methods using data integration. *BMC Bioinformatics*. 2013;14(1):223. doi:10.1186/1471-2105-14-223.

150. Day RS, McDade KK, Chandran UR, et al. Identifier mapping performance for integrating transcriptomics and proteomics experimental results. *BMC Bioinformatics*. 2011;12(1):213. doi:10.1186/1471-2105-12-213.

151. Iorns E, Lord CJ, Grigoriadis A, et al. Integrated functional, gene expression and genomic analysis for the identification of cancer targets. *PLoS One*. 2009;4(4):e5120. doi:10.1371/journal.pone.0005120.

152. Irmler M, Hartl D, Schmidt T, et al. An approach to handling and interpretation of ambiguous data in transcriptome and proteome comparisons. *Proteomics*. 2008;8(6):1165-1169. doi:10.1002/pmic.200700741.

153. Fu X, Fu N, Guo S, et al. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*. 2009;10:161. doi:10.1186/1471-2164-10-161.

154. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18(9):1509-1517. doi:10.1101/gr.079558.108.

155. Iorio M V, Croce CM. MicroRNAs in cancer: small molecules with a huge impact. *J Clin Oncol*. 2009;27(34):5848-5856. doi:10.1200/JCO.2009.24.0317.

156. Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. *Nature*. 2008;455(7209):64-71. doi:10.1038/nature07242.

157. Chen G. Discordant Protein and mRNA Expression in Lung Adenocarcinomas. *Mol Cell Proteomics*. 2002;1(4):304-313. doi:10.1074/mcp.M200008-MCP200.

158. The ENCODE Consortium. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474(7353):609-615. doi:10.1038/nature10166.

159. Lisovich A, Day RS. IdMappingRetrieval: Id Mapping Data Retrieval. 2013. http://www.bioconductor.org/packages/2.12/bioc/manuals/IdMappingRetrieval/man/IdMappingRetrieval.pdf.

160. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80. doi:10.1186/gb-2004-5-10-r80.

161. Kahlem P, Birney E. ENFIN a network to enhance integrative systems biology. *Ann N Y Acad Sci*. 2007;1115(0):23-31. doi:10.1196/annals.1407.016.

162. Maxwell GL, Hood BL, Day R, et al. Gynecologic Oncology Proteomic analysis of stage I endometrial cancer tissue : Identification of proteins associated with oxidative processes and inflammation. *Gynecol Oncol*. 2011;121(3):586-594. doi:10.1016/j.ygyno.2011.02.031.

163. Risinger JI, Allard J, Chandran U, et al. Gene expression analysis of early stage endometrial cancers reveals unique transcripts associated with grade and histology but not depth of invasion. *Front Oncol*. 2013;3(June):1-10. doi:10.3389/fonc.2013.00139.

164. Lisovich A, Day RS. IdMappingRetrieval: Id Mapping Data Retrieval. 2013.

165. Szumilas M. Explaining Odds Ratios. *J Can Acad Child Adolesc Psychiatry*. 2010;(August):227-229.